# Annotation and Classification of Sentence-level Revision Improvement

**Tazin Afrin**
University of Pittsburgh
Pittsburgh, PA 15260
tazinafrin@cs.pitt.edu

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

## Abstract

Studies of writing revisions rarely focus on revision quality. To address this issue, we introduce a corpus of between-draft revisions of student argumentative essays, annotated as to whether each revision improves essay quality. We demonstrate a potential usage of our annotations by developing a machine learning model to predict revision improvement. With the goal of expanding training data, we also extract revisions from a dataset edited by expert proofreaders. Our results indicate that blending expert and non-expert revisions increases model performance, with expert data particularly important for predicting low-quality revisions.

## 1 Introduction

Supporting student revision behavior is an important area of writing-related natural language processing (NLP) research. While revision is particularly effective in response to detailed feedback by an instructor (Paulus, 1999), human writing evaluation is time-consuming. To help students improve their writing skills, various writing assistant tools have thus been developed (Eli Review, 2014; Turnitin, 2014; Writing Mentor, 2016; Grammarly, 2016). While these tools offer instant feedback on a particular writing draft, they typically fail to explicitly compare revisions between drafts.

Our long term goal is to build a system for supporting students in revising argumentative essays, where the system automatically compares multiple drafts and provides useful feedback (e.g., informing students whether their revisions are improving the essay). One step towards this goal is the development of a machine-learning model to automatically analyze revision improvement. Specifically, given only two sentences - original and revised, our current goal is to predict if a revised sentence is better than the original.

In this paper, we focus on predicting revision improvement using non-expert (i.e., student) writing data. We first introduce a corpus of paired original and revised sentences that has been newly annotated as to whether each revision made the original sentence better or not. The revisions are a subset of those in the freely available ArgRewrite corpus (Zhang et al., 2017), with improvement annotated using standard rubric criteria for evaluating student argumentative writing. By adapting NLP features used in previous revision classification tasks, we then develop a prediction model that outperforms baselines, even though the size of our non-expert revision corpus is small. Hence, we explore extracting paired revisions from an expert edited dataset to increase training data. The expert revisions are a subset of those in the freely available Automated Evaluation of Scientific Writing (AESW) corpus (Daudaravicius et al., 2016). Our experiments show that with proper sampling, combining expert and non-expert revisions can improve prediction performance, particularly for low-quality revisions.

## 2 Related Work

Prior NLP revision analysis work has developed methods for identifying pairs of original and revised textual units in both Wikipedia articles and student essays, as well as for classifying such pairs with respect to schemas of coarse (e.g., syntactic versus semantic) and fine-grained (e.g., lexical vs. grammatical syntactic changes) revision purposes (Bronner and Monz, 2012; Daxenberger and Gurevych, 2012; Zhang and Litman, 2015; Yang et al., 2017). For example, the ArgRewrite corpus (Zhang et al., 2017) was introduced with the goal to facilitate argumentative revision analysis and automatic revision purpose classification. However, purpose classification does not ad-

*Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–246
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

dress revision quality. For example, a spelling change can both fix as well as introduce an error, while lexical changes can both enhance or reduce fluency. On the other hand, while some work has focused on correction detection in revision (Dahlmeier and Ng, 2012; Xue and Hwa, 2014; Felice et al., 2016), such work has typically been limited to grammatical error detection. The AESW shared task of identifying sentences in need of correction (Daudaravicius et al., 2016) goes beyond just grammatical errors, but the original task does not compare multiple versions of text, and also focuses on scientific writing.

In contrast, Tan and Lee (2014) created a dataset of paired revised sentences in academic writing annotated as to whether one sentence was stronger or weaker than the other. Their work directly sheds light on annotating sentence revision quality in terms of statement strength. However, their corpus focuses on the abstracts and introductions of ArXiv papers. Building on their annotation methodology, we consider paired sentences as our revision unit, but 1) annotate revision quality in terms of argumentative writing criteria, 2) use a corpus of revisions from non-expert student argumentative essays, and 3) move beyond annotation to automatic revision quality classification.

