Automated Essay Rewriting (AER): Grammatical Error Correction, Fluency Edits, and Beyond

Masato Mita[1,2,a)] Masato Hagiwara[3] Keisuke Sakaguchi[4] Tomoya Mizumoto[5]
Jun Suzuki[2,1] Kentaro Inui[2,1]

Abstract: We propose the Automated Essay Rewriting (AER) task, where computer systems make automatic edits to argumentative essays to improve their quality. AER subsumes types of edits beyond single sentences such as coherence, cohesion, and style, which are not within the scope of traditional tasks such as grammatical error correction (GEC) and fluency edits. The quantitative and qualitative analyses of a corpus specifically designed for AER reveal that these edits account for almost half of edits made by professional proofreaders. We also discuss the challenges, issues, and future direction of AER by comparing with other tasks.

Keywords: Grammatical Error Correction, Automated Essay Scoring, Proofreading

1. Introduction

The field of grammatical error correction (GEC), which has a multi-decade history, started out with the goal of detecting and correcting targeted error types and providing feedback to ESL (English as a Second Language) learners. Earlier GEC systems focused only on a small number of closed-class error types such as articles [9] and prepositions [8]. The scope of GEC was later expanded to include errors of all types, not only closed-class words, but also verb forms, subject-verb agreement, and word choice errors [11, 17]. While the field of GEC had seen its success as a number of benchmark datasets and shared tasks were enjoyed by the community, Sakaguchi et al., [18] argued thsat over-reliance on error-coded corpora and local edits lead to grammatical yet unnatural sentences. They urged the entire field to revisit the purpose of GEC and focus on improving the overall fluency of sentences.

However, grammar and fluency (or lack thereof) are not the only element when assessing the quality of essays, be it written by native or non-native speakers of the language. As demonstrated by human-written rubrics and automated essay scoring (AES) systems, essays are evaluated holistically based on a number of factors, including their content, coherence, cohesion, style, besides the accuracy of language use and mechanics. Traditional GEC and sentence-level rewriting systems, by definition, are not able to make edits that span over more than one sentence, nor can they model paragraph-level edits that require information outside the sentence in question. As demonstrated later in the paper, such edits account for more than 42% of the total edits made by professional editors to technical papers in our dataset. We believe that the field is ready to expand its focus to include such non-local linguistic phenomena.

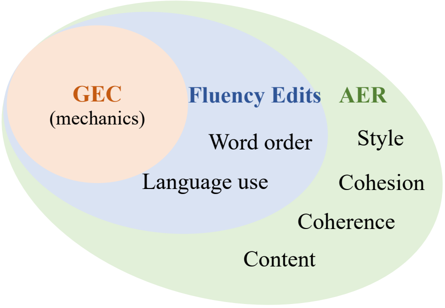

Fig. 1 Scope of GEC, Fluency Edits and AER.

If we shift our focus on to the domain of scientific writing, there has been a growing demand for assisting and automatically correcting argumentative essays. The two rounds of Helping Our Own (HOO) shared tasks [4, 5] aimed to promote the development of tools and techniques that address this demand. Most recently, Automatic Evaluation of Scientific Writing (AESW) shared task [6] was organized with its focus on assisting authors in writing scientific papers. Zhang et al., [21] developed the ArgRewrite corpus, a collection of argumentative essays and their revised drafts, where revisions between drafts are annotated with their purposes. Although the tasks for these two studies were to identify sentences in scientific works that require editing (binary classification) and classifying revision purposes (multi-class classification), respectively, it is straightforward to extend these studies to include automatic editing of sentences or even paragraphs to ensure their fit in the scientific style.

As a confluence of these couple of trends, we propose the Automated Essay Rewriting (AER) task, where the goal is to improve

[1] RIKEN Center for Advanced Intelligence Project
[2] Tohoku University
[3] Octanove Labs
[4] Allen Institute for Artificial Intelligence
[5] Future Corporation
[a)] masato.mita@riken.jp

essays by making automatic edits to them, whether or not they are written by native or non-native speakers of English. Since each revision may improve one or more aspects of the essay, including language use, fluency, cohesion and coherence, and content, the scope of AER is a superset of both GEC and fluency editing. This study focuses on one specific domain—scientific papers—to analyze the types and scopes of edits made by professional proofreaders and motivates the AER task, although it is applicable to any domain.

