Lang Resources & Evaluation (2022) 56:881–915
https://doi.org/10.1007/s10579-021-09567-z

ORIGINAL PAPER

ArgRewrite V.2: an annotated argumentative revisions corpus

Omid Kashefi[1] · **Tazin Afrin**[1] · **Meghan Dale**[2] · **Christopher Olshefski**[2] · **Amanda Godley**[2] · **Diane Litman**[1,2] · **Rebecca Hwa**[1]

Accepted: 4 November 2021 / Published online: 13 January 2022

Abstract Analyzing how humans revise their writings is an interesting research question, not only from an educational perspective but also in terms of artificial intelligence. Better understanding of this process could facilitate many NLP applications, from intelligent tutoring systems to supportive and collaborative writing environments. Developing these applications, however, requires revision corpora, which are not widely available. In this work, we present ArgRewrite V.2, a corpus of annotated argumentative revisions, collected from two cycles of revisions to argumentative essays about self-driving cars. Annotations are provided at different levels of purpose granularity (coarse and fine) and scope (sentential and subsentential). In addition, the corpus includes the revision goal given to each writer, essay scores, annotation verification, pre- and post-study surveys collected from participants as meta-data. The variety of revision unit scope and purpose granularity levels in ArgRewrite, along with the inclusion of new types of meta-data, can make it a useful resource for research and applications that involve revision analysis. We demonstrate some potential applications of ArgRewrite V.2 in the development of automatic revision purpose predictors, as a training source and benchmark.

This material is based upon work supported by the National Science Foundation (NSF) under Grant #1735752.

✉ Omid Kashefi
kashefi@cs.pitt.edu

[1] School of Computing and Information, University of Pittsburgh, 6135 Sennott Square, 210 South Bouquet Street, Pittsburgh, PA 15260, USA

[2] Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, PA 15260, USA

Keywords Revision · Subsentential revision · Revision purpose · Annotated corpus · Argumentative essay

1 Introduction

Writing is an essential human activity for organizing and understanding complex ideas (Westby et al., 2010), and revising is an important part of that process. The process of adding, deleting, rearranging or modifying words, phrases, sentences or paragraphs in one's writing not only improves the writing, but can support more complex thinking about the subject at hand (Allal et al., 2004; Flower & Hayes, 1981). Revisions to both wording and the main ideas of an essay are important but in different ways. Revisions to wording are important for improving the fluency and correctness of a text, whereas revisions to the main ideas help writers rethink and refine their argument or purpose (Beason, 1993). Researchers working on revisions have shown, however, that inexperienced writers typically focus only on changes to wording, not on the organization and main ideas (i.e., content) of an essay (Cho & MacArthur, 2010). Best practices in writing instruction, therefore, emphasize the importance of teacher and peer feedback to support effective content-level revisions (Magnifico et al., 2014).

Computational researchers have recently taken interest in writers' revision processes for both scientific reasons as well as practical ones. Scientifically, modeling how humans learn to present complex ideas has long been an active research area in artificial intelligence. Practically, a natural language processing (NLP) system that can model the revising process has many applications from intelligent tutoring systems (Jacovina & McNamara, 2016; Merrill et al., 1992; Roscoe & McNamara, 2013) to supportive writing environments (Zhang et al., 2016).

Developing these applications requires revision corpora, but only a limited set of them are available. Some extant corpora have focused on Wikipedia revisions (Bronner & Monz, 2013; Daxenberger & Gurevych, 2012); however, those revision properties and annotations are specifically designed for Wikipedia's collaborative writing environment, which hampers their applications to different and more general rewriting and revision analysis tasks. Another corpus of writing revisions is ArgRewrite V.1, which is a small collection of single-author college-level argumentative essays and their revisions, as well as a set of manually developed sentence-level annotations of revisions properties (Zhang et al., 2017). This prior version of ArgRewrite took a first step toward the creation of a more general-purpose revision corpus. However, as a pilot study, the corpus development was limited in several ways: it did not study subsentential revision; the annotation scheme did not make enough distinction between some revision purposes; the students did not receive individualized feedback before they attempted to revise their drafts; and the students' final drafts were not scored.

Therefore, in this paper, we present *ArgRewrite V.2* corpus, which aims to alleviate these limitations in a number of ways: (1) revisions, which are in English,

are annotated at both *sentential* and *subsentential* levels; (2) the annotation scheme now includes a *precision* revision purpose for changes to the specificity of the sentences; (3) a broader range of annotators with more rigorous training have coded the revisions with their argumentative purpose; (4) the corpus is almost twice as big as its predecessor, now including 258 drafts (3 drafts from each of the 86 participants) with around 3.3K sentential and 2.5K subsentential revisions; (5) the corpus comprises the *personalized feedback* given to each student and the *scores* for each draft.

Given the inclusion of new types of meta-data, ArgRewrite may facilitate broader range of writing-related research from automatic essay scoring (Amorim et al., 2018; Burstein et al., 2013; Taghipour and Ng 2016) to argumentative revision analysis (Afrin et al., 2020; Connor & Asenavage, 1994; Zhang & Litman, 2015). The relatively larger size and in-depth revision annotation of the corpus also makes it useful for supporting the development of application such as writing error detection and correction (Dahlmeier & Ng, 2011; Tetreault et al., 2010; Xue & Hwa, 2014b), sentence simplification and compression (Berg-Kirkpatrick et al., 2011; Coster & Kauchak, 2011; Filippova et al., 2015; Turner & Charniak, 2005; Vickrey and Koller 2008), paraphrasing (Barry, 2006; Berant & Liang, 2014; Kauchak & Barzilay, 2006), precision and specificity detection (Li & Nenkova, 2015; Lugini & Litman, 2018). In this work, we demonstrate an application of ArgRewrite in developing automatic revision purpose classifiers and how its properties make an interesting case for studying classification improvement through data augmentation.

2 ArgRewrite V.2: essay collection

The design choices of a corpus will have significant impact on a corpus's usefulness and applicability. The guiding principle behind the design decisions of our revision corpus was to maintain a balance between the consideration of the inevitability of writing style idiosyncrasies and a focus on ubiquitous writing and revision phenomena that exist across writers. On the one hand, we intended to capture a wide variety of revision phenomena, expressed by a diverse population of writers; on the other hand, we needed to ensure that the revisions could be reliably annotated by trained domain experts and that the resulting annotations would be useful to the community for further analyses and application development.

In this section, we discuss the data collection methodology and the essay production process, while Sect. 3 describes the annotation process; most of our discussion focuses on drafts (Draft1, Draft2, Draft3) and annotating revisions (Rev12, Rev23); the revisions are later used in an empirical NLP study. Exploring the rich auxiliary data—including scores, expert feedback, and student questionnaires—will be left to the follow-up studies.

Figure 1, illustrates an overview of the ArgRewrite V.2 corpus collection process: a group of students produced essay drafts that were subsequently semi-automatically segmented and annotated by a group of experts. *Draft1* is the initial version of the essay evaluated by an expert, resulting in a feedback text and a score,

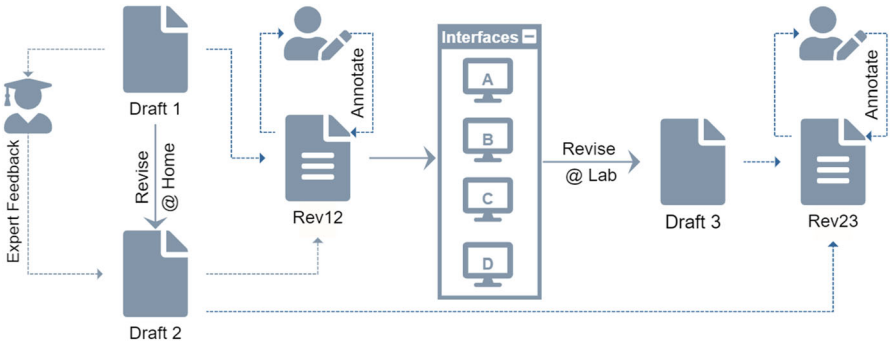

Fig. 1 ArgRewrite V.2 corpus collection process

which is not shared with the writers. The second draft (*Draft2*) was produced based on the draft and the feedback, . The first and the second draft were then semi-automatically segmented, aligned and annotated by the experts to produce the revision *Rev12*. This annotated revision was presented to the students via an interactive interface, in which the students prepared the third draft (*Draft3*) under different interfaces. As before, the drafts Draft2 and Draft3 were aligned and annotated by the experts to produce the revision *Rev23*.

2.1 Collection methodology

2.1.1 Participants

Because we wanted to collect writing samples from individuals who were relatively familiar with the basic expectations of argumentative writing, we selected a university as the pool from which to collect our data (Beach & Anson, 1988; Crammond, 1998). Recruitment materials specified that participants must be aged 18 years old and older, either native English speaker or a non-native speaker possessing sufficient English proficiency (e.g., TOEFL score 100 +). The participant recruitment process was conducted through physical and electronic flyers posted throughout the University of Pittsburgh main campus and the Carnegie Mellon University. We were able to recruit 86 participants. We expected that even within a pool of university students, we could recruit participants with a wide variety of argumentative writing skills (demographic information of participants are presented in Sect. 2.3.1).