## 3 Corpora of Revised Sentence Pairs

### 3.1 Annotating ArgRewrite

The revisions that we annotated for improvement in quality are a subset of the freely available ArgRewrite revision corpus (Zhang et al., 2017)[1]. This corpus was created by extracting revisions from three drafts of argumentative essays written by 60 non-expert writers in response to a prompt[2]. Essay drafts were first manually aligned at the sentence level based on semantic similarity. Non-identical aligned sentences (e.g., modified, added and deleted sentences) were then extracted as the revisions. Our work uses only the 940 modification revisions, as our annotation does not yet consider a sentence's context in its paragraph.

We annotated ArgRewrite revisions for improvement using the labels *Better* or *NotBetter*. *Better* is used when the modification yields an improved sentence from the perspective of argumentative writing, while *NotBetter* is used when the modification either makes the sentence worse or does not have any significant effect. Binary labeling enables us to clearly determine a gold-standard using majority voting with an odd number of annotators. Binary labels should also suffice for our long term goal of triggering tutoring in a writing assistant (e.g., when the label is *NotBetter*).

Inspired by Tan and Lee (2014), our annotation instructions included explanatory guidelines along with example annotated sentence pairs. The guidelines were crafted to describe improvement in terms of typical argumentative writing criteria. We depend on annotators' judgment for cases not covered by the guidelines. According to the guidelines[3], a revised sentence $S2$ is better than the original sentence $S1$ when: (1) $S2$ provides more information that strengthens the idea/major claim in $S1$; (2) $S2$ provides more evidence/justification for some aspects of $S1$; (3) $S2$ is more precise than $S1$; (4) $S2$ is easier to understand compared to $S1$ because it is fluent, well-structured, and has no unnecessary words; and (5) $S2$ is grammatically correct and has no spelling mistakes.

To provide context, annotators were told that the data was taken from student argumentative essays about electronic communications. We also let the annotators know the identity of the original and revised sentences ($S1$ and $S2$, respectively). Although this may introduce an annotation bias, it mimics feedback practice where instructors know which are the original versus revised sentences.

We collected 7 labels along with explanatory comments for each of the 940 revisions using Amazon Mechanical Turk (AMT). Table 1 shows examples (1, 2, and 3) of original and revised ArgRewrite sentences with their majority-annotated labels. The first revision clarifies a claim of the essay, the second removes some information and is less precise, while the third fixes a spelling mistake. As shown in Table 2, for all 940 revisions, our annotation has slight agreement (Landis and Koch, 1977) using Fleiss's kappa (Fleiss, 1971). If we only consider revisions where at least 5 out of the 7 annotators chose the same label (majority $\geq 5$), the kappa values increase to fair agreement, $0.263$. Tan and Lee (2014) achieve fair agreement (Fleiss's kappa of $0.242$) with 9 annotators labeling 500 sentence pairs for statement strength.

[1] http://argrewrite.cs.pitt.edu

[2] Prompt shown in supplemental files.

[3] The guidelines can be found in supplemental files.

| | **Original Sentence** ($S1$) | **Revised Sentence** ($S2$) | **Label** |
|---|---|---|---|
| 1 | The world has *experienced various changes throughout its lifetime.* | The world has *been defined by its revolutions - the most recent one being technological.* | Better |
| 2 | Technology is changing the *world, and in particular the* way we communicate. | Technology is changing the way we communicate. | NotBetter |
| 3 | ...Susan says by to Shelly on the 125th St... | ...Susan says by*e* to Shelly on the 125th St... | Better |
| 4 | This is numerically expensive but leads to proper results. | This is numerically expensive, but leads to proper results. | Better |
| 5 | Section 2 *formulates and solves* the balance equations. | The balance equations *are formulated and solved in* Section 2. | Better |

Table 1: Example annotated revisions from ArgRewrite (1,2,3) and AESW (4,5). The label is calculated using majority voting (out of 7 annotators) for ArgRewrite and using expert proofreading edits for AESW.

| **Data** | **#Revisions** | **#Better** | **#NotBetter** | **Fleiss's Kappa** ($\kappa$) |
|---|---|---|---|---|
| All | 940(100%) | 784(83.4%) | 156(16.6%) | 0.201(Slight) |
| Majority$\geq$ 5 | 748(79.6%) | 658(88.0%) | 90(12.0%) | 0.263(Fair) |

Table 2: Number of revisions, number of $Better$ and $NotBetter$, and Fleiss's kappa ($\kappa$) per increasing majority voting (out of 7 annotators). Percentage of revisions are shown in parenthesis.

