Annotation and Classification of Argumentative Writing Revisions

Fan Zhang
University of Pittsburgh
Pittsburgh, PA, 15260
zhangfan@cs.pitt.edu

Diane Litman
University of Pittsburgh
Pittsburgh, PA, 15260
litman@cs.pitt.edu

Abstract

This paper explores the annotation and classification of students' revision behaviors in argumentative writing. A sentence-level revision schema is proposed to capture why and how students make revisions. Based on the proposed schema, a small corpus of student essays and revisions was annotated. Studies show that manual annotation is reliable with the schema and the annotated information helpful for revision analysis. Furthermore, features and methods are explored for the automatic classification of revisions. Intrinsic evaluations demonstrate promising performance in high-level revision classification (surface vs. text-based). Extrinsic evaluations demonstrate that our method for automatic revision classification can be used to predict a writer's improvement.

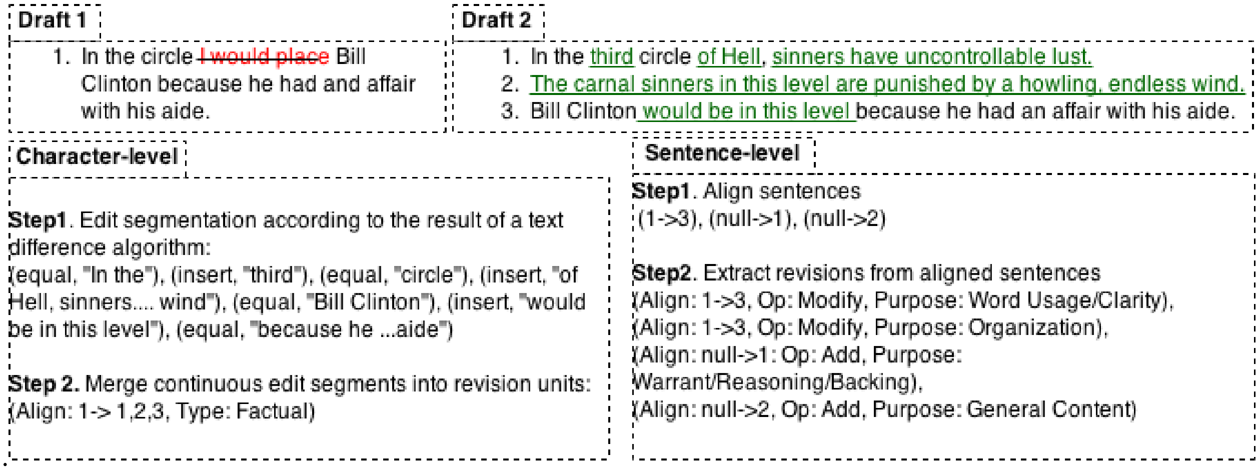
1 Introduction

Rewriting is considered as an important factor of successful writing. Research shows that expert writers revise in ways different from inexperienced writers (Faigley and Witte, 1981). Recognizing the importance of rewriting, more and more efforts are being made to understand and utilize revisions. There are rewriting suggestions made by instructors (Wells et al., 2013), studies modeling revisions for error correction (Xue and Hwa, 2010; Mizumoto et al., 2011) and tools aiming to help students with rewriting (Elireview, 2014; Lightside, 2014).

While there is increasing interest in the improvement of writers' rewriting skills, there is still a lack of study on the details of revisions. First, to find out what has been changed (defined as **revision extraction** in this paper), a typical approach is to extract and analyze revisions at the word/phrase level based on edits extracted with character-level text comparison (Bronner and Monz, 2012; Daxenberger and Gurevych, 2012). The semantic information of sentences is not considered in the character-level text comparison, which can lead to errors and loss of information in revision extraction. Second, the differentiation of different types of revisions (defined as **revision categorization**) is typically not fine-grained. A common categorization is a binary classification of revisions according to whether the information of the essay is changed or not (e.g. text-based vs. surface as defined by Faigley and Witte (1981)). This categorization ignores potentially important differences between revisions under the same high-level category. For example, changing the evidence of a claim and changing the reasoning of a claim are both considered as text-based changes. Usually changing the evidence makes a paper more grounded, while changing the reasoning helps with the paper's readability. This could indicate different levels of improvement to the original paper. Finally, for the automatic differentiation of revisions (defined as **revision classification**), while there are works on the classification of Wikipedia revisions (Adler et al., 2011; Bronner and Monz, 2012; Daxenberger and Gurevych, 2013), there is a lack of work on revision classification in other datasets such as student writings. It is not clear whether current features and methods can still be adapted or new features and methods are required.

To address the issues above, this paper makes

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 133–143,
Denver, Colorado, June 4, 2015. ©2015 Association for Computational Linguistics



Draft 1

1. In the circle ~~I would place~~ Bill Clinton because he had and affair with his aide.

Draft 2

1. In the third circle of Hell, sinners have uncontrollable lust.
2. The carnal sinners in this level are punished by a howling, endless wind.
3. Bill Clinton would be in this level because he had an affair with his aide.