2.1.2 Writing task

For the sake of consistency, all participants received the same instructions for a writing task which instructed them to develop an argument for or against self-driving cars that could serve as an op-ed piece in a local newspaper. In order to provide participants with comparable prior knowledge, each participant was provided with the same article about self-driving cars, organized according to the

"pros" and "cons" of self-driving car technology. Participants were instructed to use the article to first summarize the advantages and disadvantages of self-driving cars before moving into their argument. They were advised that "high quality" op-ed pieces typically maintain "a clear position on the issue" and use "supporting evidence" as well as explanations of that evidence, and also include a "counter-argument." They were told that such pieces also include clear organization, precise word choice and correct grammar. See Appendix A for the prompt text.

In contrast to this task, our prior study did not provide a common reading for participants to cite. We believe the common reading materials served as a unifying force, making the argumentative essays more comparable, so that the corpus focus is more on revisions than the writers' prior knowledge.

2.2 Collection process

We collected *three* drafts of argumentative essays from the participants in order to compare revision differences at different stages of rewriting. This process required each participant to take part in three sessions (refer to Fig. 1). In each session, participants were asked to write or revise their draft in approximately an hour. The sessions were organized as follows:

Draft1 This session took place at home. Participants were sent an email with a link to a pre-study questionnaire about their demographic background and self-reported writing background (see Appendix E for more details). Upon completion of the questionnaire, they were instructed to perform the writing task, described in Sect. 2.1.2.

Draft2 This session also took place at home. After a few days, each participant received personalized formative feedback from a human expert on their first draft (e.g., "Your essay's sequence of ideas is inconsistent, with some clear and some unclear progression." See Appendix D.2 for a more detailed example of personalized feedback message[1]). Feedback was provided via email and aligned with the writing criteria we later used to assess the quality of each draft (see Appendix D.3 for the scoring rubric). The feedback included 23 identified strengths and 23 weaknesses of the first draft. Participants were then asked to revise the first draft based on the feedback and resubmit the essay online.

Annotated Revisions I (Rev12) To begin the annotation process, we first aligned Draft1 and Draft2 at sentence level; this was performed semi-manually, using a method that considered word-level similarity between sentences of different drafts and their ordering (Zhang & Litman, 2014). Then, revised sentences were automatically segmented into subsentential revision units, using a method that merged linguistically related sequences of word-level edits (add, delete, or modify) into a subsentential change (Xue & Hwa, 2014a). Finally, a trained annotator manually coded the perceived purpose of each revision unit (at the sentential and subsentential level), following the annotation guideline (see Appendix B). These

[1] Personalized feedback is an augmentation over the previous version of the corpus, in which all participants received the same feedback (see Appendix D.1).

annotations served as the "Wizard of Oz" feedback for the participants' next session.

Draft3 In this third and final session, participants were asked to view one of the four ArgRewrite web-based interfaces and then write and revise their third draft in a designated computer lab at the University of Pittsburgh. Each participant was randomly assigned to one of the following four interfaces, which provided different types of feedback on how the participant had revised from Draft1 to Draft2[2].

- *Interface A* only the sentences without any further feedback
- *Interface B* sentence-level differences, as a surface or content revision
- *Interface C* sentence-level differences with fine-grained revision purpose
- *Interface D* subsentential differences with fine-grained revision purposes

During the lab session, all participants, except those assigned to Interface A, were asked to agree or disagree with the annotator-recognized revision purposes shown by the system (i.e., Rev12). The annotation verification information could be used, for example, for analysing the impact of the difference between the the system's recognized and the participant's actual revision intents. Then, all participants were asked to revise and submit their final draft and fill out a post-study questionnaire about their experiences (see Appendix F).

Annotated Revisions II (Rev23) After the participants submitted their Draft3, the revisions between Draft2 and Draft3 were coded by the trained annotator in the same process as annotating Rev12. Although our data collection stopped at Draft3, the participants' final round of revisions could also have been annotated and presented to the writers (via their preferred interface) to aid them with revising further drafts.

2.3 Statistics

Upon completion of collecting the essays, it is useful to review some corpus statistics; they help to assess whether the collected data matched the design goals we set out to achieve. Here, we look into the participants' diversity and textual statistics of the collected essays.

2.3.1 Participants

Table 1 shows the demographic statistics of the students who participated in the study. Aiming to study with a diverse population of university students, we ended up recruiting a mixture of undergraduate students (58%), graduate students (28%), and some non-students, mostly post-docs and lecturers (14%), where 80% were native and 20% were non-native English speakers, including 6 Chinese, 4 Hindi, 1

[2] Further considerations about the interface design and user interactions are outside the scope of this paper; we discussed them separately elsewhere (Afrin et al., 2021). Please refer to Appendix C for some screenshots of these interfaces.

Table 1 Participants' demographic statistics

Education level	Language proficiency		Overall
	Native	Non-Native	
Undergraduate	46	13	50
Graduate	11	4	24
Other	12	0	12
Overall	69	17	86

Table 2 Textual statistics of the ArgRewrite corpus

Draft	Essays	Paragraphs (avg)	Sentences (avg)	Words (avg)
1	86	405 (5)	2216 (26)	44,391 (516)
2	86	451 (5)	2461 (29)	48,832 (568)
3	86	488 (6)	2814 (33)	57,163 (665)
Overall	258	1344 (5)	7491 (29)	150,386 (582)

Vietnamese, 1 Tulu, 1 Telugu, 1 Japanese, 1 Korean, 1 Turkish, and 1 Kazakh native speakers.

2.3.2 Essays

Table 2 shows the textual statistics of the collected essays, including the number of essays, paragraphs, sentences, and words. The corpus includes 258 essays, collected through 2 cycles of revisions from the participants. In addition to having more essays in ArgRewrite V.2, an average of 29 sentences and 582 words per essay indicates that the essays are also much larger compared to the essays in the prior version of the corpus (53% more sentences and 30% more words per essay). We also observe that, when participants proceed with their revisions, essays become lengthier—as Draft2 has more words and sentences than Draft1 (on average, 29 sentences in Draft2 vs. 26 sentences in Draft1), and Draft3 more than Draft2 (on average, 33 sentences in Draft3 vs. 29 sentences in Draft2).

3 ArgRewrite V.2: annotation

This section discusses our annotation scheme design, the annotation process itself, as well as some statistics of the annotated corpus.

3.1 Annotation scheme

Our aim in developing a revision corpus is to understand why a writer makes certain revisions. Toward this end, we analyze the *p*urposes of edits—are they primarily to

improve readability or to convey different ideas? There are, however, many possible schemes to annotate these revisions. For example, we might opt to record fairly factual operations (text added, deleted, modified) or we might annotate the reasons for these operation, which may be more subjective. There is also the question of the appropriate scope of a revision; for example, if a relative clause is added to a noun, would that be considered an edit at the phrasal level, sentential level, and/or paragraph level? In developing the annotation scheme, we consider both the scope of the revision unit and the granularity of the purpose categories.

3.1.1 Scope of the revision unit

In the prior version of the corpus, a revision was defined as an original *sentence* paired with its revised version (Zhang et al., 2017). However, a sentence level revision can itself be a collection of multiple separate smaller revision units, which could have different revision purposes; little research has been done on the revision unit and scholars are uncertain whether a larger unit (e.g., at sentence or paragraph level) or a smaller unit (e.g., at phrase, or word, or character level) are more effective at supporting improvement in revision practices (Magnifico et al., 2014). Therefore, for each revision, we decided to provide the annotations at both sentential and subsentential (phrase) levels to expand the corpus application to revision studies at both levels. In the prior ArgRewrite corpus, semantically similar sentence pairs were aligned and annotated for one revision purpose, as a (sentential) revision. In the current version, however, we go further by segmenting sentential revisions into their subsentential revised units, which can be annotated independent of their corresponding sentential revisions.

Figure 2 shows some examples of revisions, annotated as both sentential and subsentential units. As an instance, in the first sentence pair, “While” is labeled as

Subsentential Annotation		**Sentential Annotation**
Original Draft	*Revised Draft*	
Self-driving vehicles pose many advantages and disadvantages . MODIFY: Convention/Spell/Grammar	ADD: Word-Usage While self-driving vehicles pose many advantages and disadvantages , I am not on the bandwagon for them at this time. ADD: Claim MODIFY: Convention/Spell/Grammar	MODIFY: Claim
The passengers in car with an omnipotent driver will not need to worry about emergency situations. DELETE: General Content Development		DELETE: General Content Development
This was recognized as being rather antisocial ... MODIFY: Word-Usage MODIFY: Convention/Spell/Grammar	This was recognized as being somewhat antisocial . MODIFY: Word-Usage MODIFY: Convention/Spell/Grammar	MODIFY: Convention/Spell/Grammar
	On the other hand, this behavior wasn't just an idle pursuit of the rich after all. ADD: Rebuttal	ADD: Rebuttal
An example for the case where the electronic communication is limited would be China. MODIFY: Evidence	An example for the case where the electronic communication is limited would be North Korea. MODIFY: Evidence	Modify: Evidence
An example for the case where the electronic communication is limited would be China.	An example for the case where the electronic communication is limited would be mainland China. ADD: Precision	ADD: Precision

Fig. 2 Example of revisions with sentential and subsentential annotations

ADD: Word-Usage in the beginning of the sentence, the component "I am not on the bandwagon. . . " is labeled as *ADD: Claim*, and the change of punctuation mark from period (".") in original sentence to comma (",") in the revised sentence is labeled as *MODIFY: Convention/Spell/Grammar*. Annotations at the sentential scope, however, would simply label the whole sentence pairs as *MODIFY: Claim*. Thus, depending on the scope of revision units, annotations can vary.

