Essay Revision and Corresponding Grade Change as Captured by Text Similarity and Revision Purposes

Sonia Cromp
University of Pittsburgh
snc40@pitt.edu

Diane Litman
University of Pittsburgh
dlitman@pitt.edu

ABSTRACT

Writing and revision are abstract skills that can be challenging to teach to students. Automatic essay revision assistants offer to help in this area because they compare two drafts of a student's essay and analyze the revisions performed. For these assistants to be useful, they need to provide useful information such as whether the revisions are likely to lead to an improvement in the student's grade. It is necessary to better understand the connection between revisions and grade change so that this information could be displayed in an assistant. So, this work explores the relationship between the tf-idf cosine similarity of two essay drafts and resulting essay grade change. Prior work has demonstrated that identifying the revisions between drafts, then labeling each revision with the purpose behind why the revision was performed is useful to predicting grade change. However, this process is expensive because this sort of annotation is time-consuming for humans. Moreover, classifiers achieve lower accuracy than humans when predicting purposes. Using similarity measures instead of or as supplement to revision purposes may correct these issues, as similarity can be computed automatically and without the issue of classification accuracy. As such, the correlations between grade change and the similarity measure are compared to the correlations between grade change and revision purposes with the potential use-case of an automatic writing assistant in mind. Findings suggest tf-idf cosine similarity captures overall essay and overall grade change while revision purposes capture lighter changes that fix errors or cause the essay to read better.

Keywords

writing, revision, education, similarity, NLP

1. INTRODUCTION

Assessing academic essay revision is useful in many applications, such as writing assistants that focus on revision or analyzing what makes revisions effective. As such, it is important to investigate reliable methods to quantify revisions as well as how these revisions relate to external measures such as grade change across drafts.

Prior work [16, 14] has focused on analyzing the purposes or intentions behind the revisions performed, such as labeling revisions as "Conventions" when they correct a spelling error or "Evidence" when adding an example to support a claim. Unsurprisingly, a greater quantity of revisions is associated with a greater grade change. Further, most university-level writing assignment rubrics place more importance on ideas, reasoning and evidence than spelling or adherence to writing conventions. As such, revisions to change the meaning of an essay, such as Evidence, Claims or Reasoning revisions, are associated with more grade change than minor changes such as Conventions revisions. These works have demonstrated revision purposes to be useful in assessing essay change.

However, revision purposes are time-consuming to obtain via human annotation and classifiers achieve lower classification accuracy than do human annotators. First, the sentences in the two drafts must be aligned to indicate which sentences in each draft correspond to each other. Second, revision operations are determined. Sentences removed from the old draft are marked as deleted, those inserted to the new draft marked as added and those edited but present in both drafts marked as modified. Third, each sentence that has been modified, added or deleted must be labeled with a revision purpose. For an example of an annotated essay from the ArgRewrite V.2 corpus[1], see Table 1.

Embedding-based similarity measures offer an alternative to revision purposes that can be obtained in fewer steps, automatically and without the issue of classification accuracy. However, similarity measures can also be calculated in additional ways when more annotation such as sentence alignments is available. Similarity measures likely are also able to capture additional information that cannot be detected by revision counts alone.

As an extreme example, consider two college students, A and B, revising their essays that are both N sentences long. Student A replaces one word in each sentence of their essay with a synonym while Student B re-organizes their essay to present their logic more clearly. Then, consider that an automated revision assistant were to label revision purposes using the popular binary schema from [5] that distinguishes between revisions that change meaning (such as

Old Draft	New Draft	Operation	Purpose
Anything that can save lives is good for society.		Delete	Claims
Despite the limitations in technology, self-driving will save lives.	Despite these technological limitations, self-driving will save lives.	Modify	Fluency
No other benefit matters.		Delete	General Content
Most traffic fatalities should have been prevented because the drivers simply should not have been driving.	Most traffic fatalities should have been prevented because the drivers should not have been driving.	Modify	Fluency
	Self-driving reduces fatalities a hundredfold.	Add	Reasoning
Car accidents are the top cause of death for teenagers.		Delete	Evidence

Table 1: A section of a fully annotated essay: first, sentence pairs are aligned between the old and new drafts to indicate which sentences correspond to one another (left two columns). Then, revision operations can be determined by whether a sentence was added, deleted or modified (third column). Third, each added/deleted/modified sentence pair is labeled for the purpose behind that revision (fourth column).

citing a new source of evidence or changing the thesis statement) and revisions that do not change meaning (such as fixing spelling errors or replacing a word with a synonym). Neither student performed any meaning-changing revisions. So, the revision assistant would see that Student A made N non-meaning-changing revisions and Student B made up to N non-meaning-changing revisions. To the revision assistant, it may appear as if both students performed the same types of revisions; the only difference is that Student A made many more revisions. However, Student A can likely expect a smaller grade change than can Student B because Student B's ideas are now more clearly presented and understood whereas Student A's revisions may pass nearly unnoticed to a reader of the old and new drafts. As such, the counts of revision purposes was misleading in predicting grade change. However, embedding cosine similarity could capture the difference: the similarity of Student A's old and new drafts is very close to 1 (identical) while the similarity between Student B's old and new drafts would be noticeably lower (less similar). So, a revision assistant using a similarity would be able to capture the greater change in Student B's essay.