Character-level

Step1. Edit segmentation according to the result of a text difference algorithm:
(equal, "In the"), (insert, "third"), (equal, "circle"), (insert, "of Hell, sinners.... wind"), (equal, "Bill Clinton"), (insert, "would be in this level"), (equal, "because he ...aide")

Step 2. Merge continuous edit segments into revision units:
(Align: 1-> 1,2,3, Type: Factual)

Sentence-level

Step1. Align sentences
(1->3), (null->1), (null->2)

Step2. Extract revisions from aligned sentences
(Align: 1->3, Op: Modify, Purpose: Word Usage/Clarity),
(Align: 1->3, Op: Modify, Purpose: Organization),
(Align: null->1: Op: Add, Purpose: Warrant/Reasoning/Backing),
(Align: null->2, Op: Add, Purpose: General Content)

Figure 1: In the example, words in sentence 1 of Draft 1 are rephrased and reordered to sentence 3 of Draft 2. Sentences 1 and 2 in Draft 2 are newly added. Our method first marks 1 and 3 as aligned and the other two sentences of Draft 2 as newly added based on semantic similarity of sentences. The purposes and operations are then marked on the aligned pairs. In contrast, previous work extracts differences between drafts at the character level to get edit segments. The revision is extracted as a set of sentences covering the contiguous edit segments. Sentence 1 in Draft 1 is wrongly marked as being modified to 1, 2, 3 in Draft 2 because character-level text comparison could not identify the semantic similarity between sentences.

the following efforts. First, we propose that it is better to extract revisions at a level higher than the character level, and in particular, explore the sentence-level. This avoids the misalignment errors of character-level text comparisons. Finer-grained studies can still be done on the sentence-level revisions extracted, such as fluency prediction (Chae and Nenkova, 2009), error correction (Cahill et al., 2013; Xue and Hwa, 2014), statement strength identification (Tan and Lee, 2014), etc. Second, we propose a sentence-level revision schema for argumentative writing, a common form of writing in education. In the schema, categories are defined for describing an author's revision operations and revision purposes. The revision operations can be directly decided according to the results of sentence alignment, while revision purposes can be reliably manually annotated. We also do a corpus study to demonstrate the utility of sentence-level revisions for revision analysis. Finally, we adapt features from Wikipedia revision classification work and explore new features for our classification task, which differs from prior work with respect to both the revision classes to be predicted and the sentence-level revision extraction method. Our models are able to distinguish whether the revisions are changing the content or not. For fine-grained classification, our models also demonstrate good performance for some categories. Beyond the classification task, we also investigate the pipelining of revision extraction and classification. Results of an extrinsic evaluation show that the automatically extracted and classified revisions can be used for writing improvement prediction.

2 Related work

Revision extraction To extract the revisions for revision analysis, a widely chosen strategy uses character-based text comparison algorithms first and then builds revision units on the differences extracted (Bronner and Monz, 2012; Daxenberger and Gurevych, 2013). While theoretically revisions extracted with this method can be more precise than sentence-level extractions, it could suffer from the misalignments of revised content due to character-level text comparison algorithms. For example, when a sentence is rephrased, a character-level text comparison algorithm is likely to make alignment errors as it could not recognize semantic similarity. As educational research has suggested that revision analysis can be done at the sentence level (Faigley and Witte, 1981), we propose to extract revisions at the sentence level based on semantic sentence alignment instead. Figure 1 provides an example comparing revisions annotated in our work to revisions extracted in prior work (Bronner and Monz, 2012). Our work identifies the fact that the student added new information to the essay and modified the organization of old sentences. The previous work, however, extracts all the modifications as one unit and cannot distinguish the different kinds of revisions inside the unit. Our method is similar to Lee and Webster's method (Lee and Webster, 2012), where a sentence-level revision corpus is built from college students' ESL writings. However, their corpus only includes the comments of the teachers and does not have every revision annotated.

Revision categorization In an early educational work from Faigley and Witte (1981), revisions are categorized to *text-based change* and *surface change* based on whether they changed the information of the essay or not. A similar categorization (factual vs. fluency) was chosen by Bronner and Monz (2012) for classifying Wikipedia edits. However, many differences could not be captured with such coarse grained categorizations. In other works on Wikipedia revisions, finer categorizations of revisions were thus proposed: vandalism, paraphrase, markup, spelling/grammar, reference, information, template, file etc. (Pfeil et al., 2006; Jones, 2008; Liu and Ram, 2009; Daxenberger and Gurevych, 2012). Corpus studies were conducted to analyze the relationship between revisions and the quality of Wikipedia papers based on the categorizations. Unfortunately, their categories are customized for Wikipedia revisions and could not easily be applied to educational revisions such as ours. In our work, we provide a fine-grained revision categorization designed for argumentative writing, a common form of writing in education, and conduct a corpus study to analyze the relationship between our revision categories and paper improvement.