### 3.2 Sampling AESW

The Automated Evaluation of Scientific Writing (AESW) (Daudaravicius et al., 2016) shared task was to predict whether a sentence needed editing or not. Professional proof-readers edited sentences to correct issues ranging from grammatical errors to stylistic problems, intuitively yielding 'Better' sentences. Therefore, we can use the AESW edit information to create an automatically annotated corpus for revision improvement. In addition, by randomly flipping sentences we can include 'NotBetter' labels in the corpus.

The AESW dataset was created from different scientific writing genres (e.g. Mathematics, Astrophysics) with placeholders for anonymization. We use two random samples of 5000 AESW revisions for the experiments in Section 5. "AESW all" samples revisions from all scientific genres, while "AESW plaintext" ignores sentences containing placeholders (e.g. MATH, MATHDISP) to make the data more similar to ArgRewrite. Table 1 shows two example (4 and 5) AESW revisions.

## 4 Features for Classification

We adapt many features from prior studies predicting revision purposes (Adler et al., 2011; Javanmardi et al., 2011; Bronner and Monz, 2012; Daxenberger and Gurevych, 2013; Zhang and Litman, 2015; Remse et al., 2016) as well as introduce new features tailored to predicting improvement.

Following prior work, we count each unigram across, as well as unique to, S1 or S2 (Daxenberger and Gurevych, 2013; Zhang and Litman, 2015). However, we also count bigrams and trigrams to better capture introduced or deleted argumentative discourse units.

Another group of features are based on sentence differences similar to those proposed in (Zhang and Litman, 2015), e.g., difference in length, commas, symbols, named entities, etc., as well as edit distance. However, to capture improvement rather than just difference, we also introduce asymmetric distance metrics, e.g. Kullback-Leibler divergence[4]. We also capture differences using BLEU[5] score, motivated by its use in evaluating machine-translated text quality.

Following Zhang and Litman (2015), we calculate the count and difference of spelling and language errors[6], in our case to capture improvement as a result of error corrections.

As stated in the annotation guidelines, one way a revised sentence can be better is because it is more precise or specific. Therefore, we introduce the use of the Speciteller (Li and Nenkova, 2015) tool to quantify the specificity of S1 and S2, and take the specificity difference as a new feature.

Remse et al. (2016) used parse tree based fea-

[4] Using scipy.stats.entropy on sentence vectors.

[5] Using *sentence_bleu* from nltk.translate.bleu_score module, with S1 as reference and S2 as hypothesis.

[6] Using python 'language-check' tool.

| Experiments | Precision | Recall | F1 |
|---|---|---|---|
| Majority baseline | 0.417 | 0.500 | 0.454 |
| AESW all | 0.471* | 0.470 | 0.468 |
| AESW plaintext | 0.511* | 0.515 | 0.473 |
| ArgRewrite | 0.570* | 0.534 | 0.525* |
| ArgRewrite + AESW all | 0.497* | 0.501 | 0.488* |
| ArgRewrite + AESW plaintext | **0.574*** | **0.555*** | **0.551*** |

Table 3: 10-fold cross-validation performance. * indicates significantly better than majority ($p < 0.05$). Bold indicates highest column value.

tures to capture the readability, coherence, and fluency of a sentence. Inspired by them, we calculate the difference in count of subordinate clauses (SBAR), verb phrases (VP), noun phrases (NP), and tree height in the parse trees[7] of S1 and S2.

## 5 Experiments and Results

Our goal is to examine whether we can predict improvement for non-expert ArgRewrite revisions, using AESW expert and/or ArgRewrite non-expert revisions for training. Our experiments are structured to answer the following research questions:

**Q1:** Can we use only non-expert revisions to train a model that outperforms a baseline?

**Q2:** Can we use only expert revisions to train a model that outperforms a baseline?

**Q3:** Can we combine expert and non-expert training revisions to improve model performance?

Our machine learning experiments use Random Forest (RF) [8] from Python scikit-learn toolkit (Pedregosa et al., 2011) with 10-fold cross validation. Parameters were tuned using AESW development data. Because of the ArgRewrite class imbalance (Table 2, All row), we used SMOTE (Chawla et al., 2002) oversampling for each training fold. Feature selection was also performed on each training fold. Average un-weighted precision, recall and F1 are reported and compared to majority-class baselines.