This present work explores using a tf-idf embedding cosine similarity measure to quantify the relationship between the similarity of essay drafts and grade change between drafts. For comparison, the work also presents the relationship between numbers of revision purposes and grade change. Further, these relationships are analyzed at each subsequent step of annotation shown at Table 1 - first where no annotation has been performed and the essay drafts are both just raw strings of text (revision purposes are not yet available at this stage, so only the similarity measure is performed here), then with the first annotation step of aligned sentences (where the similarity measure and the number, but not the purposes, of revisions is available), then with revision purposes labeled using either a simple but commonly-used two-class schema[5] or a finer-grained multi-class schema[16] presented in section 3.3.

In situations where both the similarity measure and revision counts are available, the correlation between similarity measure and grade change when controlling for revision counts is also considered. This information is provided because prior related work ([16, 14], discussed in next section) has demonstrated revision counts to be useful when assessing an essay's revisions for grade change. Thus, it is desirable to determine if the similarity measure provides additional information on top of revision counts when evaluating grade change, or if the similarity measure and revision counts provide the same information and controlling for one renders the other insignificant. Findings ultimately suggest that tf-idf embedding cosine similarity does well at capturing deeper, meaning-altering changes to the essay and overall essay grade change while revision purposes capture how the essay changes with respect to obeying conventions such as vocabulary choice and grammar.

Section 2 contains an overview of related work. Section 3 explains the dataset and tools employed in the analysis, which is in Section 4. A discussion of findings is in Section 5 and the conclusions are listed in Section 6.

2. RELATED WORK

Effective writing can be a difficult skill to teach, so there has been significant effort towards analyzing methods and developing tools to help in this goal. One center of attention has been on discovering what makes a revision *effective*[4, 14, 2, 12] and developing tools and writing assistants to *help* students revise effectively[13, 17]. Much of the work focuses on revisions to Wikipedia[9, 8] and students' argumentative essays[16, 7].

There are three broad categories of information that can be used to describe a revision. First, there is the type of revision operation that was performed, such as adding, deleting or modifying a sentence. [12] analyzed which revision operations are associated with essay improvement, as measured by features such as lexical diversity and amount of first, second or third person.

Second, there is the purpose or intent behind the revision that was performed, such as to fix a grammar mistake or alter the meaning of a claim. There are many revision schemata in common use. As a minimum amount of granularity when classifying revision purposes, [5] propose a binary schema of text-base revisions. One class is "Content" revisions that alter text meaning, such as changing the evidence cited or the ideas in the thesis statement, and the

other class is "Surface" revisions that do not alter meaning, such as changing citation format or fixing a spelling error in the thesis statement. Both of these categories can be further broken down into fine-grained categories. For instance, [4] used a set of 13 revision purposes on Wikipedia articles, to compare the revision strategies between articles that are featured on the Wikipedia homepage and those that have not been featured.

A third way of describing a revision is features that can be automatically gathered about the revision such as edit distance between the old and new draft or the change in word count between drafts. For instance, [3] use an array of features including Named Entity Recognition, word-level edit distance and number of inserted or deleted characters to build a classifier to distinguish between factual and fluency revisions. These sorts of features are used in classifiers that aim to predict revision purposes. For instance, [18] uses revision operation and statistics such as edit distance, word count and presence of grammatical and spelling errors as features to a revision purpose classifier for student argumentative essays. [14] used many features including number of informal words, change in character counts and punctuation to build a revision purpose classifier for Wikipedia edits.

While there has been application of similarity measures to revision studies, such as Latent Semantic Analysis (LSA)[12], Levenshtein Distance[2] and Kullback-Leibler divergence[15], these similarity measures are often used along the way to predict other information about revisions. For instance, [2] used Levenshtein Distance as a feature for predicting revision purpose. [12] used LSA to analyze which revision operations are associated with larger change in essay similarity.