3.1.2 Granularity of the revision purpose categories

Similar to our design choices for the scope of the revision units, we also annotated the revision purposes at multiple levels of granularity. We built upon the annotation schema developed for the prior version of the corpus (Zhang & Litman, 2015), wherein revisions were annotated for their edit *operation* and argumentative *purpose*. In this version of the corpus, we have updated some of our definitions and included some new categories.

Each change made to an essay was annotated with *Add*, *Delete*, or *Modify* revision operations, according to how it related to its original version in the prior draft of the essay. These operations correspond to the addition or deletion of a whole sentence, or modification of an already existing sentence during the revision. At the subsentential level, however, a modified sentence could be revised by adding a few phrases to it, or deleting or modifying some of its phrases, so may receive different annotations based on its substantial unit changes (see Fig. 2).

This corpus provides annotations at two different grain sizes. At the coarser grain size, revision units were annotated as either *Content* revisions (i.e., changes to main ideas of the essay) or *Surface* revisions (changes to the grammar, usage, or word choice). At the finer grain size, revisions were annotated for three different subcategories of surface revisions: *Word Usage* (WRD), *Spelling and Grammar* (SPL), *Organization* (ORG) revisions. Content revisions are further categorized into six subcategories: *Claim* (CLM), *Evidence* (EVD), *Reasoning* (RSN), *Rebuttal* (RBL), *General Content Development* (GCD) revisions, or *Precision* (PRN).

The latter label, *precision*, is new in this corpus and refers to words that are edited to affect the specificity of the sentence. The purpose of such revisions is deemed to be at a content level, even though the writer may change only a few words such that the edit resembles a surface change. An example of such revisions is shown in the last row of Fig. 2: the original sentence was revised by adding the word "mainland", which makes it more specific by excluding some special administrative regions from the original claim of the sentence. For more details on the annotation guideline, see Appendix B.

3.2 Annotation process

Compared to our previous study, we decided to recruit more domain experts to annotate the corpus: an expert in argumentative analysis; an expert in AI and education; as well as a computer scientist trained in argumentative analysis. Having a larger number of annotators allowed us to study a more comprehensive inter-

Table 3 Inter-annotators agreement (Fleiss' Kappa)

Revision unit	Category granularity	
	Coarse	Fine
Sentential	0.71	0.65
Subsentential	0.92	0.78

annotator agreement and annotation quality assurance. The three domain expert annotators were trained based on the annotation guideline (see Appendix B). During the training process, annotators coded 5 revised essays with sentence-level revisions and 2 revised essays for subsentential revisions from our prior corpus, then discussed their annotation intuitions and disagreements. After the annotation training, we ran a pilot version of our new study to collect 5 revised essays, then each annotator coded these essays.

Table 3 shows the inter-annotator agreement on coding the pilot study revisions with the coarse and fine-grained revision purposes at sentential and subsentential levels, calculated as Fleiss' kappa (Fleiss, 1971). As expected, annotators had higher level of agreement in coding a coarser-grained category scheme (2 categories: surface or content) compared to a finer one (9 categories: detailed revision purposes). They also agreed more on subsentential annotations than sentential annotations. One reason could be that a subsentential change is made to serve just one revision purpose, while a sentential revision might be an amalgamate of multiple smaller changes, each made for different argumentative purpose, so merging these different purposes into one clear revision purpose might be harder to distinguish each of them (see Fig. 2).

Nevertheless, since the annotators were able to reach substantial agreement on both sentential and subsentential revisions, each *Rev12* and *Rev23* file from all interfaces was randomly assigned to the annotators, so each annotator coded about one third of the ArgRewrite corpus. Revisions from Interface B and Interface C were annotated with fine-grained categories at sentence-level and revisions from Interface A and Interface D were annotated with fine-grained categories at both sentence and subsentential levels.

3.3 Statistics

Collecting a wide variety of revision phenomena that are representative of the revision behaviour of the students is an important aspect of developing a revision corpus. Table 4 shows the distribution of annotated revision *purposes* for *sentential* and *subsentential* revision unit. The corpus contains 3238 sentential (84% more than prior version) and 2596 subsentential (new in ArgRewrite V.2) revisions annotated with both fine and coarse revision purpose category labels. Also, although all essays in the corpus were annotated at the sentence level, only the essays of participants using Interfaces A and D include annotations at the sub-sentential level.

As shown, the corpus includes a variety of surface and content revisions, however, some revisions such as choosing a better word to express an idea (word

Table 4 Revision purpose statistics of sentential and subsentential revisions

Purpose	Sentential			Subsentential		
	Rev12	Rev23	Overall (avg)	Rev12	Rev23	Overall (avg)
Word usage	453	577	1030 (6)	445	654	1099 (13)
Spell/Grammar	125	114	239 (1)	137	150	267 (3)
Organization	52	25	77 (<1)	33	2	35 (<1)
Surface	630	716	1346 (8)	615	806	1421 (17)
Claim	154	80	234 (1)	89	44	133 (2)
Reasoning	262	352	614 (4)	140	243	383 (5)
Evidence	112	88	200 (1)	46	51	97 (1)
Rebuttal	22	20	42 (<1)	15	12	27 (<1)
Precision	50	35	85 (<1)	88	59	147 (2)
GCD	397	320	717 (4)	183	205	388 (5)
Content	997	895	1892 (11)	561	614	1175 (14)
Overall (avg)	1627 (19)	1892 (22)	3238 (19)	1176 (28)	1420 (34)	2596 (31)

usage), provide reasoning for a claim (reasoning), or introducing general content to develop an argument (other), are more frequent, while changes to the organization of the essay, or rebutting an idea or changing the specificity level of the essay (precision), rarely happen in students' revisions.

Sentential revisions are more inclined toward content changes (on average, 11 content change per draft vs. 8 surface revisions per draft), while it is quite the opposite for subsentential revision (on average, 14 content change per draft vs. 17 surface revisions), which may imply that a bigger content revision could be made through, or include, some smaller surface changes. This fundamental difference between purposes of the same revisions at different unit scopes may validate that our decision to annotate revisions as both sentential and subsentential units can actually make it useful for applications analyzing different units of text, which may have different annotation requirements.

In general, students made slightly more changes when revising their second draft (on average, 22 sentential or 34 subsentential revisions in Rev23 per draft) than revising their first draft (on average, 19 sentential or 28 subsentential revisions in Rev12 per draft).

Another way to look at the revisions is from the edit *operation* perspective to see if a revision is made by adding or deleting some text, or modifying some part of the essays. Table 5 shows the distribution of edit operations for sentential and subsentential revision units. There is a small difference (54 sentences) between the number of revisions that are annotated with edit operation and those that are coded with revision purpose (Table 4). Some of the revisions were annotated with more than one revision purpose, which is violating our annotation guideline, so we discarded them from the current version of the corpus.

Table 5 Revision operation statistics of sentential and subsentential revisions

Operation	Sentential			Subsentential		
	Rev12	Rev23	Overall (avg)	Rev12	Rev23	Overall (avg)
Add	555	530	1085 (6)	370	439	809 (10)
Delete	324	174	498 (3)	245	243	488 (6)
Modify	777	932	1709 (10)	568	580	1148 (14)
Overall (avg)	1656 (19)	1636 (19)	3292 (19)	1183 (28)	1262 (30)	2445 (29)

As shown in Table 5, most of the revisions involved modifying a previously written sentence (on average, modification of 10 sentences or 14 phrases of a draft), and deletion are the less popular operation for revising essays (on average, deletion of 3 sentences or 6 phrases from a draft).

From the revision operation perspective, unlike the revision purpose annotations, different drafts were revised quite similarly at both sentential and subsentential scopes (on average, 19 edit operations at sentence-level and about 29 edit operations at phrasal-level in both Rev12 and Rev23). This observation implies that different dimensions of annotation may express a different type of information and reveal different characteristics of the revision behaviour, therefore, including different type of annotations for revisions (operation, coarse-grained purpose, fine-grained purpose) can widen the usefulness of our corpus for a more diverse set of applications.

4 Corpus availability

The ArgRewrite V.2 is available from http://argrewrite.cs.pitt.edu Participant identification information is anonymous and the corpus contains:

- *Essays* 258 raw text files of the written essays (86 of each draft).
- *Annotations* 172 excel files: 86 Rev12 and 86 Rev23, grouped by the interface they are collected from.
- *Meta-Data* The corpus is shipped with students' responses to the pre-survey and post-survey questionnaires, and their annotation verification information. It also contains the score for each students' drafts (258 essays) and the expert feedback given to the Draft1 of the students (86 feedback).

