Template-guided Grammatical Error Feedback Comment Generation

Steven Coyne
Tohoku University, Japan
RIKEN, Japan
coyne.steven.charles.q2@dc.tohoku.ac.jp

Abstract

Writing is an important element of language learning, and an increasing amount of learner writing is taking place in online environments. Teachers can provide valuable feedback by commenting on learner text. However, providing relevant feedback for every issue for every student can be time-consuming. To address this, we turn to the NLP subfield of feedback comment generation, the task of automatically generating explanatory notes for learner text with the goal of enhancing learning outcomes. However, freely-generated comments may mix multiple topics seen in the training data or even give misleading advice. In this thesis proposal, we seek to address these issues by categorizing comments and constraining the outputs of noisy classes. We describe an annotation scheme for feedback comment corpora using comment topics with a broader scope than existing typologies focused on error correction. We outline plans for experiments in grouping and clustering, replacing particularly diverse categories with modular templates, and comparing the generation results of using different linguistic features and model architectures with the original dataset versus the newly annotated one. This paper presents the first two years (the master's component) of a research project for a five-year combined master's and Ph.D program.

1 Introduction

Written corrective feedback on learner text is widespread in language education, and an active area of research in the field of second language acquisition (Kang and Han, 2015). Research has shown that properly administered teacher feedback has a positive effect on language acquisition (Ferris and Roberts, 2001; Bitchener, 2008), including in electronic settings (Ene and Upton, 2014). With the rise of shared online writing environments and e-learning platforms, it has become possible for teachers to assess and comment on learner text digitally. While these advancements in computer-assisted language learning (CALL) are helping revolutionize language education, it remains true that writing frequent and context-appropriate feedback comments on essays is a time-consuming task for teachers. It would be beneficial to provide instructors with automatically generated suggestions when writing comments, allowing them to accept or edit suitable feedback comments and reject unsuitable ones. Using similar technology, it is also possible to provide such feedback comments directly to learners in an intelligent tutoring setting as well. With such use cases in mind, we turn to the task of feedback comment generation.

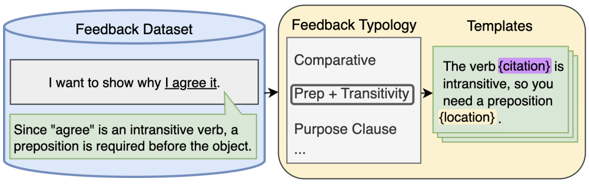
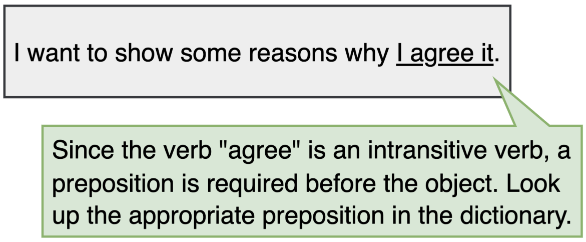

Feedback Dataset
I want to show why I agree it.
Since "agree" is an intransitive verb, a preposition is required before the object.
Feedback Typology
Comparative
Prep + Transitivity
Purpose Clause
...
Templates
The verb {citation} is intransitive, so you need a preposition {location} .

Figure 1: Visualization of the use of manually-labeled feedback comments to support the development of a template-based feedback comment generation system.

In NLP, feedback comment generation is the task of generating hints or explanatory notes for language learners (Nagata, 2019). Data consists of learner sentences, associated feedback comments, and offsets or spans to highlight where the comments were attached to the sentence. An example, taken from the ICNALE Learner Essays with Feedback Comments dataset described in Nagata et al. (2020)[1] can be seen in Figure 2. This is one of a handful of corpora about this task, along with a translated subset used in GenChal 2022 (Nagata et al., 2021) and a separate corpus developed by Lee et al. (2015) and expanded upon in Pilan et al. (2020). The commented ICNALE corpus is fairly small, as seen in Table 1, and a lack of data is

[1] The dataset is available at https://www.gsk.or.jp/en/catalog/gsk2019-b

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 94–104
May 2-4, 2023 ©2023 Association for Computational Linguistics



I want to show some reasons why I agree it.

Since the verb "agree" is an intransitive verb, a preposition is required before the object. Look up the appropriate preposition in the dictionary.

Figure 2: Example of an English learner's sentence with an annotator's feedback comment on a targeted span. Note that feedback comments in the source dataset are written in Japanese, but presented here in English.

one of the major challenges of feedback comment generation.