The contribution of this present work is twofold. First, it considers tf-idf cosine similarity as a method of quantifying grade change between drafts, rather than using the similarity measure to predict revision purposes and subsequently using revision purposes to quantify grade change. See Figure 1 and Figure 2 for a visualization of this change. Searching through the literature, an application of similarity measures in this way does not appear to have been done before. This specific similarity measure was chosen because embedding cosine similarity is a very simple, not state of the art, method of calculating similarity compared to methods like LSA or Kullback-Leibler divergence. So, if cosine similarity is shown to be a better predictor for grade change than revision purposes, when using a high-quality corpus ([1], introduced in Section 3) that has been human-annotated for sentence alignments and revision purposes, then this pattern may be able to also hold when using more advanced similarity measures or lower-quality corpora such as automatically annotated ones with less reliable revision purpose labels. Further, only tf-idf is presented in this work, but the same experiments were performed with sent2vec[10] and BERT[11] embeddings, which yielded similar results. Because of the similar results, only the tf-idf embedding is reported in this paper for simplicity and to demonstrate that even this simpler embedding is able to perform favorably compared to revision purposes. The second contribution of this work is that it explores the information revealed by tf-idf cosine similarity as subsequently more additional information becomes available: first on its own, then with essay drafts that have had their sentences aligned, then with coarse-grained revision purposes labeled, then finally with fine-grained revision purposes labeled.

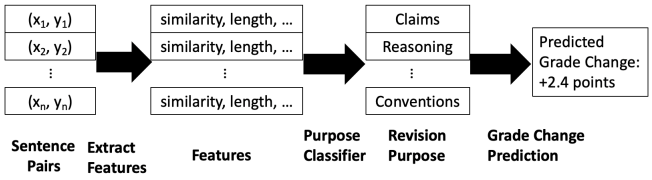

Figure 1: The method in prior works: using similarity measures and other features to predict revision purposes, then using the number of revision purposes to predict grade change.

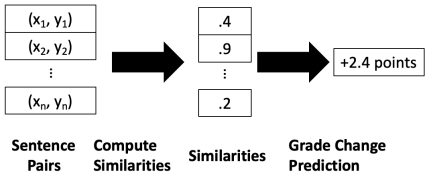

Figure 2: One of the methods explored in this work: using similarity measures to predict grade change.

3. METHODOLOGY

3.1 Data

The ArgRewrite V.2 corpus[1] was used as the dataset for this work. 86 recruited graduate and undergraduate university students wrote argumentative essays about self-driving cars and revised their essays two times for a total of three drafts. Each draft has been graded as described in Section 3.2, sentence purposes aligned between drafts and revision purposes annotated as described in section 3.3. For the present analysis, only revisions between drafts 1 and 2 were used. An example section of one essay and its annotations is given in Table 1.

3.2 Grades

Human graders assigned scores to each essay draft using a 10-category rubric, with each category being evaluated on a scale from 1 (poor) to 4 (excellent), for minimum possible score of 10 and maximum possible score of 40. The names of these categories and the description for the 4-point/excellent category are provided in Table 2. All drafts were scored separately by two annotators and the Quadratic Weighted Kappa was 0.537[1].

Two additional grade categories beyond those in Table 2 are also added in the analysis: Average, which is the average score across all the rubric categories for some draft, and Total, which is the total score for a draft on the 40-point scale. For information on how grades changed between the drafts, see Table 3.

3.3 Revision Purposes

The fine-grained revision purposes used by the ArgRewrite V.2 corpus are able to capture more of the variation in revisions by breaking the coarse categories of Surface and Content as described by [5] each into nine smaller sub-categories,

Category	Requirement for Maximum (4) Points
Response to Prompt	The author responds to all parts of the prompt and the entire essay is focused on the prompt.
Thesis	The author provided a clear, nuanced and original statement that acted as a specific stance for or against self-driving cars.
Claims	The author makes multiple, distinct claims that are clear, and align with both their thesis statement and the given reading. They fully support the author's argument.
Evidence	The author provides specific and convincing evidence for each claim,and most evidence is given through detailed examples, direct quotations, or detailed examples from the provided reading. The source of the evidence is credible and acknowledged/cited where appropriate
Reasoning	All claims are supported with clear reasoning that shows thoughtful, elaborated analysis.
Organization	The essay has an introduction, body and conclusion and a logical sequence of ideas. Each paragraph makes a distinct claim.
Rebuttal	The essay explains a different point of view and elaborates why it is not convincing or correct.
Precision	Throughout the essay, word choices are specific and convey precise meanings(e.g., "Self-driving cars are dangerous because the technology is still not advanced enough to address the ethical decisions drivers must make.")
Fluency	All sentences are clear because of correct and appropriate word choices and sentence structure.
Conventions	The author makes few or no grammatical or spelling errors throughout their piece, and the meaning is clear.