5 Example usage: revision purpose classification

The corpus will be useful for developing a variety of applications, from revision analysis to predicting whether a text chunk is expressing a general or specific piece of information. Additionally, the corpus affords the examination of a variety of

feedback types and the types of revisions that follow. Scholars in composition and educational research might find it useful to map patterns in revisions to the four different interfaces. Most readers of this paper, however, will likely be interested in the computational uses of the corpus, which we outline below. In this section, we demonstrate one example usage of the corpus—the development of revision purpose classifiers, a component of an argumentative revision analysis system. The variety of revision unit scope and purpose granularity levels allows us to study a variety of revision classification tasks with different settings. Therefore, we experiment on a *binary* classification task (Sect. 5.3) and a *multi-class* classification task (Sect. 5.4), both trying to predict the purpose of the sentential and subsentential revisions.

Since our main objective is to demonstrate the usefulness of our corpus for NLP applications, we do not develop highly domain specific features or complex models for the classification tasks. Instead, we opt for some features (Sect. 5.1) and models (Sect. 5.2) that are widely applicable to many NLP applications (Burstein et al., 2001; Daxenberger & Gurevych, 2013; Jabreel & Moreno, 2018; Zhang & Litman, 2016).

5.1 Features

We use a mixture of features to represent textual (length and position), syntactic (part-of-speech), semantic (embedding), and discourse (transition words) aspects of a revision as follow:

- *Length* the length of the sentence in number of its words.
- *Position* the index (location) of the sentence in the essay's sentences.
- *Embedding* the vector representation of the sentence encoded using *Universal Sentence Encoder* (USE), which is a pre-trained transformer-based encoder of greater-than-word length text (Cer et al., 2018).
- *Part-Of-Speech* the *term frequency* representation of the sentence words' part-of-speech (POS) tags, predicted using `spaCy`[3]. See Appendix G for more details on how we generate the POS term frequency representation.
- *Transition words* the term frequency representation of the transition words in the sentence. See Appendix H for the complete list of transition words we used
- to represent the discourse aspect of the revisions.

Each *sentential* revision is represented as the pair of <old-sentence, new-sentence>, which could be either between Draft1 and Draft2 (Rev12), or between Draft2 and Draft3 (Rev23). These sentences then transform into feature space. Each *subsentential* revision is represented as the pair of <old-phrase, new-phrase>, which could be either between Draft1 and Draft2 (Rev12), or between Draft2 and Draft3 (Rev23). To take the context of revised phrases into account, we extend the subsentential revision representation to include the sentences in which the phrases are used as: <old-phrase|old-sentence, new-phrase|new-sentence>. Each context sentence and subsentential revision is transformed into feature space in the same

[3] https://spacy.io/.

process as for sentential revision representation, except for the *position* feature of subsentential revisions, which is their starting index in the context sentence. Note that in our experiments, we assume revisions are pre-segmented and pre-aligned at the desired revision scope level, based on the classification task settings.

5.2 Training settings

The choice of classifier model, tuning, datasets, and evaluation methodology of our experiments are as follow:

- *Classifier model*
  - *XGB* We opt to use XGBoost (Chen & Guestrin, 2016) as the learning algorithm in all classification tasks. For each classification task, we explore a range of hyperparameter, including: `number of the estimators` $\in \{250, 500, 750, 1000\}$, `maximum depth` $\in \{3, 4, 5\}$, and `learning rate` $\in \{0.1, 0.05, 0.01\}$. We pick the final setting through a randomized parameter search with cross-validation process (Bergstra & Bengio, 2012).
  - *Majority* To better understand the classification result of each task and check whether we trained reasonable models, we compare the results with a simple majority classifier baseline, which assigns the most frequent revision purpose of the dataset to all revisions.

- *Sentential dataset* The sentential revision dataset contains 3238 training examples collected from all four interfaces, with coarse and fine revision purpose annotation levels. We use this dataset to train the sentential binary and multi-class revision purpose classifiers.
- *Subsentential dataset* The subsentential revision dataset contains 2596 training examples collected from interface A and D, with coarse and fine revision purpose annotation levels. We use this dataset to train the subsentential binary and multi-class revision purpose classifiers.
- *Evaluation* Classifiers are evaluated in a fivefold cross-validation process using *average unweighted F-score* and *Accuracy* measure.

5.3 Binary classification

In this section, we experiment with the task of predicting whether the purpose behind a revision is to make a content-level change or a surface-level change. Given that the ArgRewrite V.2 contains purpose annotation for sentential and subsentential annotations, we also investigate how coarse-grained revision purpose prediction tasks may differ for different revision scopes. For the sentential classification, we trained the models on the *sentential dataset*, and for the subsentential classification, we trained the models on the *subsentential dataset*. To better understand the contribution of different features (see Sect. 5.1), we experiment with three

Table 6 The F-score and accuracy of binary revision purpose classification, †indicates significantly better than Features ($p < 0.05$)

Scope	Model	Surface	Content	AVG	ACC
Sentence	Majority	0.00	0.73	0.37	0.58
	Features	0.89	0.91	0.90	0.90
	USE	0.91	0.92	0.92	0.92
	Features + USE	0.92	0.94	0.93	0.93†
	Cross-Task (USE)	0.70	0.72	0.71	0.71
Subsent.	Majority	0.76	0.00	0.38	0.61
	Features	0.88	0.86	0.87	0.87
	USE	0.89	0.88	0.88	0.89
	Features + USE	0.90	0.90	0.90	0.91†
	Cross-Task (USE)	0.09	0.76	0.42	0.62

classification settings: (1) training only on semantic features of the revisions (referred to as USE), (2) training on textual, syntactic, and discourse features (referred to as Features), and (3) training using all of the features (referred to as Features + USE). This ablation test can help to investigate the impact of semantics and how a pre-trained language model performs on the task.

Moreover, we study a cross-task experiment—*how does the model trained on sentential revisions performs on predicting the purpose of the subsentential revisions, and the other way around?* Since some features in the subsentential setting (e.g. positions, or the context sentence) are not applicable to the sentential setting, for this experiment, we use the model that is trained only on the embedding from USE, which is independent of tasks and domains. To collect comparable results with other classification settings, we evaluate the model through fivefold cross-validation, where in each fold, the training set, and the test set are picked from the corresponding splits from the sentential dataset and subsentential dataset, respectively.

Table 6 shows the average unweighted F-score and the accuracy (ACC) of our binary classification experiments for different revision scopes and settings[4]. In general, we can observe that while predicting if a revision is a content or a surface change, it is slightly easier at sentential-level than subsentential-level. Both supervised models can achieve a high classification performance, which outperforms the majority baseline by a big margin, in all classification settings. The embedding-only classifiers (USE) perform slightly better than the features-only classifiers (Features), while the classifiers that are trained on both (Features+USE) significantly outperform the feature-only classifiers. However, the difference between USE and Features+USE is not significant, which suggests the domain-independent approach of using only embeddings from a pre-trained language model might also produce comparable results on predicting whether a revision purpose is a content or surface change. Additionally, feature-only classifiers also produce promising classification results, suggesting this problem

[4] Hyperparameters: estimators = 500, maximum depth = 4, and learning rate = 0.05.

can also be addressed by more traditional solutions without sacrificing classification performance.

As we expected the classification performance drops in cross-task evaluation, however, results also imply that the model trained on the sentential revisions could be used to predict the coarse-grained revision purpose for the subsentential revision with acceptable performance (F1: .71), while the subsentential model performs only as good as a majority classifier on predicting the purpose of the sentential revision. One possible reason for this could be that the subsentential revisions that are labeled as content changes are relatively longer than surface changes (on average, 9 words compared to 5 words). Thus, this model will predict a content label for (almost) all of the sentential revisions, which are longer than an average subsentential revision.

5.4 Multiclass classification

In this section, we experiment with the task of predicting the fine-grained purpose of a revision and investigate how it may be influenced by the scope of the revision. Similar to our binary classification experiment, for the sentential classification we train the models on the *sentential dataset*, and for the subsentential classification we trained the models on the *subsentential dataset*, but this time with fine-grained revision purposes as the supervision.

We also perform an ablation test to investigate the contribution of different types of features to the task, and a cross-task experiment to see how does the pre-trained fine-grained revision purpose classifiers perform on predicting a purpose for revisions with different scope levels. For the same reasons mentioned in our binary cross-task experiment (see Sect. 5.3), here we also use the models that are trained only on the embeddings. To collect comparable results with other classification settings, we evaluate the model through fivefold cross-validation, where in each fold, the training set and test set are picked from the corresponding splits from the sentential dataset and subsentential dataset, respectively.

Table 7 shows the detailed average unweighted F-score and accuracy (ACC) of our fine-grained revision classification experiments for different revision scopes and settings[5]. Intuitively, the multi-class classification is harder than the corresponding binary classification task, and here we also observe the multi-class classification experiments yield lower revision purpose prediction results than the binary classification experiments. However, in contrast to our findings in binary classification experiments, we observe that predicting fine-grained revision purpose yields higher results for classifying subsentential revisions compared to sentential revisions. This observation is counter-intuitive because the *subsentential dataset* contains fewer training examples than the *sentential dataset* (3.2K vs. 2.6K, see Sect. 5.2). Referring back to the inter-annotator agreements for annotating sentential and subsentential revisions, it seems that annotating revisions at the subsentential level is much easier than annotating them as sentential units. A sentential level change might be the result of multiple subsentential changes, which do not necessarily have the same intended purposes, therefore, an amalgamation of

[5] Hyperparameters: `estimators` = 750, `maximum depth` = 5, and `learning rate` = 0.05.