In this paper, we motivate and describe the new task AER, which addresses cross-sentence phenomena in contrast to with-sentence phenomena which are currently at in the focus of the GEC. Additionally, we report on a study designed to indicate to which extend edits/revisions made by professional editors fall in to the categories of GEC, Fluency Edits and the newly proposed AER. Furthermore, we suggest particular methods and evaluations which may be used to address the new task.

2. Automated Essay Rewiring

First, we define the terminology for the AER task: an *edit* is the smallest unit of operation made to text. It is one of insertion, deletion, substitution, or transposition of one or more contiguous words. A *revision* is an atomic, cohesive change made to an essay in order to improve its quality while preserving its content. A revision may contain more than one edit spanning over more than one sentence. For example, if the subject of a sentence is changed in order to improve coherence, its verb should also be changed so that it agrees with the subject. These two edits constitute one revision.

Now we formally define the task of Automated Essay Rewriting (AER): it is a task where, given a target augmentative essay, computer systems make automatic revisions to it to improve the essay's overall quality measured by an appropriate rubric, while preserving its content (Table 1). Because a typical rubric for augmentative essays measures the quality based on several factors, including mechanics, language use, coherence, and cohesion, all revisions that improve any of these can be the scope of AER. Note that some of these factors also fall under the scope of existing tasks including GEC (grammar and language use) and fluency edits (language use and intra-sentence development), while others do not (e.g., style, content, and coherence). Therefore, the AER task is a superset of GEC and fluency edits. Figure 1 depicts this relationship and the scopes of each task.

3. Corpus Analysis

There still remains a question as to how much of actual change made to augmentative essays falls under the scope of AER as well as the other two traditional tasks. In order to establish AER as a task and assess its scope in an empirical way, we have created and analyzed a corpus of argumentative essays comprised of approximately 100 drafts and their revised versions edited by professional technical proofreaders.

3.1 Corpus Creation

Our goal here is to create a preliminary corpus for the AER task and to analyze the linguistic nature of revisions made to the essays: how many of them can be captured by GEC and fluency edits versus AER? What are the challenges for this AER task compared to the other two tasks? The following factors are considered when creating the corpus:

Original: Community-based Question Answering services, such as Yahoo! Answers, OKWave and Baidu Zhidao, have become popular web services. In these services, a user posts a question and other users answer it. The questioner chooses one of the answers as the best answer. These services have many threads consisting of one question and a number of answers, and the number of threads grows day by day. The threads are stored and anyone can read them. When a user has a question, if there is a similar question in the service, he or she can refer to the answers to the similar question. Herefrom, these services are useful for not only the questioner but also other users having a similar question.
Revised: Community-based Question Answering services, such as Yahoo! Answers, OKWave, and Baidu Zhidao, have become popular web services. As the name suggests, on such services, a user posts a question, other users answer it, and the original questioner selects the best answer. Typically, such services have an increasing number of threads comprising a single question and multiple answers. The threads are stored and are publicly available. If a user posts a question similar to one stored in the system, they can refer to the answers to the stored question.

Table 1 Example of Original Essay and Revised Draft

- The corpus creation design must solicit revisions that are not necessarily contained within single sentences.
- The domain of the corpus must be narrow.

We assume keeping the corpus domain narrow is the key to the task designthe evaluation of AER assumes the existence of (implicit) domain-dependent rubrics, and there could be too many revision possibilities without restricting the domain.

Hence, we focus on the domain of scientific papers. Specifically, we use introduction sections from ACL papers authored by non-native English speakers and treat them as independent "essays". The rationale for using introduction sections is that the discourse structure is generally considered to be more important there than it is in other sections, and it is easier to solicit diverse types of revisions.

For each essay, we collected a complete sets of revisions by a professional editor. The editor was instructed to make whatever changes necessary to the text so that it appears as if it had been written by an English native speaker in the scientific paper domain. In total, we collected 104 pairs of original essays and their revised versions. Table 2 shows the statistics of the collected essays.

3.2 Edit Analysis

In order to clarify what kind of revisions were made, we randomly sampled 15 pairs of original essays and their revised ver-

Num. of original essays	104
Num. of revised versions	104
Num. of paragraphs	631
Num. of sentences	2,287
Num. of words	57,410

Table 2 Statistics of the corpus

Scope	Edit type	Definition	# edit	%
GEC	Mechanics	edits that aimed to fix spelling/grammar mistakes	100	19.1
Fluency Edits	Language use	edits that aimed to increase sentence fluency	175	33.4
	Word order	edits that switched the order of words	29	5.5
AER	Style	edits that aimed to adapt the style	138	26.3
	Cohesion	edits that aimed to make the essay more cohesive	13	2.5
	Content	edits that changed the information of the essay	31	5.9
	Coherence	edits that aimed to increase paragraph fluency	36	6.9
	Others	other types of content revisions	2	0.4

Table 3 Distribution of edit types.

sions and annotated each revision with its type. The annotators were the authors of this paper.