To answer Q1, we train a model using only ArgRewrite data. Table 3 shows that this model outperforms the majority baseline, significantly so for Precision and F1. Compared to all other models (Figure 1), this model can identify 'Better' revisions with the highest recall, and can identify 'NotBetter' revisions with the highest precision. However, for our long-term goal of building an effective revision assistant tool, intuitively we will also need to identify 'NotBetter' revisions with higher recall, which is very low for this model.

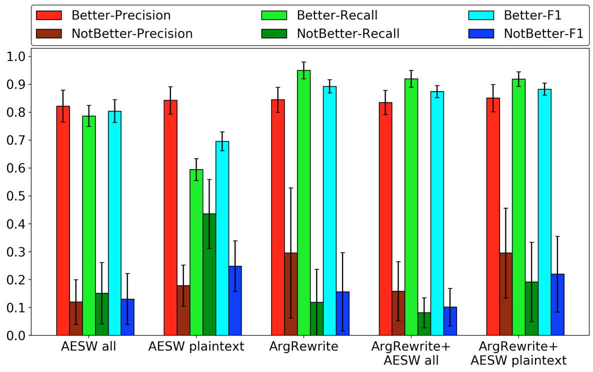

Figure 1: Precision, Recall, and F1 by class label.

To answer Q2, we train only on AESW data but test on the same ArgRewrite folds as above. For both AESW revision samples (before and after removing the placeholders), only Precision is significantly better than the baseline. However, Figure 1 shows that AESW plaintext has significantly higher ($p < 0.05$) Recall than any other model in predicting 'NotBetter' revisions (which motivates Q3 as a way to address the limitation noted in Q1).

To answer Q3, during each run of cross-validation training we inject the AESW data in addition to the 90% ArgRewrite data, then test on the remaining 10% as before. As can be seen from Table 3, AESW plaintext combined with ArgRewrite shows the best classification performance using all three metrics. It also has improved Recall for 'NotBetter' revisions compared to training only on ArgRewrite data. This result indicates that selective extraction of revisions from AESW data helps improve model performance, especially when classifying low-quality revisions.

Finally, to understand feature utility, we compute average feature importance in the 10-folds for each experiment. Top important features include unigrams, trigrams, length difference, language errors, edit distance, BLEU score, specificity difference, and parse-tree features. For example, length difference scores in the top 5 for all experiments. This is intuitive as the annotation guidelines state that adding evidence can make a better revision. Other features such as differences in language errors, specificity scores, and BLEU scores show more importance when training on combined ArgRewrite and AESW data than when training on only ArgRewrite. Surprisingly, spelling error corrections show low importance.

[7] https://nlp.stanford.edu/software/lex-parser.shtml

[8] Random Forest outperformed Support Vector Machines.

| **Original Sentence** ($S1$) | **Revised Sentence** ($S2$) | **Label Distribution** | **Sample Comments** |
|---|---|---|---|
| A 1,000-word letter is considered long, and takes days, if not weeks, to reach the recipient. | A 1,000-word letter is considered long, and takes days, if not weeks, to reach the recipient, with risks of getting lost along the way. | 3 vs 4 | **NotBetter:** S1 is clearer than S2 and the 'risks along the way' could be included as a second sentence to increase readability. **Better:** S2 provides more information that strengthens the idea/major claim in S1. |
| People can't feel the atmosphere of the conversation. | Also, people can't feel the atmosphere of the conversation. | 3 vs 4 | **NotBetter:** Either sentence is fine, but sentence two is not any better. **Better:** Assuming this sentence originally came from the context of a larger part of text, I imagine the continuation included here improves the flow of the original context. |
| With respect to personal life, social networking provides us opportunities to interact with people from different areas, such as Facebook and Twitter. | With respect to personal life, social networkings provide us opportunities to interact with people from different areas, such as Facebook and Twitter. | 1 vs 6 | **NotBetter:** S2 includes incorrect grammar. **Better:** S1 flows better than S2. |

Table 4: Misclassified $NotBetter$ revisions from ArgRewrite along with label distribution ($\#Better$ vs $\#NotBetter$) and sample annotator comments.