Additional challenges were revealed by Hanawa et al. (2021) and the participants of GenChal 2022. First, generation is confounded by a many-to-one issue in which multiple comments which ultimately concern the same topic may use different wording. Consider the following pair of sentences:

> **We reached to the station.*
>
> Because the verb "reach" is a transitive verb, the preposition "to" is not required.
>
> **I reached to New York.*
>
> "Reach" is a transitive verb. This verb does not require a preposition prior to the object.

The targeted error is the same, but the comments are superficially different. This diversity can result in mixed generations which are less clear, as shown in Hanawa et al. (2021).

Furthermore, there are a large number of very specific comments relating to particular words and their collocations. In relatively inflexible systems such as the neural retrieval model seen in Nagata (2019), these are rarely output, since the same words would have to occur with the same errors to produce a match. In more flexible generation systems, such comments show a great deal of diversity and contribute to the mixed output problem.

Finally, generation systems can produce inaccurate or misleading comments which can lead learners astray, as reported by Hanawa et al. (2021). It is important to constrain these false generations, which can have a negative learning effect or reduce confidence in the system.

This research seeks to improve the generation of educationally effective feedback comments by addressing the above challenges. We outline plans to group feedback comments with a set of annotations which focus on the "topic" of each comment, based on its communicative purpose and its connection to an issue in the sentence when applicable. We identify highly variable or noisy feedback comment categories and replace such categories with modular templates. We also describe experiments with textual features and generation architectures to be used in testing the effects of the above approaches. It is hoped that these contributions can enable additional research into feedback comment generation for language learning.

2 Related Work

Pedagogical feedback comments have long been studied in the field of education, including in the context of language learning. There is considerable debate about what kind of feedback works best and why, which includes dimensions such as directness (Ferris and Roberts, 2001), presence of metalinguistic terms (Bitchener, 2008), and hedging (Baker and Hansen Bricker, 2010). While there are some detractors (Truscott, 1996), written feedback has generally been found effective for language learning (Kang and Han, 2015).

Turning towards the online environment of our task, we must consider systems which already exist. There are various tools for grammatical error correction (GEC) and writing assistance, perhaps the most notable of which is Grammarly[2]. We define the purpose of these tools as *writing assistance*, in which the goal is to improve the content of the document. This overlaps with, but is distinct from, the purpose of this work, which we define as *learning assistance*. Our goal is to help learners notice and understand their errors, not just correct them. Defining and suggesting changes in sentences is a necessary step in the process, but it is done with an eye towards a long term learning effect. The generation goal is therefore different. In our case, it is acceptable if we do not produce a comment for every error in the sentence, since we prioritize precision to avoid misleading students, and because a large number of overlapping and uncoordinated feedback comments can overwhelm and demotivate students (Lee, 2013). On the other hand, a GEC or writing support system like Grammarly ideally has something to offer for all issues in a sentence. We also place more emphasis on explaining what is wrong and particularly why, rather than

[2] https://www.grammarly.com/

Dataset Information	General	Preposition	Combined
Sentences	43568	28829	72397
Feedback Comments	26592	5693	32285
Commented Sentences	19991	4931	24922
Comment/Sentence Ratio	0.459	0.171	0.344
Most Comments/Sentence	14	6	14

Table 1: Information about the "ICNALE Learner Essays with Feedback Comments" dataset. It is divided into two sub-corpora, one with comments on general topics, and the other focusing on preposition use.

what specific edits should be made, which the comments in this task often hint at rather than provide outright. This would quickly become frustrating in a writing support environment, but in the context of education, such comments have been found to be effective for long-term learning (Sheen, 2007; Bitchener and Knoch, 2010).

In the context of natural language processing, feedback generation was first formally defined in Nagata (2019), followed by the release of a dataset containing learner sentences and feedback comments, constructed from essays from The International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2013). The dataset contains triples consisting of English sentences written by learners, feedback comments added by professionals, and the offsets designating the target span of the comment. There were also developments by Lai and Chang (2019), who created a system which constructed feedback templates from word collocations, and Pilan et al. (2020), who added additional annotations to a corpus of textual revisions (Lee et al., 2015) and investigated the revision outcomes of various kinds of feedback comments.