As the annotation scheme, we used a set of eight revision type labels inspired by common essay rubrics used in automated essay scoring (AES) [10, 14, 19]: language use, style, mechanics, coherence, content, word order, cohesion, and others.

Table 3 shows the definition of the eight revision types and their distribution in the annotated corpus. We can confirm that only 19.1% of total revisions are covered by traditional GEC, 38.9% by fluency edits, and 42.0% can be captured only by AER. This result indicates that this corpus contains many AER-specific revisions that require the content knowledge and/or at least some information outside the sentence.

It is noteworthy that, despite the fact that the proofreader was not explicitly instructed to do so, we can observe a number of AER-specific revisions, including ones related to content, coherence, cohesion, and style. Some examples of revisions covered by AER are presented in Table 4.

4. Looking into the Future

As we saw in the previous section, AER covers a wide range of revisions (e.g., style, content, and coherence) in addition to local edits that are covered by traditional tasks. GEC and fluency edits are already challenging enough, both in terms of building systems and evaluating them. What are some challenges we could face if we were to build AER systems and automatically evaluate them? Is it even a problem solvable by automated algorithms? In the remainder of this section, we discuss these challenges in terms of models and evaluation and show some prospects for establishing AER as a task.

4.1 Model

As with earlier GEC systems, it may be beneficial to think of AER systems in term of individual revision types and what models and approaches are effective for each one of them.

For style-related revisions, style transfer techniques [15, 16] may be applied, which transform texts using monolingual sequence-to-sequence models while preserving their meaning given the domain or style (e.g., technical papers and argumentative essays).

On the other hand, other revision types (cohesion, content, and coherence) involve more than one sentence and/or need to consider the global discourse structure. Although sequence-to-sequence models may be sufficient to deal with simple, almost monotonic revisions such as sentence splitting and merging, we don't expect others to be solved by simply applying existing techniques. One possible approach is to use context-aware machine translation (MT) models [7, 20], which are designed so that the information flows from the extended context to the translation model.

As we see, individual revisions can be addressed by existing techniques and extensions, and their holistic combination may establish a good starting point for building a first AER system.

4.2 Evaluation

Automatic evaluation of AER systems can be almost as challenging as, if not more challenging than, that of traditional GEC and fluency edit systems. A common approach to automatically evaluating GEC models is to use reference-based metrics, where gold-standard references are manually created for a given test set and the system output is scored by comparing it with the corresponding references in terms of the metrics, including: MaxMatch [3], ERRANT [2], and GLEU [13]. However, these metrics assume that the set of references are correct and complete, which is far from the truth especially in the context of AER – the number of potential edits in AER is simply too large to be enumerated compared to potential corrections in GEC. A potentially better alternative is to use reference-less metrics that do not require gold-standard references [1, 12].

Ultimately, evaluation of AER systems itself is a research challenge that could be as difficult as building high-quality automated essay scoring (AES) systems. Fortunately, we can borrow from the considerable amount of work done in the AES field and also from the insights gained from the community-driven evaluation effort in, e.g., the Workshop on Machine Translation (WMT).

5. Conclusion

We have proposed the task of AER, where the goal is to improve argumentative essays by making automatic revisions while preserving the content. While the traditional text rewriting tasks, including GEC and fluency edits, focus only on intra-sentential linguistic phenomena, the AER task covers revisions that require content knowledge and/or some information outside the sentence that is being edited. We believe that this task is a natural progression of GEC and fluency edits and that the field is mature enough to expand its focus to this novel, challenging, yet important task.