Table 2: Grading rubric categories for the ArgRewrite V.2 corpus [1] and the description to earn maximum points in that category.

with Surface containing Conventions, Organization and Fluency, and Content containing Precision, Claims, Evidence, Reasoning, Rebuttal and General Content [17]. Each sentence pair with a revision operation of Add, Delete or Modify is assigned one purpose. See Table 5 for details of the fine-grained categories and Table 4 for information on the number of occurrences of each revision purpose. Revision

	Draft 1	Draft 2	Better	Same
Prompt	3.12 (0.62)	3.20 (0.59)	9	71
Thesis	2.38 (0.76)	2.52 (0.69)	23	54
Claims	2.44 (0.75)	2.58 (0.71)	18	65
Evidence	2.01 (0.68)	2.16 (0.72)	19	64
Reasoning	2.12 (0.64)	2.23 (0.63)	19	63
Org.	2.47 (0.71)	2.69 (0.70)	25	55
Rebuttal	2.03 (0.93)	2.29 (0.94)	26	57
Precision	2.40 (0.60)	2.46 (0.61)	12	71
Fluency	2.32 (0.68)	2.38 (0.65)	14	68
Convetnns	2.32 (0.60)	2.35 (0.55)	13	67
Average	2.36 (0.43)	2.48 (0.43)	59	18
Total	23.64 (4.32)	24.84 (4.31)	59	18

Table 3: Statistics about the 86 essays' grades: Average score in drafts 1 and 2 (standard deviation in parenthesis), number that improved in the given category and that stayed the same in the given category. When the number of "Better" and "Same" essays do not sum to 86, the remaining essays received lower scores on draft 2 than draft 1.

	Occurrences	Essays with at least one
Any	1658	86
Surface	631	86
Content	1027	86
Conventns	124	49
Org.	52	44
Fluency	455	79
Precision	51	32
Claims	166	58
Evidence	115	44
Reasoning	290	65
Gen. Con.	381	70

Table 4: Number of occurrences of each type of revision in the left column: any revision at all, coarse-grained revisions and fine-grained revisions. Not all essays contain all fine-grained revision categories, so the number of essays that contain each revision is shown in the second column.

purpose annotations of this corpus were performed by three annotators, with a Fleiss' kappa[6] of 0.65 on a sample of five essays that all annotators labeled independently prior to labeling disjoint sets of the remaining essays.

The revision purpose categories and grading rubric categories bear some resemblance to one another, with most revision purposes corresponding to one of the rubric categories. In particular, the Conventions, Organization, Fluency, Precision, Claims, Evidence, Reasoning and Rebuttal revision purposes each align with the rubric categories of the same names. The definitions of these revision purposes and the criteria for the rubric categories correspond such that an Evidence-purpose revision, for instance, is more likely to cause a change in the Evidence rubric category than any other rubric category. The remaining rubric categories of Response to Prompt and Thesis do not clearly align with any revision purpose, although they can be thought of as aligning most closely with Claims revisions. By the same logic that the revision purposes can be sorted into Surface

versus Content categories, so too can the rubric categories: rubric categories that align with Surface or Content revision purposes respectively can be considered Surface-related or Content-related rubric categories. Meanwhile, the General Content revision purpose does not align with any certain rubric category, but can be thought of as revisions that may correspond to any of the Content-related rubric categories such as Thesis or Evidence. While the rubric categories are never grouped together in the tests performed (e.g. considering "Content" and "Surface" grade categories), these concepts are still useful for analysis and interpreting the results.

	Purpose	Definition
Surface	Conventions	fix grammar/spelling mistakes
	Organization	switch the order of sentences
	Fluency	make the essay read better
Content	Precision	slight changes alter essay meaning
	Claims	change the thesis of the essay
	Evidence	change evidence support for thesis
	Reasoning	change the reasoning for the thesis
	Rebuttal	change the rebuttal of the thesis
	Gen. Content	other types of content revisions

Table 5: Definitions of revision purposes [16].

3.4 Tf-idf Cosine Similarity Measure

The tf-idf cosine similarity of two strings x and y in a corpus containing W unique words is calculated in two steps: first, vector embedding representations e_x and e_y are calculated for each string. The number of elements in each vector equals W. To make the embedding for a string s, the number of occurrences in the string s (Term Frequency, TF) are counted for each word in the corpus. This results in a length-W vector TF_s where the i-th element of TF_s contains the number of occurrences in string s of the i-th unique word of the corpus. Next, the number of documents (Document Frequency) that each word occurs in are counted and used inverse-proportionally to calculate the Inverse Document Frequency (IDF) resulting in another length-W vector IDF. Lastly, the tf-idf vector e_s to represent string s equals the element-wise multiplication of TF_s and IDF.

After obtaining the embeddings e_x and e_y for two strings x and y, the cosine similarity between the vectors is computed. Cosine similarity is a measure of how the directions/angles of vector e_x and e_y compare to one another, with 1 being identical, 0 being orthogonal (signifying no correlation between the meanings of the two strings) and -1 being opposite (antonymous strings such as "up" and "down"). Because tf-idf embeddings never contain negative numbers, tf-idf cosine similarity actually varies between 0 and 1 and does not consider antonymy. Cosine similarity is calculated as

$$similarity = cos(\theta) = \frac{e_x \cdot e_y}{|e_x||e_y|} = \frac{e_x \cdot e_y}{W^2}.$$

In this sense, longer-length revisions are captured by similarity measures as having lower similarity, because they change more words and presumably the overall meaning of the sentence. In this work, embeddings and similarities were calculated using the Gensim package [10].