Hanawa et al. (2021) performed experiments on the ICNALE feedback dataset, revealing challenges faced by models using neural retrieval, simple generation, and retrieve and edit (Hashimoto et al., 2018) architectures. Namely, the retrieval model can not generalize, the retrieve-and-edit model over-edited in an unconstrained manner, and the simple generation and retrieve-and edit models both produced mixed or misleading outputs. Following that, there was a shared task on feedback comment generation in GenChal 2022. Teams demonstrated various modeling and preprocessing techniques, particularly that it is possible to extract detailed linguistic features from the sentences and comments using existing NLP tools such as parsers and GEC systems and use them to enhance feedback comment output.[3] We discuss several such options in section 3.4.

3 Research Plan

Based on the above literature, we have identified two major challenges in this task:

1. Superficial diversity of comments. For a given error, there are any number of ways to describe or explain it, and any number of ways to phrase a suggestion. This manifests as superficial differences among multiple comments that are effectively the same in meaning, presenting a challenge when counting or classifying feedback comments. Such comments could be grouped into one category or "topic." Consider the examples below:

> **It should be a clean places for service everyone that comes to have in there.*
>
> The preposition "for" cannot be followed by the base form of a verb. Use a to-infinitive instead.
>
> **Sometimes students from the outside city will do this for earn some money.*
>
> The preposition "for" indicating the purpose of the sentence are followed by nouns including gerunds. You cannot put a verb in its original form after a preposition. Use to-infinitive to indicate the purpose.

The learner sentences have little in common beyond the presence of "for + base form of a verb," and the feedback comments are different in length and detail. The latter contains several words such as "gerunds" and "purpose" which the former lacks. There is also a difference in terminology: "base form" vs. "original form." It is possible that these differences may cause the two comments to be treated somewhat differently by models. It is ideal if they can be assigned to the same group.

[3] Participants' systems can be viewed on the GenChal 2022 website: `https://fcg.sharedtask.org/links/`

2. Unreliability of outputs. In the context of educational feedback comments, it is very important to provide accurate advice. Misleading outputs can confuse learners and lead them astray with false information. Furthermore, noticing inaccurate comments can erode trust in the system. The following example is taken from Hanawa et al. (2021):

> **I disagree <u>to</u> you.*
>
> Since the verb "disagree" is a transitive verb, the object does not require the preposition "to."

In this case, the comment is incorrect because "disagree" is in fact an intransitive verb. Furthermore, instead of removing "to," we should replace it with "with." The model may have found the correct word to change, but suggests both the wrong operation and the wrong reasoning. This example is also a case where the correction relies on a specific word's use and collocations. It would be beneficial to identify which kinds of comments are particularly likely to face these issues, and address them in a targeted manner. This would first necessitate some form of grouping the comments.

3.1 Feedback Topic Tagging

To address these points, we decided to manually tag all sentence-comment pairs in the dataset with a "topic." This is distinct from error typing, since it must include a broader scope to encapsulate what an instructor's comment is about. This can include comments on more abstract issues in learner text.

Currently, NLP tools can identify errors and predict an edit, but not necessarily describe the underlying rule. Just using the edit information to generate feedback might give us a comment that applies to the target text span in some way, but not necessarily match the advice we want to give, especially if we want to comment on something with a broader scope. Consider the following example:

> *If I <u>will</u> have chance, I must do part time job.*
>
> In the if clause "if... then", we do not use the auxiliary verb "will" to express the future. In the if clause, let's express the future with the present tense of the verb.

The correction is to change the verb's tense from future to present, but the reason is more complex, relying on a conditional clause. The topic of the comment could be thought of as "conditional." However, if we consider existing NLP frameworks for grammatical correction, we only find much more local categories. The most popular error typology in NLP is ERRANT (Bryant et al., 2017), which compares erroneous sentences and their corrections. ERRANT would characterize this as "U:VERB:TENSE," which does not take the broader picture of a conditional clause into consideration. GECToR (Omelianchuk et al., 2020), a sequence tagger which predicts grammatical corrections, tags this as "$DELETE," addressing only the edit operation of removing "will."

This is no indictment of these systems - indeed, the scale of the errors they consider is reasonable for the tasks they were designed for. Rather, we highlight the difference in scope between GEC and the present task, in which it is desirable to include broader structures in our analysis. Therefore, we seek to include information such as "conditional" in our labels. Furthermore, since many GEC tags can be returned automatically by present tools, it is prudent to include complementary information, and we thus use a set of categories which do not always focus on the same phenomena. These labels are based on the "topic" of the feedback, i.e. what the comment is about. In the very common case where a feedback comment targets an error, these topic labels will often overlap with error typologies, but they include broader-scoped perspectives of the errors which extend beyond the level of edit operations, perhaps focusing on the learner's attempted grammatical pattern (e.g. "conditional"). They also incorporate major types of teacher feedback which are not sufficiently covered by automatic systems, such as redundancy, parallelism, transitions, run-on sentences, fragments, tone, and idiom errors.