## 6 Discussion

Although AESW-plaintext helped classify NotBetter revisions, performance is still low. Table 4 shows some example NotBetter revisions misclassified as Better by most models. The first two examples were also difficult for humans to classify.

In the first example, one annotator for Better (the minority label) points out that the revision provides more information. We speculate that our models might similarly rely too heavily on length and classify longer sentences as Better, since as noted above, length difference was a top 5 feature in all experiments. In fact, for the best model (ArgRewrite+AESW plaintext), the length difference for predicted Better revisions was $4.81$, while for predicted NotBetter revisions it was $-3.99$.

In the second example, one of the annotators who labeled the revision as Better noted that the added word 'Also' indicates a larger context not available to the annotators. This suggests that including revision context could help improve both annotation and classification performance.

The third revision was annotated as NotBetter by 6 annotators. We looked into our features and found that the 'language-check' tool in fact was able to catch this grammatical mistake. Yet only the model using just ArgRewrite for training was able to correctly classify this revision, as all models using AESW data misclassified.

## 7 Conclusion and Future Work

We created a corpus of sentence-level student revisions annotated with labels regarding improvement with respect to argumentative writing.[9] We used this corpus to build a machine learning model for automatically identifying revision improvement. We also demonstrated smart use of an existing corpus of expert edits to improve model performance.

In the future, we would like to improve interrater reliability by collecting expert annotations rather than using crowdsourcing. We would also like to examine how the accuracy of our feature extraction algorithms impacted our feature utility results. Finally, we would like to improve our use of the AESW data, e.g., by automatically clustering revisions for more targeted sampling. Optimizing how many AESW revisions to use and how to balance labels in AESW sampling are also areas for future research.

## Acknowledgments

We thank Prof. Rebecca Hwa for her guidance in setting up a pilot predating our current annotation procedure. We would also like to thank Luca Lugini, Haoran Zhang and members of the PETAL group for their helpful suggestions and the anonymous reviewers for their valuable advice. This research is funded by NSF Award 1550635.

## 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

[9] Freely available for research usage at: http://www.petal.cs.pitt.edu/data.html

*Computational Linguistics and Intelligent Text Processing*, CICLing '11, pages 277–288, Berlin, Heidelberg. Springer Berlin Heidelberg.

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, Avignon, France. Association for Computational Linguistics.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 568–572, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vidas Daudaravicius, Rafael E. Banchs, Elena Volodina, and Courtney Napoles. 2016. A report on the automatic evaluation of scientific writing shared task. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*, pages 53–62.

Johannes Daxenberger and Iryna Gurevych. 2012. A corpus-based study of edit categories in featured and non-featured wikipedia articles. In *Proceedings of the 24th International Conference on Computational Linguistics*, COLING '12, pages 711–726, Mumbai, India.

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*, EMNLP '13, pages 578–589, Seattle, Washington, USA. Association for Computational Linguistics.

The Eli Review. 2014. Eli review, https://elireview.com/support/. [online; accessed 03-18-2018].

Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic extraction of learner errors in esl sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Grammarly. 2016. http://www.grammarly.com. [online; accessed 03-18-2018].

Sara Javanmardi, David W. McDonald, and Cristina V. Lopes. 2011. Vandalism detection in wikipedia: A high-performing, feature-rich model and its reduction through lasso. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, pages 82–90, New York, NY, USA. ACM.

J. Richard Landis and Gary Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, pages 2281–2287.

Trena M. Paulus. 1999. The effect of peer and teacher feedback on student writing. *Journal of Second Language Writing*, 8(3):265 – 289.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Madeline Remse, Mohsen Mesgar, and Michael Strube. 2016. Feature-rich error detection in scientific writing using logistic regression. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*, pages 162–171. Association for Computational Linguistics.

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, MD, USA.

Turnitin. 2014. http://turnitin.com/. [online; accessed 03-18-2018].

The Writing Mentor. 2016. Ets writing mentor, https://mentormywriting.org/, [online; accessed 03-18-2018].

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. Association for Computational Linguistics.

Diyi Yang, Aaron Halfaker, Robert E. Kraut, and Eduard H. Hovy. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages

2000–2010. Association for Computational Linguistics.

Fan Zhang, Homa Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 133–143, Denver, Colorado. Association for Computational Linguistics.