Table 7 The F-score and accuracy of fine-grained revision purpose classification, †indicates significantly better than `Features` ($p < 0.05$), ‡indicates significantly better than `Features+USE` ($p < 0.05$)

Scope	Model	Surface			Content						AVG	ACC
		WRD	SPL	ORG	CLM	RSN	EVD	RBL	PRN	GCD		
Sentence	Majority	0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.29
	Features	0.79	0.34	0.32	0.25	0.48	0.42	0.35	0.45	0.56	0.44	0.58
	USE	0.78	0.38	0.35	0.33	0.58	0.45	0.37	0.54	0.59	0.49	0.62
	Features+USE	0.79	0.39	0.35	0.37	0.60	0.48	0.37	0.54	0.60	0.51	0.63†
	+DA	0.78	0.38	0.47	0.44	0.60	0.57	0.56	0.66	0.59	0.56	0.68‡
	Cross-Task (USE)	0.61	0.00	0.57	0.73	0.71	0.77	0.67	0.00	0.45	.045	0.55
Subsentence	Majority	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.44
	Features	0.82	0.63	0.33	0.40	0.67	0.27	0.45	0.35	0.50	0.49	0.62
	USE	0.83	0.65	0.33	0.48	0.60	0.46	0.45	0.35	0.46	0.54	0.66
	Features+USE	0.83	0.71	0.33	0.44	0.67	0.46	0.45	0.45	0.56	0.57	0.70†
	+DA	0.85	0.73	0.51	0.74	0.67	0.63	0.88	0.86	0.58	0.67	0.78‡
	Cross-Task (USE)	0.09	0.00	0.12	0.25	0.48	0.23	0.35	0.00	0.44	0.22	0.34

different revision purposes uniting under their most prominent revision purpose (see Appendix B for more details on annotation process). As a result, classification models may find it harder to predict the purposes for sentential revisions, as opposed to subsentential revisions, which are atomic revisions with only one clear revision purpose, so are relatively easier to annotate and classify.

Similar to the binary classification experiments, we also observe that, while the `Features+USE` classifiers do not perform significantly better than the embedding-only classifiers, they significantly outperform the feature-only classifiers. Moreover, the classification performance of the models drops when cross-evaluating them on predicting the purpose for different revision scopes. Similar to our observations in binary classification experiment, the model trained on the sentential revisions performs better in predicting fine-grained revision purpose for the substantial revisions than the other way around. This is in accordance with our intuition about the difference between the length of subsentential content and surface revisions, which may cause the subsentential models to predict a content-level purpose for (almost) all sentential revisions.

5.4.1 Data augmentation

In the binary classification problem, the training examples are either surface or content revisions, so each of them comprises a reasonable amount of training examples, however, in our multi-class classification problem, training examples are distributed into nine classes, so compared to the binary case, we have fewer training examples for each class, while some, may seriously lack training examples (e.g., there are only 42 and 85 sentential training examples for the rebuttal and precision class, respectively). This training examples scarcity could be the main cause of the relatively lower prediction accuracy of fine-grained revision purposes, especially for under-represented revision purposes. In order to investigate this, we study how data augmentation may help to improve the fine-grained revision purpose prediction performance by providing more training examples for under-represented classes.

We use a customized version of the synonym replacement (SR) augmentation strategy—randomly pick a content word from the sentence and replace it with a synonym chosen at random (Wei & Zou, 2019), as our augmentation strategy to generate training examples. In general, we generate up to *4* (on average: 3.4) augmented examples by substituting about 20% of its content-words with their synonyms, which are retrieved using `sense2vec` contextual word embedding (Trask et al., 2015), for each examples of the underrepresented revision purposes, namely *claims*, *rebuttal*, *evidence*, *precision*, and *organization*. Our data augmentation strategy is discussed in detail elsewhere (Kashefi & Hwa, 2020).

During the cross-validation, each time, we expanded the training fold with new augmented examples and evaluate the model on the test fold. The rows indicated by `+DA` in Table 7 show that incorporating data augmentation to generate more training examples can improve the fine-grained revision purpose classification at both sentential and subsentential levels. Aside from overall classification improvements, we can also observe an average F-score improvement of around 30% and

70% for classifying the underrepresented sentential and subsentential revision purposes, respectively, when the training set is augmented with more samples for them. Therefore, with more training examples, the fine-grained purpose of revisions could also be precisely predicted.

6 Related work

Early studies describe revision as a recursive process that involves both lexical and semantic changes (Fitzgerald, 1987; Flower & Hayes, 1981; Sommers, 1980). Those studies also show that effective writers' revision strategies differ from those of novice writers (Flower & Hayes, 1981). Hence, more and more studies have focused on understanding students' revision efforts. However, research on writing revision is inadequate in NLP. Prior NLP research on writing has focused on analysis of a single drafts as opposed to multiple iterations of the same composition. Such studies have focused on, for example, esssay scoring (Attali & Burstein, 2006; Taghipour & Ng, 2016), discourse structure analysis (Burstein et al., 2003; Falakmasir et al., 2014) and paraphrase detection (Barron-Cedeno et al., 2013; Dolan & Brockett, 2005; Tan & Lee, 2014; Vila et al., 2015), grammatical or semantic error correction (Dahlmeier et al., 2013; Kashefi et al., 2018; Yannakoudakis et al., 2011). The most closely related work to ours that has focused on revision are the bodies of literature on Wikipedia user edits or student academic essay revision.

Most related to our work is the Wikipedia revision analysis and categorization (Bronner & Monz, 2013; Daxenberger & Gurevych, 2013; Sarkar et al., 2019). Revision categorization of user edits from Wikipedia focus on both coarse-level (Bronner & Monz, 2013) and fine-grained (Daxenberger & Gurevych, 2012; Jones, 2008; Yang et al., 2017) categories. Although coarse-level categories (e.g., surface vs. content) can be generalized for academic writing, some fine-grained Wikipedia categories (e.g., vandalism) are specific to wiki scenarios. In academic writing, previous studies instead use fine-grained revision categories more suitable for student argumentative writing (Toulmin, 2003; Zhang & Litman, 2015). The above studies focus on investigating the reliability of manually annotating and automatically classifying the revision categories. Other related works for categorizing revisions include measuring statement strength of revised sentences in academic writing (Tan & Lee, 2014), sentence-level revision improvement in argumentative writing (Afrin & Litman, 2018), modeling revision requirement in wiki instructions (Bhat et al., 2020), etc.

There are many NLP-based writing assistant tools that were developed over the last few years. Such tools usually focus on grammar error correction of a single draft, few also provide high-level semantic error suggestions. For example, Grammarly (2016) provides feedback on grammatical error correction and fluency or word-usage, ETS-writing-mentor (Writing Mentor, 2016) provides feedback to reflect on higher-level essay properties such as coherence, convincingness, etc. Other tools such as EliReview (Eli Review, 2014), Turnitin (2014) are focused on peer feedback, plagiarism detection than focusing on student revision analysis. In

contrast, our ArgRewrite revision assistant tool is focused on students' revision between two drafts of an essay. The prior version of our system provided feedback based on detailed revision categorization at the sentence-level (Zhang et al., 2016). Our new system, ArgRewrite V.2, augmented the prior work by developing two additional interfaces for *binary sentential* (Interface B) and *fine-grained subsentential* (Interface D) revision categorization. Impact of different interfaces on students' writings are evaluated using both survey and writing improvement data (Afrin et al., 2021).

7 Conclusion

We have introduced ArgRewrite V.2, a corpus of revisions that are collected from argumentative essays written by university students in response to a writing prompt, and revised in response to some revision feedback. Revisions are semi-automatically aligned at both sentential and subsentential units, and each revision unit, then, manually annotated by domain experts with its coarse and fine-grained purpose category.

Aside from the annotated revisions, ArgRewrite V.2 also includes additional meta-data such as participants' demographic and self-regulation survey, as well as evaluative feedback on the drafts. To demonstrate the potential of ArgRewrite as a resource for revision analysis and other NLP applications, we explored usages of the corpus in a variety of automatic revision purpose prediction tasks.

Appendix: Writing prompt

Students are asked to read a brief article about self-driving cars, and then write a short argumentative essay in response to the following prompt:

> In this argumentative writing task, imagine that you are writing an op-ed piece for the Pittsburgh City Paper about self-driving cars. The editor of the paper has asked potential writers, like you, to gather information about the use of self-driving cars, and argue whether they are beneficial or not beneficial to society.
>
> In your writing, first, briefly explain both the advantages and disadvantages of self-driving cars. Then, you will choose a side, and construct an argument in support of self-driving cars as beneficial to society, or against self-driving cars as not beneficial to society. A high quality op-ed piece maintains a clear position on the issue and uses supporting ideas, strong evidence from the reading, explanations of your ideas and evidence, and a counter-argument. Furthermore, a high quality op-ed piece is clearly organized, uses precise word choices, and is grammatically correct.