The current tags are presented in section 3.1.2. These were developed by first consulting previous typologies (see section 3.1.1), considering which categories are most likely to be used by English language teachers,[4] then checking them against the comments in the ICNALE feedback dataset, adapting to the data in the corpus. As a first step, we considered a subset consisting of of 250 sentences each from the General and Preposition sub-corpora, each with exactly one feedback comment. Comments extending across multiple sentences were excluded, since they exceed the sentence-level scope of this work. Sentences were then sampled with a particular random seed. The proposed tag set may

[4] The author worked in English education for five years, and drew on that experience in the process.

evolve further by the time all sentences in the data have been considered.

3.1.1 Existing Tag Systems

When designing the annotation system, consideration was given to existing tag sets from the fields of NLP, corpus linguistics, and second language acquisition (SLA).

The error typology used in the NUCLE dataset (Dahlmeier et al., 2013) contains 28 tags. Some examples work well for this task, such as redundancy, which may not be identified as a grammatical error per se, but which a critical teacher may certainly comment on. However, it has some categories which are too broad in some cases, such as Mec, which concerns spelling, punctuation, and capitalization, among others. These would have quite different feedback comments. There are also some very distinct subcategories of each topic, which may warrant more granular labels.

The system used in the Cambridge Learner Corpus (Nicholls, 2003) and seen in the First Certificate in English (FCE) dataset (Yannakoudakis et al., 2011) is more fine-grained, and contains detailed descriptions of various errors. It is also modular, with edit type (whether words are missing, unnecessary or should be replaced) as well as the relevant part of speech. It has 77 tags, making it quite expressive. Some of the tags are for quite rare or esoteric errors. Additionally, we find this system too linguistically oriented for this task, its original purpose being to describe a corpus of errors rather than the topic of teacher feedback.

ERRANT was created with both of the above in mind, and strikes a good balance between them, having quickly become the standard for GEC research. However, it too was created for the task of GEC specifically, and thus its error tags do not extend to the broader topics seen in the educational realm as ours do. This is understandable, because it is simply a tool for another (albeit related) field.

Meanwhile, educational researchers have also been considering learners and their errors, creating some typologies of their own. Error analysis studies such as Watcharapunyawong and Usaha (2013) and Darus and Subramaniam (2009) tend to use very broad categories, often with no section outlining their reasoning for them. These may be too broad to use for this task.

The most direct source of educational feedback topics may be the various sets of error code annotations used by English teachers. These tend to include many of the more abstract categories we wish to address, such as redundancy, parallelism, and idiom. While there does not seem to be a well-accepted correction code standard in literature, there are a variety of systems shared online, many covering similar topics. One example is the system used for writing programs at the University of California, Irvine (UCI Writing Center, 2008).

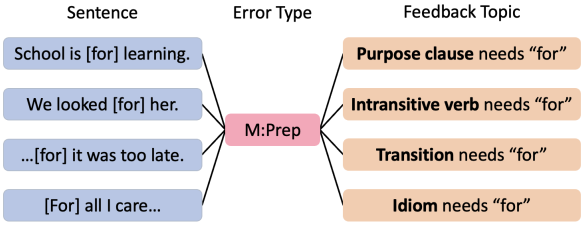


Figure 3: One grammatical error type (as identified by ERRANT) can be associated with several different underlying reasons, each with distinct comments.

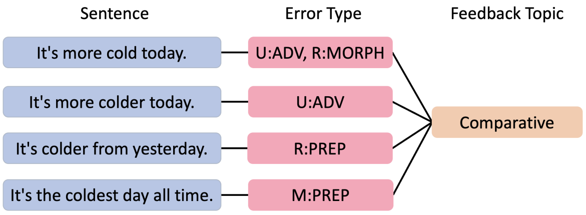


Figure 4: A variety of error types may be associated with a common attempted construction. "Grammatical Pattern" feedback topics seek to model this kind of broader phenomenon seen in learner errors.

3.1.2 Proposed Tag System

The proposed system is shown in Tables 2 and 3. The tags are divided into three levels of abstraction. The most concrete are "Operational" tags, which reflect direct changes to one or a few words in the text. These are expected to correlate very closely with existing error typologies. Examples include punctuation, spelling, and "missing noun." There are cases where this kind of straightforward word-level edit to the text is indeed the best summary of a feedback comment's content.