4. ANALYSIS

Each dataset and each educational data mining project has different requirements and different resources available. For instance, the numbers of revision purposes have been demonstrated in prior works to be useful in assessing revision's impacts on grade change. However, aligning sentence pairs between old and new drafts and subsequently labelling revision purposes requires time and effort that may not be possible for all datasets and all projects. Similarity measures are able to be used with any essays dataset, without needing any annotation, and capture slightly different information than do numbers of revision purposes.

As such, for each subsequent degree that annotation is performed, this analysis explores what information is provided by tf-idf cosine similarity or revision purposes when assessing rubric category grade change. The different levels of annotation are: (1) raw non-annotated drafts, (2) sentences aligned, (3) revisions labeled with coarse-grained Surface/ Content revision purposes and (4) revisions labeled with fine-grained 9-class revision purposes. In all cases, $N = 86$ for these tests because there are 86 essays, each associated with one score in each of the rubric categories. Afterwards, some patterns in which rubric categories are best assessed at which level of dataset annotation will be highlighted.

4.1 Non-annotated Data

Document-level similarity is the simplest and least expensive method of relating grade change to essay similarity because it does not require sentences to be aligned. Each draft is treated as one string, an embedding is created to correspond to each of the two strings and the similarity between the embeddings is calculated. Each draft is treated as one unit and there is no need to align the sentences between the drafts or identify sentences' revision purposes. A potential use-case for document-level similarity might be a revision assistant for students that gives real-time feedback as they work, because this data can be computed quickly, accurately and automatically as a student works on their essay.

First, the similarity between old and new draft are found for each of the 86 essays, using the process described in Section 3.4. Then, the Pearson correlation was computed between similarity and grade change in each of the rubric categories (correlation between similarity and Claims grade change, between similarity and Average rubric grade change, etc.). This method shows a Pearson correlation of $r = -0.2189$ with Average rubric category grade change ($p = 0.0428$). As such, essay drafts with embeddings that are more similar to each other tend to experience less grade change. For specific rubric categories, only Reasoning grade change is also significantly correlated ($r = -0.2480, p = 0.0213$) with document-level similarity. See Table 6 for all significant correlations with rubric categories.

	Essay-Level Similarity	
Grade	r	p
Reasoning	-0.2480	0.0213
Average	-0.2189	0.0428

Table 6: Pearson correlation between essay tf-idf embedding cosine similarity and rubric category grade change. Only significant results shown.

Non-annotated data is where cosine similarity measures show the greatest advantage over revision counts, because revision counts cannot even be used without some amount of annotation. Further, even if annotations are available, this essay-level similarity is still a useful way to quickly assess and summarize the revision of an essay as demonstrated by the significant level of correlation between essay-level similarity and Average rubric grade change.

4.2 Sentence-aligned Data

Aligning sentence pairs enables a further degree of analysis between similarity measures and grades where the unit of comparison is at the sentence level. Without using a measure of similarity, it is possible to simply examine the correlation between the total number of revised sentences and rubric grade changes. The total number of revisions between two drafts is significantly correlated with grade change in the precision ($r = 0.2495, p = 0.0205$) and fluency ($r = 0.2172, p = 0.0446$) rubric categories. See Table 7 for all significant correlations with rubric categories.

	Number of Revisions	
Grade	r	p
Precision	0.2495	0.0205
Fluency	0.2172	0.0446

Table 7: Pearson correlation between number of revisions and rubric category grade change. Only significant ($p \leq 0.05$) pairs shown.

These results can be contrasted with the findings of the similarity measure in the previous subsection. While the essay-level similarity measure is significantly correlated with Reasoning grade change and Average category grade change, the number of revision counts is significantly correlated with Precision and the Surface-related category of Fluency. As such, the revision purpose may be more useful in applications where Surface-level information is desired, whereas the essay-level cosine similarity may be better for gaining an overall picture of how the essay has changed.

Now that sentence alignments are available, the similarity measure can also be calculated by finding the similarity between each old-new sentence pair and then averaging over all sentence pairs. An old draft-new draft sentence pair (x, y) that is not modified between drafts ($x == y$) has a similarity of 1, sentences deleted from the first draft ($y == \varnothing$) or added to the second draft ($x == \varnothing$) have a similarity of 0 (signifying no correlation between the empty string and the added or deleted sentence) and similarity of a modified sentence ($x \neq y \neq \varnothing$) may be calculated by creating embeddings for each sentence version and then finding the cosine similarity between the embeddings. Using Gensim tf-idf embeddings[10], this method demonstrates a Pearson correlation of $r = -0.2692$ ($p = 0.0122$) with Average rubric grade change, which is slightly stronger and more significant than the essay-level correlation without using aligned sentence pairs. Sentence-level similarity is also significantly correlated with grade change in the Reasoning ($r = -0.2962, p = 0.0056$) and Claim ($r = -0.2441, p = 0.0235$) rubric categories. See Table 8 for all significant correlations.