Annotation guideline

Alignment annotation

The essays for each draft are tokenized into sentences. The sentences are enumerated from 1 to N according to their occurrence in the essay as 'Sentence Index'. For 'Aligned Index', each sentence in the revised draft is assigned the index of its aligned sentence in the original draft. Also, each sentence in the original draft is assigned the index of the aligned sentence in the revised draft. If a sentence is newly added, it will be marked as ADD. If a sentence is deleted from the old draft, it will be marked as DELETE.

Rules

1. Every sentence should either be aligned (one-to-one, one-to-many, many-to-one) or marked as ADD or DELETE. Only the alignment from the Old Draft to New Draft contains "DELETE" and only alignment from the New Draft to Old Draft contains "ADD".

2. For one-to-one case, align the sentences if the revised sentence is either replication or modification of the original sentence with one or several of the following changes:

   (a) Addition/deletion of some content within the sentence
   (b) Modification of words, phrases
   (c) Restatement of the ideas of the sentence

   The aligned sentences should be either syntactically or semantically close and within the same/similar context (i.e., the paragraphs the sentences belong to should be similar)

   – *Syntactically similar* The two sentences look explicitly similar to each other. (i.e., the difference between the two sentences should be a small ratio of the whole sentence. For example, a sentence with less than 10 words should have at most 2 words that are different (Does not count the change of words in the same stem, e.g. change− > changes)).
   – *Semantically similar* The two sentences describe the same information, or the revised sentence adds/deletes information on the basis of the original sentence

3. For many-to-one and one-to-many cases, only align when multiple sentences are syntactically similar to some part of the one target sentence. When multiple sentences can be combined without major addition/deletion/modification of words/phrases to construct the aligned sentence. Or, when one sentence can be divided to construct the aligned sentences without major addition/deletion/modification of words/phrases. It should also be explicit and better to align the target

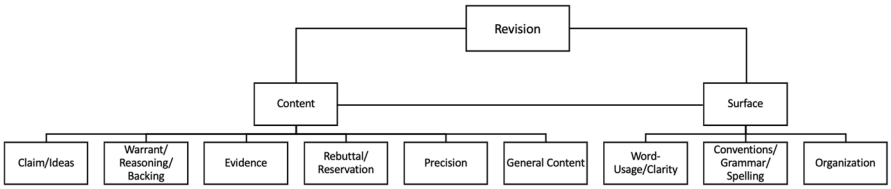

Fig. 3 Revision purpose schema

sentence to the group of sentences than to align the target sentence to one or some of the sentences.

Revision purpose annotation

Each aligned sentence (including ADD and DELETE) should have ONE major revision purpose. As shown in the Fig. 3, each revision purpose can be classified as two higher-level changes—surface and text/content. These change can be further categorized into 9 major revision purposes. The annotator is required to annotate ONLY the major revision purpose. Annotator has to obey the following rules to decide the major revision purpose type.

Rules

1. *Importance orders of revision purposes (Higher to lower)*
   The importance of different revision purpose type is different, when there are multiple revision purpose types in one revised sentence, make sure that the more important one is selected. The following sub-rules explains more specific details for cases where the decision of the appropriate revision purpose can be difficult.
   - *Claim/Ideas vs. Warrant/Reasoning/Backing*
     An essay can have one major claim and several sub-claims to support the major claim. These sub-claims are usually in the form of reasoning to support the major claim. Thus the differentiation of sub-claim and reasoning for the major claim can be ambiguous. We ask the annotators to think of the Claim and Reasoning as a hierarchical tree structure. The leaves of the tree are marked as "Warrant/Reasoning/Backing" while the others are marked as "Claim/Ideas". Hence, there should be no "Warrant/Reasoning/Backing" without a "Claim/Ideas" seen before. In specific, if the major idea of the essay is further supported or objected by other sentences, it is considered as a Claim. If the sentence cannot be classified as Evidence/Rebuttal of the Claim, but the sentence contains elements backing or reasoning for or against the Claim, it should be annotated as "Warrant/Reasoning/Backing".

- *General Content vs. Warrant/Reasoning/Backing*
  Differentiating General Content and Reasoning can be difficult as they both often occur after the author proposes a claim. To differentiate the two categories, the annotator is required to distinguish whether the author is suggesting his position for his claim in the sentences or not. If the annotator senses the author's sentiment position towards his claim, then it should be "Warrant/Reasoning/Backing", whereas it should be "General Content".
- *Evidence vs. Warrant/Reasoning/Backing*
  These two categories are similar as they both provide support to the authors' claim. The annotators are required to distinguish these two categories according to whether the sentences are stating facts. The facts can be (1) Citation: the citation of papers, reports, news and books. (2) Example: facts of history or personal experiences. (3) Scientific proof. If there are facts involved, it is marked as Evidence, otherwise it is marked as Warrant.
- *Conventions/Grammar/Spelling vs. Word Usage/Clarity*
  These two genres are similar as they do not change the content of the text and improve the quality of the text. The annotators are required to make the judgment according to the question: Are there spelling/grammar mistakes in the original draft and has this mistake been addressed in the new draft? If the ONLY a mistake is addressed, it should be marked as "Grammar/Spelling".
- *Precision vs Word-Usage/Clarity*
  These two genres are not similar but annotating can be confusing. When there is a word/phrase change in the sentence that significantly change the specificity level of a sentence to make it more specific or general as a content revision, then it is a precision change, otherwise it will be a word-usage change.
- *Claim/Idea vs Word-Usage/Clarity*
  These two genres are not similar but annotating can be confusing when there is a word/phrase change in the claim of the essay affecting major claim. If the change of the sentence affects/changes the claim of the essay, it should be annotated as Claim/Ideas instead of Word-Usage/Clarity. Because a change in the claim affects the subsequent changes of warrant/evidence. A feedback of the revision as claim change would help writers understand and think about the changes of the essay better than a feedback of a word usage.
- *Organization vs General Content Development*
  Although these two categories seem very different, annotators need to be very careful while annotating these two. General content changes are usually heavy changes in the sentence (compared to Word-Usage) or added and deleted sentences. If merged or split sentences do not have major change in words, it should be Organization. However, if those sentence have major change in words so that it is better to consider them as individual sentence rather than aligned sentence, then it should be annotated as General Content. Sometimes reordered sentences maybe aligned as DELETE and then ADD. In those cases, it should be considered as Organization rather than DELETE General Content and then ADD General Content.

2. *Focus on WHAT than WHERE*
   It is not necessarily that revisions made on the thesis of the paragraph are Claim/Idea changes, the type of the change should be determined according to what the author really has changed. For example, in a Claim sentence of a paragraph, if the author added a clause in the new sentence for reasoning the claim, the change would be a Warrant/Reasoning/Backing change; if the author only replaced some word with a more appropriate form of word, the annotator should mark it as Word usage change. However, if the change affects the claim it should be a Claim/Ideas change as stated before.
3. *Read and understand the prompt before the annotation*
   Sometimes the annotation of revision purpose could be different according to what the author is really targeting. So it is critically important that the annotator read and understand the prompt before the annotation. For example, in a regular essay, a sentence change from "Fidel Castro would be a good example for this case" to "Saddam Hussein would be a good example for this case" would typically be "Evidence". However, if the prompt of the essay writing assignment is "Put the contemporaries at different levels of Hell", then the annotation would be "Claim/Ideas".
   We have developed an annotation tool to ease the annotation of alignments, the tool automatically breaks the text to sentences and the annotator only needs to do the annotation on the interface. After the alignment completes, the annotator can select the type of the revision purpose. Check out more details of the tool in the annotation too manual.

Interfaces

Each participant was randomly assigned to one of the following four interfaces, which provided different types of feedback on differences between Draft1 and Draft2, including the size of the revision unit span and the granularity of the revision purpose category. For more details refer to (Afrin et al., 2021).

- *Interface A* The 20 participants assigned to this condition were shown only the changed sentences without any further feedback (Fig. 4a);
- *Interface B* The 22 participants assigned to this condition were shown sentence-level differences, as either a surface or content revision (Fig. 4b);
- *Interface C* The 22 participants assigned to this condition were shown sentence-level differences with fine-grained revision purposes (Fig. 4c);
- *Interface D* The 22 participants assigned to this condition were shown subsentential differences with fine-grained revision purposes (Fig. 4d)

(a) Interface A **(b)** Interface B

(c) Interface C **(d)** Interface D

Fig. 4 Screenshot of Different Conditions, where warmer colors indicate content revisions and colder colors indicate surface revisions. **a** No Feedback; **b** Sentence-Level feedback with coarse-grained (surface vs. Content) revision purposes; **c** Sentence-Level feedback with fine-grained revision purposes; **d** Subsentential-Level feedback with fine-grained revision purposes

Feedback

Prior study's feedback

The same feedback given to all students in the prior version of the corpus (Zhang & Litman, 2015):

> Strengthen the essay by adding one more example or reasoning for the claim; then add a rebuttal to an opposing idea; keep the essay at 400 words.

Personalized feedback example

An example of a personalized feedback message:

> Thank you for your participation in the study. Your draft has been read, and feedback from an expert writing instructor is written below. We advise that you use this feedback when you revise.
>
> The strengths of your essay include:

– All claims have relevant supporting evidence, though that evidence may be brief or general.

– You respond to one, but not all parts of the prompt. However, your entire essay is focused on the prompt.