Revision classification Features and methods are widely explored for Wikipedia revision classifications (Adler et al., 2011; Mola-Velasco, 2011; Bronner and Monz, 2012; Daxenberger and Gurevych, 2013; Ferschke et al., 2013). Classification tasks include binary classification for coarse categories (e.g. factual vs. fluency) and multi-class classification for fine-grained categories (e.g. 21 categories defined by Daxenberger and Gurevych (2013)). Results show that the binary classifications on Wikipedia data achieve a promising result. Classification of finer-grained categories is more difficult and the difficulty varies across different categories. In this paper we explore whether the features used in Wikipedia revision classification can be adapted to the classification of different categories of revisions in our work. We also utilize features from research on argument mining and discourse parsing (Burstein et al., 2003; Burstein and Marcu, 2003; Sporleder and Lascarides, 2008; Falakmasir et al., 2014; Braud and Denis, 2014) and evaluate revision classification both intrinsically and extrinsically. Finally, we explore end-to-end revision processing by combining automatic revision extraction and categorization via automatic classification in a pipelined manner.

3 Sentence-level revision extraction and categorization

This section describes our work for sentence-level revision extraction and revision categorization. A corpus study demonstrates the use of the sentence-level revision annotations for revision analysis.

3.1 Revision extraction

As stated in the previous section, our method takes semantic information into consideration when extracting revisions and uses the sentence as the basic semantic unit; besides the utility of sentence revisions for educational analysis (Faigley and Witte, 1981; Lee and Webster, 2012), automatic sentence segmentation is quite accurate. Essays are split into sentences first, then sentences across the essays are aligned based on semantic similarity.[1] An added sentence or a deleted sentence is treated as aligned to *null* as in Figure 1. The aligned pairs where the sentences in the pair are not identical are extracted as revisions. For the automatic alignment of sentences,

[1] We plan to also explore revision extraction at the clause level in the future. Our approach can be adapted to the clause level by segmenting the clauses first and aligning the segmented clauses after. A potential benefit is that clauses are often the basic units of discourse structures, so extracting clause revisions will allow the direct use of discourse parser outputs (Feng and Hirst, 2014; Lin et al., 2014). However, potential problems are that clauses contain less information for alignment decisions and clause segmentation is noisier.

we used the algorithm in our prior work (Zhang and Litman, 2014) which considers both sentence similarity (calculated using TF*IDF score) and the global context of sentences.

3.2 Revision schema definition

As shown in Figure 2, two dimensions are considered in the definition of the revision schema: the author's behavior (**revision operation**) and the reason for the author's behavior (**revision purpose**).

Revision operations include three categories: *Add*, *Delete*, *Modify*. The operations are decided automatically after sentences get aligned. For example, in Figure 1 where Sentence 3 in Draft 2 is aligned to sentence 1 in Draft 1, the revision operation is decided as *Modify*. The other two sentences are aligned to null, so the revision operations of these alignments are both decided as *Add*.

The definitions of revision purposes come from several works in argumentative writing and discourse analysis. *Claims/Ideas*, *Warrant/Reasoning/Backing*, *Rebuttal/Reservation*, *Evidence* come from *Claim*, *Rebuttal*, *Warrant*, *Backing*, *Grounds* in Toulmin's model (Kneupper, 1978). *General Content* comes from *Introductory material* in the essay-based discourse categorization of Burstein et al. (2003). The rest come from the categories within the surface changes of Faigley and Witte (1981). Examples of all categories are shown in Table 1. These categories can further be mapped to surface and text-based changes defined by Faigley and Witte (1981), as shown in Figure 2.

Note that while our categorization comes from the categorization of argumentative writing elements, a key difference is that our categorization focuses on revisions. For example, while an evidence revision must be related to the evidence element of the essay, the reverse is not necessarily true. The modifications on an evidence sentence could be just a correction of spelling errors rather than an evidence revision.

3.3 Data annotation

Our data consists of the first draft (Draft 1) and second draft (Draft 2) of papers written by high school students taking English writing courses; papers were revised after receiving and generating peer feedback. Two assignments (from different teachers) have been annotated so far. Corpus C1 comes from an AP-level course, contains papers about Dante's Inferno and contains drafts from 47 students, with 1262 sentence revisions. A Draft 1 paper contains 38 sentences on average and a Draft 2 paper contains 53. Examples from this corpus are shown in Table 1. After data was collected, a score from 0 to 5 was assigned to each draft by experts (for research prior to our study). The score was based on the student's performance including whether the student stated the ideas clearly, had a clear paper organization, provided good evidence, chose the correct wording and followed writing conventions. The class's average score improved from 3.17 to 3.74 after revision. Corpus C2 (not AP) contains papers about the poverty issues of the modern reservation and contains drafts from 38 students with 495 revisions; expert ratings are not available. Papers in C2 are shorter than C1; a Draft 1 paper contains 19 sentences on average and a Draft 2 paper contains 26.

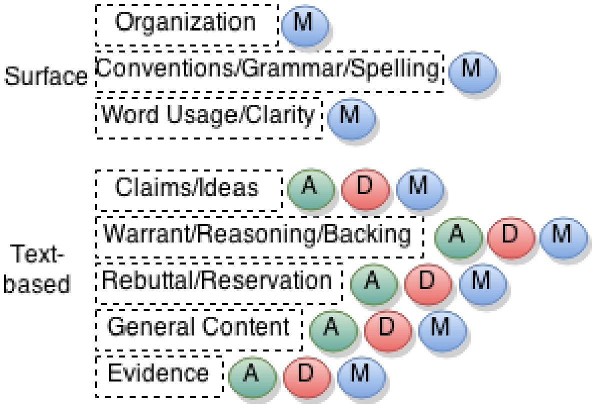


Figure 2: For the revision purpose, 8 categories are defined. These categories can be mapped to surface and text-based changes. Revision operations include *Add*, *Delete*, *Modify* (A, D, M in the figure). Only text-based changes have *Add* and *Delete* operations.