Controlling for the total number of revisions yields a correlation between average sentence tf-idf cosine similarity and Average rubric grade change of $r = -0.2720$ ($p = 0.0113$). Further, when controlling for total number of revisions, there are significant correlations between tf-idf cosine similarity and Claim rubric grade change ($r = -0.2261, p = 0.0363$), Reasoning grade change ($r = -0.2723, p = 0.0112$) and Rebuttal grade change ($r = -0.3352, p = 0.0016$) even though the Rebuttal category was not significantly correlated with sentence-level similarity when not controlling for number of revisions. Conventions grade change is almost significantly correlated, with a correlation of $r = 0.2120$ and significance of $p = 0.0501$. Interestingly, the correlation with Conventions (as well as a few other, non-significant categories), is positive. This positive correlation means that greater cosine similarity (meaning more similar essay drafts) is correlated with more grade improvement in these rubric categories. See Table 8 for all significant correlations. Perhaps students who focus more on revising for Conventions see greater cosine similarity and greater grade increases in the Conventions category, at the cost of less improvement in other categories like Claims and Reasoning. As a result, these other categories have negative correlations with similarity. This pattern of Conventions-focused revision would be similar to Student A in the example of Section 1.

	Sentence Pair Similarity		Sentence Pair Similarity, Control Revision Count	
Grade	r	p	r	p
Claim	-0.2441	0.0235	-0.2261	0.0363
Reasoning	-0.2962	0.0056	-0.2723	0.0112
Rebuttal			-0.3352	0.0016
Average	-0.2692	0.0122	-0.2720	0.0113

Table 8: Correlation between rubric grade change and average sentence tf-idf cosine similarity, with (on left) or without (on right) controlling for number of revisions. Insignificant correlations not shown.

At this level, there seems to be no overlap between the rubric categories significantly correlated with number of revisions (Precision and Fluency) and the rubric categories significantly correlated with average sentence-level similarity (Reasoning and Claims). Controlling for the number of revisions allows similarities also to be significantly correlated with Rebuttal grade change. Further, similarity measures at this level are significantly correlated with Average rubric category grade change. As such, while both revision counts and similarity measures provide useful information when sentences have been aligned across drafts, revision counts may be better suited to getting a general, overall preview of the degree of essay change and degree of grade change. Revision counts are more significantly correlated with Surface-related categories such as Conventions. Meanwhile, similarity measures are more significantly correlated with Content-level categories such as Reasoning.

4.3 Coarse-grained Revision Purposes

When annotating the dataset, the next step after aligning sentences between drafts can be annotating the revisions with coarse revision purposes, which means distinguishing between Surface and Content. A potential use-case for this level of annotation might be to applications where there is

more time available to do more dataset annotation, but a higher degree of accuracy is desired than when annotating for fine-grained revision purposes.

Examining the number of coarse-grained revision purposes per essay, the number of Surface revisions is not significantly correlated with grade change in any rubric category. The number of Content revisions has a Pearson correlation of $r = 0.2588$ ($p = 0.0161$) with Reasoning rubric grade change, $r = 0.2512$ ($p = 0.0917$) with Precision grade change and $r = 0.2235$ ($p = 0.0386$) with fluency grade change. No other rubric categories show a significant correlation of $p \leq 0.05$. This is the first level of annotation where revision counts are significantly correlated with a Content-related rubric category (Reasoning).

	Content Revision Count	
Grade	r	p
Reasoning	0.2588	0.0161
Precision	0.2512	0.0197
Fluency	0.2235	0.0386

Table 9: Pearson correlation between rubric grade change and number of content-purpose revisions. Insignificant correlations not shown.

Reasoning grade change is significantly correlated with the number of Content revisions, but is not correlated with the total number of revisions as shown in Table 7. Meanwhile, the number of Surface revisions is not significantly correlated with any grade change, including in rubric categories that correspond to fine-grained surface-level revision purposes (i.e. Conventions, Fluency and Organization). The number of Content revisions is significantly correlated with Fluency, despite Fluency being a surface-level revision purpose.

New patterns arise when using cosine similarity but controlling for the number of Surface or Content revisions. When controlling for just for the number of Content revisions, the correlation between cosine similarity (calculated as the average of sentence pairs' embeddings' similarities) and grade change is $r = -0.2543$ ($p = 0.0181$). When controlling for the count of Surface revisions, the correlation is $r = -0.2665$ ($p = 0.0131$). Controlling for the number of Surface revisions also results in significant correlation of similarity to Claims grade and Reasoning grade. Controlling for the number of Content revisions gives significant correlation of similarity to Thesis grade and Rebuttal grade. See Table 10 for all significant correlations between cosine similarity and rubric categories when controlling for number of Surface or Content revisions.