Areas to improve in your essay include:

– You provided a statement that somewhat show your stance for or against self-driving cars, but it is unclear, or is just a restatement of the prompt.
– Your essay's sequence of ideas is inconsistent, with some clear and some unclear progression.
– Your essay does not include a rebuttal.

Scoring rubric

Each participants is given a personalized feedback in the form of lists of 2–4 strengths and 2–4 weaknesses that characterized their first draft of the essay based on the following scoring rubric (Table 8):

Pre-study questionnaires

– Are you an undergraduate or graduate student?

– What is your current year of study?
– Is English your native language?
– What is your native language?
– When writing an essay/paper for a class, how many drafts (that are not required by the class) do you typically write?
– Overall, how confident are you with your writing?
– Please tell us how comfortable you feel about writing in the English language versus writing in your primary language.
– What aspects of writing do you think you are good at?[Click all that apply]
– What aspects of writing do you think you can improve?[Click all that apply]
– I typically set aside routine, planned times to complete writing tasks.
– I typically create an outline of my writing before I begin any writing task.
– I typically seek out feedback from others on my writing.
– I typically plan time for multiple revisions of my writing.
– I typically set revision goals for myself to meet the requirements of a writing task.
– The revision goals I set for myself focus mostly on developing the content or thesis.
– The revision goals I set for myself focus mostly on surface level changes (e.g., grammar, spelling, organization and word clarity).
– While I am revising, I typically look back at or think about my previous draft(s) to refine my essay.

Table 8 Argumentative essay rubric

	1-Poor	2-Developing	3-Proficient	4-Excellent
Response to prompt	The essay is off topic, and does not consider or respond to the prompt in any way	The essay addresses the topic, but the entire essay is not focused on the prompt. The author may get off topic at points	The author responds to one, but not all parts of the prompt, but the entire essay is focused on the prompt	The author responds to all parts of the prompt and the entire essay is focused on the prompt
Thesis	The author did not include a statement that clearly showed the author's stance for or against self-driving cars	The author provided a statement that somewhat showed the author's stance for or against self-driving cars, though it may be unclear or only a restatement of the essay prompt	The author provided a brief statement that reflects a thesis, and is indicative of the stance the author is taking toward self-driving cars	The author provided a clear, nuanced and original statement that acted as a specific stance for or against self-driving cars
Claims	The author's claims are difficult to understand or locate	The author's claims are present, but are unclear, not fully connected to the thesis or the reading, or the author makes only one claim multiple times	The author makes multiple, distinct, and clear claims that align with either their thesis or the given reading, but not both	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 for claims	The author does not provide any evidence to support thesis/claims	Less than half of claims are supported with relevant or credible evidence or the connections between the evidence and the thesis/claims is not clear	All claims have relevant supporting evidence, though that evidence may be brief or general. The source of the evidence is credible and acknowledged/cited where appropriate	The author provides specific and convincing evidence for each claim, and most evidence is given through detailed personal examples, relevant direct quotations, or detailed examples from the provided reading. The source of the evidence is credible and acknowledged/cited where appropriate
Reasoning	The author provides no reasoning for any of their claims	Less than half of claims are supported with reasoning or the reasoning is so brief, it essentially repeats the claim. Some reasoning may not appear logical or clear	All claims are supported with reasoning that connect the evidence to the claim, though some may not be fully explained or difficult to follow.	All claims are supported with clear reasoning that shows thoughtful, elaborated analysis

Table 8 continued

	1-Poor	2-Developing	3-Proficient	4-Excellent
Reordering/	Organization	The sequence of ideas/claims is difficult to follow and the essay does not have an introduction, conclusion, and body paragraphs that are organized clearly around distinct claims	The essay's sequence of ideas is inconsistent, with some clear and some unclear progression of ideas OR the essay is missing a distinct introduction OR conclusion	The essay has a clear introduction, body, and conclusion and a logical sequence of ideas, but each claim is not located in its own separate paragraph
The essay has an	introduction, body and conclusion and a logical sequence of ideas. Each paragraph makes a distinct claim			
Rebuttal	The essay does not include a rebuttal	The essay includes a rebuttal in the sense that it acknowledges another point of view, but does not explore possible reasons why this other viewpoint exists	The essay includes a rebuttal in the form of an acknowledgement of a different point of view and reasons for that view, but does not explain why those reasons are incorrect or unconvincing	The essay explains a different point of view and elaborates why it is not convincing or correct
Precision	Throughout the essay, word choices are overly informal and general (e.g., "I don't like self-driving cars because they have problems.")	Word choices are mostly overly general and informal, though at times they are specific	Word choices are mostly specific though there may be a few word choices that make the meaning of the sentence vague	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	A majority of sentences are difficult to understand because of incorrect/ inappropriate word choices and sentence structure	A noticeable number of sentences are difficult to understand because of incorrect/ inappropriate word choices and sentence structure, although the author's overall point is understandable	Most sentences are clear because of correct and appropriate word choices and sentence structure	All sentences are clear because of correct and appropriate word choices and sentence structure

Table 8 continued

	1-Poor	2-Developing	3-Proficient	4-Excellent
Conventions/ Grammar/ Spelling	The author makes many grammatical or spelling errors throughout their piece that interfere with the meaning	The author makes many grammatical or spelling errors throughout their piece, though the errors rarely interfere with meaning	The author makes few grammatical or spelling errors throughout their piece, and the errors do not interfere with meaning	The author makes few or no grammatical or spelling errors throughout their piece, and the meaning is clear

- While I am revising, I typically look back at or think about feedback from others to refine my essay.
- While I am revising, I typically think about the reader's expectations.
- While I am revising, I typically address grammatical errors.
- While I am revising, I typically try to develop the content or thesis.
- When I make a revision, I reread the sentence, paragraph, or whole essay to see whether my revision improved the essay.
- I can meet the requirements of a writing task without revising.
- I am confident in my writing and revising abilities.

Post-study questionnaires

Followings are the questions asked from all students, regardless of the interface they assigned to:

- The system allows me to have a better understanding of my previous revision efforts.
- I find the system easy to use.
- My interaction with the system is clear and understandable.
- The system helps me to recognize the weakness of my essay.
- The system encourages me to make more revisions (quantity) than I usually do.
- The system encourages me to make more meaningful revisions (quality) than I usually do.
- Overall the system is helpful to my writing.
- I put a lot of effort into writing and revising this essay.
- How could the system be more helpful?

Following questions are only asked form the students who were assigned to *Interface A*:

- What led you to notice that some parts of your essay needed to be revised?
- Was this revision process similar to how you normally revise your essays?

Following questions are only asked form the students who were assigned to *Interface B*:

- I found the overview page to be useful.
- The description of the purpose of my revisions inspired me to make more revisions.
- I found it useful to see my revision purposes highlighted in different colors (i.e., Warm and cold colors)
- I found the revision map visualization useful.
- I found the small window of revision details to be useful.

- In general, I found it helpful to know whether my revision was a surface or content level change.
- My revision purposes were most often indicated correctly by the system.
- I trust the feedback that the system gave me.
- What influenced your decision to make revisions to Draft3?

In addition to all the question that are asked from the students in Interface B, students who were assigned to *Interface C* are also asked the following question:

- I found it helpful to have the specific purposes of my revisions indicated (e.g., claim, evidence, warrant, etc.).

In addition to all the question that are asked from the students in Interface C, students who were assigned to *Interface D* are also asked the following question:

- The system accurately highlighted each, specific area of text that I revised (this area of text could be as small as a word, or as large as a sentence).

Term frequency representation

The term frequency representation is a vector with the size of the total number of classes. The `spaCy` library recognizes 19 different POS tags, so the term frequency representation of the POS is an array with length 19, where each index represents a POS tag, and the number at each index represents the total number of words in a sentence that has that POS tag.

For example, consider the following sentences and the associated POS of its word:

this	is	a	revised	sentence
DET	VERB	DET	VERB	NOUN

The POS term frequency representation of this sentence would be [00000020100 00000200], where each index represent the number of words with one of the 19 POS tags, for example, the number at index 0 represents the number of ADJECTIVEs (there is none of the in the sentence so the value is 0), the 6th index represents the number of DETs (there are two words with this tag so the value is 2), the 8th index represents NOUNs (there is one word with this tag so the value is 1), and the 16th index represents the number of VERBs (we have two verbs so the values is 2).

Transition words

Table 9 includes the list of words we used for calculating the term frequency representation of transition words as a feature for revision purpose classification tasks. We collected these words from multiple transition word lists published by the

Table 9 Transition words used for training revision purpose classifiers

Group	Words
Reasoning	Consequently, clearly, then, furthermore, additionally, moreover, because, besides, also
Evidence	(as an) illustration, e.g., (for) example, (for) instance, specifically, (to) demonstrate, (to) illustrate
Rebuttal	However, but, yet, although, despite, (in) contrast, nevertheless, nonetheless, notwithstanding, (on the) contrary, otherwise, though, yet
Conclusion	Therefore, hence, conclusion, consideration, indeed, finally, lastly
Details	Specifically, especially, (in) particular, (to) explain, (to) list, (to) enumerate, (in) detail, namely, including
Causation	Accordingly, so, because, consequently, hence, since, therefore, thus

writing centers of some universities[6] and filtered for those words and categories that we though might correspond to our revision purpose categories.