Two steps were involved in the revision scheme annotation of these corpora. In the first step, sentences between the two drafts were aligned based on semantic similarity. The kappa was 0.794 for the sentence alignment on C1. Two annotators discussed about the disagreements and one annotator's work was decided to be better and chosen as the gold standard after discussion. The sentence alignment on C2 is done by one annotator after his annotation and discussion of the sentence alignment on C1. In

Codes	*Claims/Ideas*: change of the position or claim being argued for *Conventions/Grammar/Spelling*: changes to fix spelling or grammar errors, misusage of punctuation or to follow the organizational conventions of academic writing
Example	Draft 1: (1, “**Saddam Hussein and Osama Bin Laden come** to mind when mentioning wrathful people”) Draft 2: (1, “**Fidel Castro comes** to mind when mentioning wrathful people”)
Revisions	(1->1, Modify, “claims/ideas”), (1->1, Modify, “conventions/grammar/spelling”)
Codes	*Evidence*: change of facts, theorems or citations for supporting claims/ideas *Rebuttal/Reservation*: change of development of content that rebut current claim/ideas
Example	Draft 1: (1, “In this circle I would place Fidel.”) Draft 2: (1, “In the circle I would place Fidel”), (2, “**He was annoyed with the existence of the United States and used his army to force them out of his country**”), (3, “**Although Fidel claimed that this is for his peoples’ interest, it could not change the fact that he is a wrathful person.**”)
Revisions	(null->2, “Add”, “Evidence”), (null->3, “Add”, “Rebuttal/Reservation”)
Codes	*Word-usage/Clarity*: change of words or phrases for better representation of ideas *Organization*: changes to help the author get a better flow of the paper *Warrant/Reasoning/Backing*: change of principle or reasoning of the claim *General Content*: change of content that do not directly support or rebut claims/ideas
Example	As in Figure 1

Table 1: Examples of different revision purposes. Note that in the second example the alignment is not extracted as a revision when the sentences are identical.

the second step, revision purposes were annotated on the aligned sentence pairs. Each aligned sentence pair could have multiple revision purposes (although rare in the annotation of our current corpus). The full papers were also provided to the annotators for context information. The kappa score for the revision purpose annotation is shown in Table 2, which demonstrates that our revision purposes could be annotated reliably by humans. Again one annotator's annotation is chosen as the gold standard after discussion. Distribution of different revision purposes is shown in Tables 3 and 4.

3.4 Corpus study

To demonstrate the utility of our sentence-level revision annotations for revision analysis, we conducted a corpus study analyzing relations between the number of each revision type in our schema and student writing improvement based on the expert paper scores available for C1. In particular, the number of revisions of different categories are counted for each student. Pearson correlation between the number of revisions and the students' Draft 2 scores is calculated. Given that the student's Draft 1 and Draft 2 scores are significantly correlated ($p < 0.001$, $R = 0.632$), we controlled for the effect of Draft 1 score by regressing it out of the correlation.[2] We expect surface changes to have smaller impact than text-based changes as Faigley and Witte (1981) found that advanced writers make more text-based changes comparing to inexperienced writers.

As shown by the first row in Table 5, the overall number of revisions is significantly correlated with students' writing improvement. However, when we compare revisions using Faigley and Witte's binary categorization, only the number of text-based revisions is significantly correlated. Within the text-based revisions, only *Claims/Ideas*, *Warrant/Reasoning/Backing* and *Evidence* are significantly correlated. These findings demonstrate that revisions at different levels of granularity have different relationships to students' writing success,

[2] Such partial correlations are one common way to measure learning *gain* in the tutoring literature, e.g. (Baker et al., 2004).

Revision Purpose	**Kappa (C1)**	**Kappa (C2)**
Surface		
Organization	1	1
Conventions	0.74	0.87
Word-usage	1	1
Text-based		
Claim	0.76	0.89
Warrant	0.78	0.85
Rebuttal	1	1
General Content	0.76	0.80
Evidence	1	1

Table 2: Agreement of annotation on each category.

which suggests that our schema is capturing salient characteristics of writing improvement.

While correlational, these results also suggest the potential utility of educational technologies based on fine-grained revision analysis. For teachers, summaries of the revision purposes in a particular paper (e.g. "The author added more reasoning sentences to his old claim, and changed the evidence used to support the claim.") or across the papers of multiple students (e.g. "90% of the class made only surface revisions") might provide useful information for prioritizing feedback. Fine-grained revision analysis might also be used to provide student feedback directly in an intelligent tutoring system.