At this level, similarity measures are significantly correlated with several content-oriented rubric categories such as thesis and reason, as well as average rubric grade. Meanwhile, the number of content revisions is significantly correlated with two content-oriented categories (reasoning and precision) and one surface-oriented category (fluency). The number of surface revisions is not significantly related to any rubric category grade change. As such, the similarity measure seems to be doing the best at capturing deeper, content-level changes and the number of content revisions captures a general view of changes. However, the similarity measure is also significantly correlated with average category grade change, unlike the number of surface, content or total revisions. As a result, at this level of annotation, the best overall understanding of a student's revision pattern as a predictive measure for grade change would be to calculate the number of content revisions and the average sentence-level cosine similarity, with or without controlling for number of content revisions.

	Control for Surface Count		Control for Content Count	
Grade	r	p	r	p
Thesis			-0.2606	0.0154
Claim	-0.2369	0.0281		
Reasoning	-0.2877	0.0072		
Rebuttal			-0.4141	0.0001
Average	-0.2665	0.0131	-0.2543	0.0181

Table 10: Pearson correlation between average sentence cosine similarity and rubric categories when controlling for number of surface or content revisions. Insignificant correlations not shown.

Rubric Category, Revision Purpose	r	p
Claim, Claims	0.3195	0.0027
Evidence, Evidence	0.2578	0.0166
Reasoning, General Content	0.3226	0.0024
Average, Evidence	0.2642	0.0140

Table 11: Correlation between the similarity of sentences with a specific revision purpose and rubric grade change, controlling for number of sentences revised for this purpose. Only significant ($p \leq 0.05$) correlations shown.

4.4 Fine-grained Revision Purposes

Labeling revisions for fine-grained revision purposes, although more intensive and giving lower annotator agreement, provides more information when correlating with grade change. For counts of fine-grained revision purposes, the results for all significant ($p \leq 0.05$) correlations with rubric grade changes are summarized in Table 11. Due to the great number of combinations of 12 rubric categories and 9 revision purposes, only correlations between corresponding categories (for instance, between Organization grade category and Organization revision purpose) and correlations with Average and Total rubric grade change were performed.

Only three revision purposes are significantly correlated with grade change in their relevant rubric categories: Claims ($r = 0.3195, p = 0.0027$), Evidence ($r = 0.2578, p = 0.0166$) and General Content with Reasoning rubric category ($r = 0.3226$, $p = 0.0024$). Further, the Evidence revision purpose is significantly correlated with Average rubric grade change ($r = 0.2642, p = 0.0140$). This is the first significant correlation between any variety of revision count and Average rubric grade change.

The next test that was performed involved considering just the correlation between a specific rubric category and cosine similarity between sentence pairs revised for a specific fine-

grained revision purpose, when controlling for the number of occurrences of that revision purpose. For instance, when considering the subset of essays that contain at least one sentence revised for Fluency, the correlation between Fluency rubric grade change and average tf-idf cosine similarity of old-new sentence pairs revised for Fluency is $r = -0.1988$ ($p = 0.0790$) when controlling for the number of Fluency revisions. However, no correlations were found to be significant in this complicated test. As such, this very high degree of detail focusing on specific rubric categories and related revision purposes is not useful. Perhaps the test is so focused on minuscule portions of essays that it loses sight of the context in which the revision is situated. For instance, adding a Reasoning sentence to a very long essay may make a very small difference to the overall comprehensibility of the essay's reasoning, while adding a Reasoning sentence to a very short essay may help to bridge a hole in the reasoning that was caused by the essay's brevity and failure to give detailed explanations. As such, this individual sentence may contribute less to the first essay's grade change than it does to the second essay's grade change. Another consideration about this test is that not all revision purposes occur in all essays (see Table 4 for details), so the sample size of essays included in these tests is often smaller than the full 86 essays included in all other tests.

At this level of annotation where the essays have been annotated for fine-grained revision purposes, the most useful indicator of grade change appears to be the counts of revision purposes. However, the rubric categories (Reasoning, Precision and Fluency) that are significantly correlated with the counts of revision purposes are already significantly correlated with other tests that do not require fine-grained revision purposes. As a result, when looking to use revisions to predict grade change, the additional effort to annotate fine-grained revision purposes instead of coarse-grained ones may not be worthwhile.