References

Afrin, T., & Litman, D. (2018). *Annotation and classification of sentence-level revision improvement* (pp. 240–246). New Orleans, Louisiana.

Afrin, T., Kashefi, O., Olshefski, C., Litman, D., Hwa, R., & Godley, A. (2021). *Effective interfaces for student-driven revision sessions for argumentative writing* (pp. 1–13)

Afrin, T., Wang, E. L., Litman, D., Matsumura, L. C., & Correnti, R. (2020). *Annotation and classification of evidence and reasoning revisions in argumentative writing* (pp. 75–84)

Allal, L., Chanquoy, L., & Largy, P. (2004). *Revision cognitive and instructional processes. Studies in writing*. Springer.

Amorim, E., Cançado, M., & Veloso, A. (2018). Automated essay scoring in the presence of biased ratings. In *NAACL* (pp. 229–237).

Attali, Y., & Burstein, J. (2006). The automated essay scoring with e-rater vol 2. *Journal of Technology, Learning, and Assessment, 4*(3).

Barron-Cedeno, A., Vila, M., Marti, A., & Rosso, P. (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics, 39*(4), 917–947.

Barry, E. S. (2006). Can paraphrasing practice help students define plagiarism? *College Student Journal, 40*(2), 377–384.

Beach, R., & Anson, C. M. (1988). The pragmatics of memo writing. *Written Communication, 5*(2), 157–183.

Beason, L. (1993). Feedback and revision in writing across the curriculum classes. *Research in the Teaching of English, 27*(4), 395–422.

Berant, J., & Liang, P. (2014). Semantic parsing via paraphrasing. In *ACL* (pp. 1415–1425).

Berg-Kirkpatrick, T., Gillick, D., & Klein, D. (2011). Jointly learning to extract and compress. In *ACL* (pp. 481–490).

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*(2), 281–305.

Bhat, I., Anthonio, T., & Roth, M. (2020). Towards modeling revision requirements in wikiHow instructions. In *EMNLP* (pp. 8407–8414).

Bronner, A., & Monz, C. (2013). User edits classification using document revision histories. In *EACL* (pp. 356–366).

Burstein, J., Marcu, D., Andreyev, S., & Chodorow, M. (2001). Towards automatic classification of discourse elements in essays. In *ACL* (pp. 98–105).

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems, 18*(1), 32–39.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The E-rater automated essay scoring system. In J. Burstein & M. D. Shermis (Eds.), *Handbook of automated essay evaluation. Current applications and new directions* (pp. 55–67). Routledge.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St John, R., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., Sung, Y. H., Strope, B., & Kurzweil Google Research Mountain View, R. (2018). *Universal sentence encoder*. Computing Research Repository. Retrieved from http://arxiv.org/abs/1803.11175

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *KDD* (pp. 785–794).

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and instruction, 20*(4), 328–338.

Connor, U., & Asenavage, K. (1994). Peer response groups in ESL writing classes: How much impact on revision? *Journal of Second Language Writing, 3*(3), 257–276.

[6] https://writing.wisc.edu/handbook/style/transitions/
http://writing2.richmond.edu/writing/wweb/trans1.html
https://writingcenter.unc.edu/tips-and-tools/transitions/.

Coster, W., & Kauchak, D. (2011). *Learning to simplify sentences using wikipedia. In: Workshop on Monolingual Text-to-Text Generation* (pp 1–9).
Crammond, J. G. (1998). The uses and complexity of argument structures in expert and student persuasive writing. *Written Communication, 15*(2), 230–268.
Dahlmeier, D., & Ng, H. T. (2011). Grammatical error correction with alternating structure optimization. In *ACL* (pp. 915–923).
Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a large annotated corpus of learner english: The NUS Corpus of Learner English (pp. 22–31).
Daxenberger, J., & Gurevych, I. (2012). A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *COLING* (pp. 711–726).
Daxenberger, J., & Gurevych, I. (2013). Automatically classifying edit categories in wikipedia revisions. In *EMNLP* (pp. 578–589).
Dolan, W. B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In: *International Workshop on Paraphrasing*.
Eli Review, T. (2014). Retrieved December 01, 2021 from https://elireview.com
Falakmasir, M., Ashley, K., Schunn, C., & Litman, D. (2014). Identifying thesis and conclusion statements in student essays to scaffold peer review (pp. 254–259).
Filippova, K., Alfonseca, E., Colmenares, C. A., Kaiser, L., & Vinyals, O. (2015). Sentence compression by deletion with lstms. In *EMNLP* (pp. 360–368).
Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57*(4), 481–506.
Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382.
Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365.
Grammarly. (2016). Retrieved December 01, 2021 from http://www.grammarly.com
Jabreel, M., & Moreno, A. (2018). EiTAKA at SemEval-2018 Task 1: An ensemble of N-Channels ConvNet and XGboost regressors for emotion analysis of Tweets. In *SemEval* (pp. 193–199).
Jacovina, M., & McNamara, D. (2016). *Intelligent tutoring systems for literacy: Existing technologies and continuing challenges. Intelligent Tutoring Systems: Structure, Applications and Challenges* (pp. 153–174).
Jones, J. (2008). Patterns of revision in online writing: A study of wikipedia's featured articles. *Written Communication, 25*(2), 262–289.
Kashefi, O., & Hwa, R. (2020). Quantifying the evaluation of heuristic methods for textual data augmentation. In *WNUT-EMNLP* (pp. 200–208).
Kashefi, O., Lucas, A. T., & Hwa, R. (2018). Semantic pleonasm detection (pp. 225–230).
Kauchak, D., & Barzilay, R. (2006). Paraphrasing for automatic evaluation. In *NAACL* (pp. 455–462).
Li, J. J., & Nenkova, A. (2015). Fast and accurate prediction of sentence specificity. In *AAAI* (pp. 2281–2287).
Lugini, L., & Litman, D. (2018). Predicting specificity in classroom discussion. In *Workshop on innovative use of NLP for building educational applications* (pp. 52–61).
Magnifico, A., McCarthey, S., Kline, S., & Kennett, K. (2014). *Reconsidering peer feedback in argumentative essays*. American Educational Research Association.
Merrill, D. C., Reiser, B. J., Ranney, M., & Trafton, J. G. (1992). Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *The Journal of the Learning Sciences, 2*(3), 277–305.
Roscoe, R. D., & McNamara, D. S. (2013). Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology, 105*(4), 1010.
Sarkar, S., Reddy, B. P., Sikdar, S., & Mukherjee, A. (2019). StRE: Self attentive edit quality prediction in Wikipedia. In *ACL* (pp. 3962–3972).
Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. *College Composition and Communication, 31*(4), 378–388.
Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In *EMNLP* (pp. 1882–1891).
Tan, C., & Lee, L. (2014). A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication (pp. 403–408).
Tetreault, J., Foster, J., & Chodorow, M. (2010). Using parse features for preposition selection and error detection (pp. 353–358).

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.
Trask, A., Michalak, P., & Liu, J. (2015). *sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings*. Retrieved from http://arxiv.org/abs/151106388
Turner, J., & Charniak, E. (2005). Supervised and unsupervised learning for sentence compression. In *ACL* (pp. 290–297).
Turnitin. (2014). Reteieved December 01, 2021 from http://turnitin.com/
Vickrey, D., & Koller, D. (2008). Sentence simplification for semantic role labeling. In *ACL* (pp. 344–352).
Vila, M., Rodríguez, H., & Martí, M. A. (2015). Relational paraphrase acquisition from Wikipedia: The WRPA method and corpus. *Natural Language Engineering, 21*(3), 355–389.
Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP* (pp. 6382–6388).
Westby, C., Culatta, B., Lawrence, B., & Hall-Kenyon, K. (2010). Summarizing expository texts. *Topics in Language Disorders, 30*(4), 275–287.
Writing Mentor, T. (2016). *ETS writing mentor*. Retrieved December 01, 2021 from https://mentormywriting.org/
Xue, H., & Hwa, R. (2014a). Improved correction detection in revised ESL sentences (pp. 599–604).
Xue, H., & Hwa, R. (2014b). Redundancy detection in ESL writings. In *EACL* (pp. 683–691).
Yang, D., Halfaker, A., Kraut, R., & Hovy, E. (2017). Identifying semantic edit intentions from revisions in wikipedia. In *EMNLP* (pp. 2000–2010).
Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *ACL* (pp. 180–189).
Zhang, F., & Litman, D. (2015). Annotation and classification of argumentative writing revisions. In *Workshop on innovative use of NLP for building educational applications* (pp. 133–143).
Zhang, F., & Litman, D. (2016). Using context to predict the purpose of argumentative writing revisions. In *NAACL* (pp. 1424–1430).
Zhang, F., & Litman, D. J. (2014). Sentence-level rewriting detection. In *Workshop on innovative use of NLP for building educational applications* (pp. 149–154).
Zhang, F., Hashemi, HB., Hwa, R., & Litman, D. (2017). A corpus of annotated revisions for studying argumentative writing. In *ACL* (pp. 1568–1578).
Zhang, F., Hwa, R., Litman, D., & B Hashemi, H. (2016). ArgRewrite: A web-based revision assistant for argumentative writings. In *NAACL: Demonstrations* (pp. 37–41).