4 Revision classification

In the previous section we described our revision schema and demonstrated the utility of it. This section investigates the feasibility of automatic revision analysis. We first explore classification assuming we have revisions extracted with perfect sentence alignment. After that we combine revision extraction and revision classification in a pipelined manner.

4.1 Features

As shown in Figure 3, besides using unigram features as a baseline, our features are organized into *Location*, *Textual*, and *Language* groups following prior work (Adler et al., 2011; Bronner and Monz, 2012; Daxenberger and Gurevych, 2013).

Baseline: unigram features. Similarly to Daxenberger and Gurevych (2012), we choose the count of unigram features as a baseline. Two types of uni-

Rev Purpose	**# Add**	**# Delete**	**#Modify**
Total	800	96	366
Surface	0	0	297
Organization	0	0	35
Conventions	0	0	84
Word-usage	0	0	178
Text-based	800	96	69
Claim	80	23	8
Warrant	335	40	14
Rebuttal	1	0	0
General	289	23	42
Evidence	95	10	5

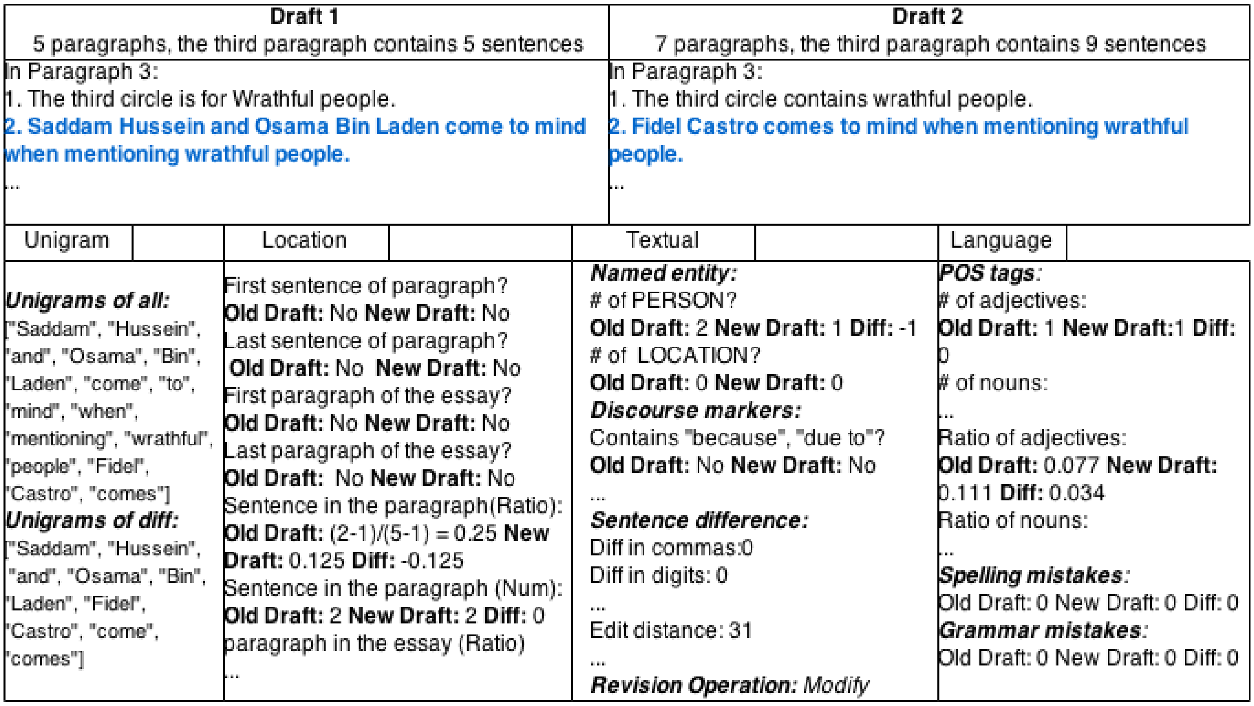
Table 3: Distribution of revisions of corpus C1.

Rev Purpose	**# Add**	**# Delete**	**#Modify**
Total	280	53	162
Surface	0	0	141
Organization	0	0	1
Conventions	0	0	29
Word-usage	0	0	111
Text-based	280	53	21
Claim	42	12	4
Warrant	153	23	10
Rebuttal	0	0	0
General	60	13	6
Evidence	25	5	1

Table 4: Distribution of revisions of corpus C2.

Revision Purpose	**R**	**p**
# All revisions (N = 1262)	0.516	<**0.001**
# Surface revisions	0.137	0.363
# Organization	0.201	0.180
# Conventions	-0.041	0.778
# Word-usage/Clarity	0.135	0.371
# Text-based revisions	0.546	<**0.001**
# Claim/Ideas	0.472	**0.001**
# Warrant	0.462	**0.001**
# General	0.259	0.083
# Evidence	0.415	**0.004**

Table 5: Partial correlation between number of revisions and Draft 2 score on corpus C1 (partial correlation regresses out Draft 1 score); rebuttal is not evaluated as there is only 1 occurrence.