5. DISCUSSION

Generally, the similarity measure is more significantly correlated with Content-related rubric category grade change and with Average rubric grade change, while numbers of revisions are more significantly related with Surface-related rubric category grade change. This is the case in Sections 4.1, 4.2 and 4.3. Significant correlations present at high degrees of annotation, such as those between fine-grained revision purposes and rubric categories in Section 4.4, tend to be present at coarser degrees of annotation as well. As such, when aiming to predict grade change, the most productive level of annotation may be aligned sentences like in Section 4.2 or coarse-grained revision purposes as in Section 4.3. These levels of analysis capture significant correlations between revision and a wide range of the different rubric categories for essays in this dataset, with revision counts capturing Surface-related rubric category grade change and similarity measures capturing Content-related and overall average change.

A caveat associated with this conclusion is that the ArgRewrite corpus is entirely human-annotated and all sentence alignments and revision purpose labels are the gold standard between two annotators. However, many datasets and applications do not have detailed human annotation available. As such, the accuracy of this dataset's sentence alignments and revision purpose annotations is at the upper bound of possible accuracy. The correlations between revision counts/purposes and rubric categories are, therefore, an upper bound that may not be possible in datasets that have been annotated by a classifier. This caveat also holds for the similarity measure in cases where the measure is being calculated on data that has some sort of annotation, such as sentence alignments for average sentence-level similarity.

A second caveat is that this analysis does not take into account that some students scored higher than others on the first draft. For instance, a student who nearly receives a perfect score on the first draft has little room to improve whereas a student with a low score initially has ample improvement opportunities. Potential ways to combat this issue would be using some variety of corrected learning gain score or finding the correlation between similarity score and second draft score after controlling for first draft score. Lastly, it may be worthwhile to apply a post-hoc control such as Bonferroni correction to the significance tests.

6. CONCLUSION

This work indicates that tf-idf embedding cosine similarity captures overall essay grade change and essay revisions that lead to rubric grade change in Content-related categories, while revision purposes capture change in more Surface-oriented rubric categories. Future work is needed to demonstrate whether this pattern extends to additional datasets, particularly datasets where the sentence alignment and revision purposes have been automatically labeled by a classifier. Further, the dataset in this analysis contained only argumentative essays about self-driving cars, so further work would need to examine these findings for datasets with other writing styles and topics.

7. ACKNOWLEDGMENTS

Special thanks to the entire ArgRewrite group, and particularly Tazin Afrin for her detailed comments on a draft of this paper. This work is supported by National Science Foundation (NSF) grant 1735752 to the University of Pittsburgh. The opinions expressed are those of the authors and do not represent the views of the Institute.

8. REFERENCES

[1] T. Afrin, O. Kashefi, C. Olshefski, D. Litman, R. Hwa, and A. Godley. Effective interfaces for student-driven revision sessions for argumentative writing. *ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.

[2] T. Afrin and D. Litman. Annotation and classification of sentence-level revision improvement. *arXiv preprint arXiv:1909.05309*, 2019.

[3] A. Bronner and C. Monz. User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 356–366, 2012.

[4] J. Daxenberger and I. Gurevych. A corpus-based study of edit categories in featured and non-featured wikipedia articles. In *Proceedings of COLING 2012*, pages 711–726, 2012.

[5] L. Faigley and S. Witte. Analyzing revision. *College composition and communication*, 32(4):400–414, 1981.

[6] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[7] J. Lee, C. Y. Yeung, A. Zeldes, M. Reznicek, A. Lüdeling, and J. Webster. Cityu corpus of essay drafts of english language learners: a corpus of textual revision in second language writing. *Language Resources and Evaluation*, 49(3):659–683, 2015.

[8] A. Max and G. Wisniewski. Mining naturally-occurring corrections and paraphrases from wikipedia's revision history. In *LREC*, 2010.

[9] M. Potthast. Crowdsourcing a wikipedia vandalism corpus. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 789–790, 2010.

[10] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[11] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[12] R. D. Roscoe, M. E. Jacovina, L. K. Allen, A. C. Johnson, and D. S. McNamara. Toward revision-sensitive feedback in automated writing evaluation. In *EDM*, pages 628–629. Citeseer, 2016.

[13] A. Shibani, S. Knight, and S. Buckingham Shum. Understanding students' revisions in writing: from word counts to the revision graph. Technical report, Technical report, Connected Intelligence Centre, University of Technology Sydney, 2018.

[14] D. Yang, A. Halfaker, R. Kraut, and E. Hovy. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010, 2017.

[15] T. Zesch, M. Wojatzki, and D. Scholten-Akoun. Task-independent features for automated essay grading. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 224–232, 2015.

[16] F. Zhang, H. B. Hashemi, R. Hwa, and D. Litman. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578, 2017.

[17] F. Zhang, R. Hwa, D. Litman, and H. B. Hashemi. Argrewrite: A web-based revision assistant for argumentative writings. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Demonstrations*, pages 37–41, 2016.

[18] F. Zhang and D. Litman. Annotation and classification of argumentative writing revisions. *Grantee Submission*, 2015.