As mentioned earlier, there are several open research questions that need to be addressed. We seek support from the commu-

Style	
Existing fine-grained entity type classification systems have **used** approaches.	Existing fine-grained entity type classification systems have **employed various** approaches.
Cohesion	
Several studies have examined the canonical word order of Japanese double object constructions, ranging from theoretical studies to empirical **ones** based on psychological experiments and brain science.	Several studies have examined the canonical word order of Japanese double object constructions, ranging from theoretical studies to empirical **studies** based on psychological experiments and brain science.
Content	
However, no previously proposed system has attempted to learn to compose the representations of an entity context recursively.	However, **to the best of our knowledge,** no previously proposed system has attempted to learn to compose the representations of an entity context recursively.
Coherence	
For example, we show actual responses generated by a vanilla seq2seq-based DRG system trained on Twitter conversations in Table 1. **The** responses have inconsistent style as if the system had multiple personalities.	For example, we show actual responses generated by a vanilla seq2seq-based DRG system trained on Twitter conversations in Table 1. **As can be seen, the** responses have inconsistent style as if the system had multiple personalities..

Table 4 Examples of edit types covered by AER.

nity especially in terms of contributing datasets and setting up the benchmark. We invite discussion from all interested parties to establish AER as a community-driven effort.

Reference

[1] Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. "Reference-based Metrics can be Replaced with Reference-less Metrics in Evaluating Grammatical Error Correction Systems". In: *Proceedings of IJCNLP*. 2017.

[2] Christopher Bryant, Mariano Felice, and Ted Briscoe. "Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. July 2017, pp. 793–805.

[3] Daniel Dahlmeier and Hwee Tou Ng. "Better Evaluation for Grammatical Error Correction". In: *Proceedings of NAACL*. 2012, pp. 568–572.

[4] Robert Dale, Ilya Anisimoff, and George Narroway. "HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task". In: *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, 2012, pp. 54–62.

[5] Robert Dale and Adam Kilgarriff. "Helping Our Own: The HOO 2011 Pilot Shared Task". In: *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*. Association for Computational Linguistics, 2011, pp. 242–249.

[6] Vidas Daudaravicius. "Automated Evaluation of Scientific Writing: AESW Shared Task Proposal". In: *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, 2015, pp. 56–63.

[7] Voita Elena et al. "Context-Aware Neural Machine Translation Learns Anaphora Resolution". In: *Proceedings of ACL*. 2018, pp. 1264–1274.

[8] Rachele De Felice and Stephen G. Pulman. "A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English". In: *Proceedings of COLING*. 2008, pp. 169–176.

[9] Na-Rae Han, Martin Chodorow, and Claudia Leacock. "Detecting Errors in English Article Usage by Non-Native Speakers". In: *Natural Language Engineering* 12.2 (2006), pp. 115–129.

[10] Derrick Higgins et al. "Evaluating multiple aspects of coherence in student essays". In: *Proceedings of NAACL-HLT*. 2004, pp. 181–192.

[11] John Lee and Stephanie Seneff. "Correcting Misuse of Verb Forms". In: *Proceedings of ACL-08: HLT*. June 2008, pp. 174–182.

[12] Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. "There's No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction". In: *Proceedings of EMNLP*. 2016, pp. 2109–2115.

[13] Courtney Napoles et al. "Ground Truth for Grammatical Error Correction Metrics". In: *Proceedings of ACL*. 2015, pp. 588–593.

[14] Isaac Persing and Vincent Ng. "Modeling prompt adherence in student essays". In: *Proceedings of ACL*. 2014, pp. 1534–1543.

[15] Shrimai Prabhumoye et al. "Style Transfer Through Back-Translation". In: *Proceedings of ACL*. 2018, pp. 866–876.

[16] Sudha Rao and Joel Tetreault. "Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer". In: *Proceedings of NAACL-HLT*. 2018, pp. 129–140.

[17] Alla Rozovskaya and Dan Roth. "Building a State-of-the-Art Grammatical Error Correction System". In: *Transactions of the Association for Computational Linguistics* 2 (Dec. 2014), pp. 419–434.

[18] Keisuke Sakaguchi et al. "Reassessing the Goals of Grammatical Error Correction: Fluency Instead of Grammaticality". In: *Transactions of the Association for Computational Linguistics* (2016).

[19] Kaveh Taghipour and Hwee Tou Ng. “A neural approach to automated essay scoring”. In: *Proceedings of EMNLP*. 2016, pp. 1882–1891.

[20] Longyue Wang et al. “Exploiting Cross-Sentence Context for Neural Machine Translation”. In: *Proceedings of EMNLP*. 2017, pp. 2816–2821.

[21] Fan Zhang et al. “A corpus of annotated revisions for studying argumentative writing”. In: *Proceedings of ACL*. 2017, pp. 1568–1578.