Draft 1 5 paragraphs, the third paragraph contains 5 sentences	**Draft 2** 7 paragraphs, the third paragraph contains 9 sentences
In Paragraph 3: 1. The third circle is for Wrathful people. **2. Saddam Hussein and Osama Bin Laden come to mind when mentioning wrathful people.** ...	In Paragraph 3: 1. The third circle contains wrathful people. **2. Fidel Castro comes to mind when mentioning wrathful people.** ...

Unigram	Location	Textual	Language
Unigrams of all: ["Saddam", "Hussein", "and", "Osama", "Bin", "Laden", "come", "to", "mind", "when", "mentioning", "wrathful", "people", "Fidel", "Castro", "comes"] ***Unigrams of diff:*** ["Saddam", "Hussein", "and", "Osama", "Bin", "Laden", "Fidel", "Castro", "come", "comes"]	First sentence of paragraph? **Old Draft:** No **New Draft:** No Last sentence of paragraph? **Old Draft:** No **New Draft:** No First paragraph of the essay? **Old Draft:** No **New Draft:** No Last paragraph of the essay? **Old Draft:** No **New Draft:** No Sentence in the paragraph(Ratio): **Old Draft:** (2-1)/(5-1) = 0.25 **New Draft:** 0.125 **Diff:** -0.125 Sentence in the paragraph (Num): **Old Draft:** 2 **New Draft:** 2 **Diff:** 0 paragraph in the essay (Ratio) ...	***Named entity:*** # of PERSON? **Old Draft:** 2 **New Draft:** 1 **Diff:** -1 # of LOCATION? **Old Draft:** 0 **New Draft:** 0 ***Discourse markers:*** Contains "because", "due to"? **Old Draft:** No **New Draft:** No ... ***Sentence difference:*** Diff in commas:0 Diff in digits: 0 ... Edit distance: 31 ... ***Revision Operation:*** *Modify*	***POS tags:*** # of adjectives: **Old Draft:** 1 **New Draft:**1 **Diff:** 0 # of nouns: ... Ratio of adjectives: **Old Draft:** 0.077 **New Draft:** 0.111 **Diff:** 0.034 Ratio of nouns: ... ***Spelling mistakes:*** Old Draft: 0 New Draft: 0 Diff: 0 ***Grammar mistakes:*** Old Draft: 0 New Draft: 0 Diff: 0

.

Figure 3: An example of features extracted for the aligned sentence pair (2->2).

grams are explored. The first includes unigrams extracted from all the sentences in an aligned pair. The second includes unigrams extracted from the differences of sentences in a pair.

Location group. As Falakmasir et al. (2014) have shown, the positional features are helpful for identifying thesis and conclusion statements. Features used include the location of the sentence and the location of paragraph .[3]

Textual group. A sentence containing a specific person's name is more likely to be an example for a claim; sentences containing "because" are more likely to be a sentence of reasoning; a sentence generated by text-based revisions is possibly more different from the original sentence compared to a sentence generated by surface revisions. These intuitions are operationalized using several feature groups: *Named entity features*[4] (also used in Bronner and Monz (2012)'s Wikipedia revision classification task), *Discourse marker features* (used by Burstein et al. (2003) for discourse structure identification), *Sentence difference features* and *Revision operation* (similar features are used by Daxenberger and Gurevych (2013)).

Language group. Different types of sentences can have different distributions in POS tags (Daxenberger and Gurevych, 2013). The difference in the number of spelling/grammar mistakes[5] is a possible indicator as Conventions/Grammar/Spelling revisions probably decrease the number of mistakes.

4.2 Experiments

Experiment 1: Surface vs. text-based As the corpus study in Section 3 shows that only text-based revisions predict writing improvement, our first experiment is to check whether we can distinguish between the surface and text-based categories. The classification is done on all the non-identical aligned sentence pairs with *Modify* operations[6]. We choose 10-fold (student) cross-validation for our experi-

[3] Since Add and Delete operations have only one sentence in the aligned pair, the value of the empty sentence is set to 0.

[4] Stanford parser (Klein and Manning, 2003) is used in named entity recognition.

[5] The spelling/grammar mistakes are detected using the languagetool toolkit (https://www.languagetool.org/).

[6] *Add* and *Delete* pairs are removed from this task as only text-based changes have *Add* and *Delete* operations.

N = 366	Precision	Recall	F-score
Majority	32.68	50.00	37.12
Unigram	45.53	49.90	46.69
All features	**62.89**∗	**58.19**∗	**55.30**∗

Table 6: Experiment 1 on corpus C1 (Surface vs. Text-based): average unweighted precision, recall, F-score from 10-fold cross-validation; ∗ indicates significantly better than majority and unigram.

N = 162	Precision	Recall	F-score
Majority	31.57	40.00	33.89
Unigram	50.91	50.40	51.79
All features	**56.11**∗	**55.03**∗	**54.49**∗

Table 7: Experiment 1 on corpus C2.

ment. Random Forest of the Weka toolkit (Hall et al., 2009) is chosen as the classifier. Considering the data imbalance problem, the training data is sampled with a cost matrix decided according to the distribution of categories in training data in each round. All features are used except *Revision operation* (since only *Modify* revisions are in this experiment).

Experiment 2: Binary classification for each revision purpose category In this experiment, we test whether the system could identify if revisions of each specific category exist in the aligned sentence pair or not. The same experimental setting for surface vs. text-based classification is applied.

Experiment 3: Pipelined revision extraction and classification In this experiment, revision extraction and Experiment 1 are combined together as a pipelined approach[7]. The output of sentence alignment is used as the input of the classification task. The accuracy of sentence alignment is 0.9177 on C1 and 0.9112 on C2. The predicted *Add* and *Delete* revisions are directly classified as text-based changes. Features are used as in Experiment 1.

4.3 Evaluation

In the intrinsic evaluation, we compare different feature groups' importance. Paired t-tests are utilized to compare whether there are significant differences in performance. Performance is measured using unweighted F-score. In the extrinsic evaluation, we repeat the corpus study from Section 3 using the predicted counts of revision. If the results in the intrinsic evaluation are solid, we expect that a similar conclusion could be drawn with the results from either predicted or manually annotated revisions.

Intrinsic evaluation Tables 6 and 7 present the results of the classification between surface and text-based changes on corpora C1 and C2. Results show that for both corpora, our learned models significantly beat majority and unigram baselines for all of unweighted precision, recall and F-score; the F-score for both corpora is approximately 55.

Tables 8 and 9 show the classification results for the fine-grained categories. Our results are not significantly better than the unigram baseline on *Evidence* of C1, C2 and *Claim* of C2. While the poor performance on *Evidence* might be due to the skewed class distribution, our model also performs better on *Conventions* where there are not many instances. For the categories where our model performs significantly better than the baselines, we see that the location features are the best features to add to unigrams for the text-based changes (significantly better than baselines except *Evidence*), while the language and textual features are better for surface changes. We also see that using all features does not always lead to better results, probably due to over fitting. Replicating experiments in two corpora also demonstrates that our schema and features can be applied across essays with different topics (Dante vs. poverty) written in different types of courses (advanced placement or not) with similar results.

For the intrinsic evaluation of our pipelined approach (**Experiment 3**), as the revisions extracted are not exactly the same as the revisions annotated by humans, we only report the unweighted precision and unweighted recall here; C1 (p: 40.25, r: 45.05) and C2 (p: 48.08, r: 54.30). Paired t-test shows that the results significantly drop compared to Tables 6 and 7 because of the errors made in revision extraction, although still outperform the majority baseline.

Extrinsic evaluation According to Table 10 , the conclusions drawn from the predicted revisions and annotated revisions are similar (Table 5). Text-based changes are significantly correlated with writing improvement, while surface changes are not. We can also see that the coefficient of the predicted text-

[7] We leave pipelined fine-grained classification to the future.

N = 1261	**Text-based**				**Surface**		
Experiments	**Claim**	**Warrant**	**General**	**Evidence**	**Org.**	**Word**	**Conventions**
Majority	39.24	32.25	29.38	27.47	25.49	27.75	31.23
Unigram	65.64	63.24	69.21	60.40	49.23	62.07	56.05
All features	66.20	**70.76**∗	72.65∗	60.57	**54.01**∗	73.79∗	70.95∗
Textual+unigram	**71.54**∗	68.13∗	70.76	59.73	52.62∗	**75.92**∗	**71.98**∗
Language+unigram	67.76∗	66.27∗	69.23	59.81	49.21	65.01∗	69.62∗
Location+unigram	69.90∗	67.78∗	**72.94**∗	59.14	49.25	62.40	66.85∗

Table 8: Experiment 2 on corpus C1: average unweighted F-score from 10-fold cross-validation; ∗ indicates significantly better than majority and unigram baselines. *Rebuttal* is removed as it only occurred once.

N = 494	**Text-based**				**Surface**	
Experiments	**Claim**	**Warrant**	**General**	**Evidence**	**Word**	**Conventions**
Majority	24.89	32.05	28.21	27.02	13.00	32.67
Unigram	54.34	64.06	55.00	56.99	49.56	60.09
All features	50.22	67.50∗	56.50	53.90	56.07∗	77.78∗
Textual+unigram	52.19	65.79	55.74	56.08	54.19∗	76.08∗
Language+unigram	50.54	**68.24**∗	56.42	56.15	**58.83**∗	**78.92**∗
Location+unigram	53.20	66.45∗	**58.08**∗	52.57	51.55	75.39∗

Table 9: Experiment 2 on corpus C2; *Organization* is removed as it only occurred once.

Predicted purposes	**R**	**p**
#All revisions (N = 1262)	0.516	<**0.001**
#Surface revisions	0.175	0.245
#Text-based revisions	0.553	<**0.001**
Pipeline predicted purposes	**R**	**p**
#All (predicted N = 1356)	0.509	<**0.001**
#Surface revisions	0.230	0.124
#Text-based revisions	0.542	<**0.001**

Table 10: Partial correlation between number of predicted revisions and Draft 2 score on corpus C1. (Upper: Experiment 1, Lower: Experiment 3)

based change correlation is close to the coefficient of the manually annotated results.

5 Conclusion and current directions

This paper contributes to the study of revisions for argumentative writing. A revision schema is defined for revision categorization. Two corpora are annotated based on the schema. The agreement study demonstrates that the categories defined can be reliably annotated by humans. Study of the annotated corpus demonstrates the utility of the annotation for revision analysis. For automatic revision classification, our system can beat the unigram baseline in the classification of higher level categories (surface vs. text-based). However, the difficulty increases for fine-grained category classification. Results show that different feature groups are required for different purpose classifications. Results of extrinsic evaluations show that the automatically analyzed revisions can be used for writer improvement prediction.

In the future, we plan to annotate revisions from different student levels (college-level, graduate level, etc.) as our current annotations lack full coverage of all revision purposes (e.g., “Rebuttal/Reservation” rarely occurs in our high school corpora). We also plan to annotate data from other educational genres (e.g. technical reports, science papers, etc.) to see if the schema generalizes, and to explore more category-specific features to improve the fine-grained classification results. In the longer-term, we plan to apply our revision predictions in a summarization or learning analytics systems for teachers or a tutoring system for students.

Acknowledgments

We would like to thank Wencan Luo, Christian Schunn, Amanda Godley, Kevin Ashley and other members of the SWoRD group for their helpful feedback and all the anonymous reviewers for their suggestions.

This research is funded by IES Award R305A120370 and NSF Award #1416980.

References

B. Thomas Adler, Luca De Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, and Andrew G. West. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'11, pages 277–288, Berlin, Heidelberg. Springer-Verlag.

Ryan Shaun Baker, Albert T Corbett, Kenneth R Koedinger, and Angela Z Wagner. 2004. Off-task behavior in the cognitive tutor classroom: when students game the system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 383–390.

Chloé Braud and Pascal Denis. 2014. Combining natural and artificial examples to improve implicit discourse relation identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1694–1705, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Amit Bronner and Christof Monz. 2012. User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 356–366, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jill Burstein and Daniel Marcu. 2003. A machine learning approach for identification of thesis and conclusion statements in student essays. *Computers and the Humanities*, 37(4):pp. 455–467.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the write stuff: Automatic identification of discourse structure in student essays. *Intelligent Systems, IEEE*, 18(1):32–39.

Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. 2013. Robust systems for preposition error correction using wikipedia revisions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 507–517, Atlanta, Georgia, June. Association for Computational Linguistics.

Jieun Chae and Ani Nenkova. 2009. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 139–147. Association for Computational Linguistics.

Johannes Daxenberger and Iryna Gurevych. 2012. A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *Proceedings of COLING 2012*, pages 711–726, Mumbai, India, December. The COLING 2012 Organizing Committee.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically classifying edit categories in Wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 578–589, Seattle, Washington, USA, October. Association for Computational Linguistics.

Elireview. 2014. Eli review. `http://elireview.com/support/`. [Online; accessed 11-17-2014].

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College composition and communication*, pages 400–414.

Mohammad Hassan Falakmasir, Kevin D Ashley, Christian D Schunn, and Diane J Litman. 2014. Identifying thesis and conclusion statements in student essays to scaffold peer review. In *Intelligent Tutoring Systems*, pages 254–259. Springer.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of The 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, USA, June*.

Oliver Ferschke, Johannes Daxenberger, and Iryna Gurevych. 2013. A survey of nlp methods and resources for analyzing the collaborative writing process in wikipedia. In *The Peoples Web Meets NLP*, pages 121–160. Springer.

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

John Jones. 2008. Patterns of revision in online writing a study of wikipedia's featured articles. *Written Communication*, 25(2):262–289.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Charles W Kneupper. 1978. Teaching argument: An introduction to the toulmin model. *College Composition and Communication*, pages 237–241.

John Lee and Jonathan Webster. 2012. A corpus of textual revisions in second language writing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 248–252. Association for Computational Linguistics.

Lightside. 2014. lightside revision assistant. `http://lightsidelabs.com/ra/`. [Online; accessed 11-17-2014].

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, pages 1–34.

Jun Liu and Sudha Ram. 2009. Who does what: Collaboration patterns in the wikipedia and their impact on data quality. In *19th Workshop on Information Technologies and Systems*, pages 175–180.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning sns for automated japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155. Asian Federation of Natural Language Processing.

Santiago M Mola-Velasco. 2011. Wikipedia vandalism detection. In *Proceedings of the 20th international conference companion on World wide web*, pages 391–396. ACM.

Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. 2006. Cultural differences in collaborative authoring of wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Nat. Lang. Eng.*, 14(3):369–416, July.

Chenhao Tan and Lillian Lee. 2014. A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 403–408, Baltimore, Maryland, June. Association for Computational Linguistics.

Jaclyn M. Wells, Morgan Sousa, Mia Martini, and Allen Brizee. 2013. Steps for revising your paper. `http://owl.english.purdue.edu/owl/resource/561/05`.

Huichao Xue and Rebecca Hwa. 2010. Syntax-driven machine translation as a model of esl revision. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1373–1381. Association for Computational Linguistics.

Huichao Xue and Rebecca Hwa. 2014. Improved correction detection in revised esl sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–604, Baltimore, Maryland, June. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2014. Sentence-level rewriting detection. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 149–154, Baltimore, Maryland, June. Association for Computational Linguistics.