The next level of abstraction we call "Grammatical Patterns." These are essentially designed as a teacher's perspective of the violated "grammar point" that underlies the writer's error. They can thus serve to summarize a large portion of comments that target errors in an educational setting. If compared to GEC error types or the operational tags, these are expected to display complex mapping behaviors with many-to-one and one-to-many relationships, as demonstrated in Figures 3 and 4.

Tags at the highest level of abstraction are appropriately called "Abstract" tags, which may map to "any" or "none" of the theoretical errors in a sentence. An example is "unclear," which teachers apply to certain sentences which can display any of a vast variety of issues. Praise and complex rewrites are also in this category, as are comments pointing out language transfer. Specialized approaches may be necessary to best generate these comments, if a system's designers intend to include them at all.

3.2 Grouping Comments

Once comments have been classified by topic, it will be possible to run a variety of NLP tools on the dataset and explore the co-occurrence of their outputs with each of the tags. These include sequence taggers for parts of speech and dependencies as well as error correction systems. If it is discovered that some feedback comment types correlate very strongly with certain parse patterns or GEC error types, those system outputs may be useful as predictive features for the feedback comments.

Returning to the proposed use cases, it may be desirable in educational settings to focus feedback on a limited number of categories, or to adapt the system to the learner's level using a framework such as the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) or even specific curriculum goals laid out by a school board. This would also allow customization by users. Categorizing the comments is a useful first step towards realizing such options.

Additionally, it will be easier to explore automatic clustering once the comments have human-annotated categories. Nagata and Hanawa (2021) attempted to address the superficial diversity issue by clustering the comments with textual similarity, but the interpretability of the resulting categories is limited. It would be useful to compare such results to class labels added by a human.

In addition to surface similarity, we will experiment with clustering based on semantic similarity or with a topic-modeling approach as seen in Grootendorst (2022). Topic labels can be identified in the feedback comment text, and placed into hierarchical clusters. Given that there are many synonyms for grammatical terms in the feedback comments, we hypothesize that semantic or topic modeling will perform better than surface similarity. We will compare clusters to the manual tags, potentially revealing additional topic subtypes which can help improve the tagging logic.

Furthermore, it will be interesting to observe whether the clustering and classification strategies described above can generalize to other feedback comment data. If so, it may suggest that the strategies are sound. If not, useful observations may result which could suggest improvements to either the tagging logic or the application of these techniques. Any dataset with pairs of learner sentences and associated feedback comments made by humans can be a candidate. Presently, the only other suitable dataset we are aware of is the one described in Lee et al. (2015) and expanded in Pilan et al. (2020). We may additionally create our own dataset of feedback on learner sentences as part of future work on this topic, as noted in section 4.

3.3 Templates

To address the superficial diversity issue, we seek to replace the outputs of highly diverse comment categories with generalized templates. The manual tagging step will allow us to identify the feedback categories most in need of such attention. Tentatively, it seems that there are a large number of comments which contain content very specific to a single word, pair, or triple, often taking a form like the following:

> We do not use **«a»** with **«b»** to express "meaning of collocation." Think of an alternative **<(part of speech of a)>**.

Comments like these form a long tail of rare examples in the dataset, and the data may simplify significantly if they are unified into a limited number of semi-automatically generated templates with slot-filling. The slots can be filled with words from the sentence and information from open-source lexical resources. This can also help with the reliability challenge in this task, since we can more tightly control the output in these cases, and filter candidate comments if they do not contain certain words present in the original sentence.

The word designated **«b»** above is likely to prove hardest to handle. It is a non-erroneous word being combined erroneously in the original sentence. Lai and Chang (2019) call this is the "problem word," and Nagata and Hanawa (2020) call it the "attachment word." A collection of collocations or other lexical resources may be necessary to determine this word and its relationship to others in the sentence or its theoretical correction. It will also

Abstract Tags

Tag Name	Example
Fragment	**Obligation at home and at campus.**
Idiom	[**There's** → **That's**] the way it goes.
Language Transfer	I like riding [**jet** → **roller**] coasters.
Praise	(Various kinds of praise and encouragement)
Rewrite	(Used for explicit, complex revision suggestions)
Tone	It's maybe [**cause** → **because**] my work experience less than other people.
Unclear	**If home is not richness economically, everybody is only just doing it.**

Grammatical Pattern Tags

Tag Name	Example
Comparative	Maybe you will study [**more hard** → **harder**] in the class.
Causative	It will ruin our concentration and make everything [**getting**] worse.
Conditional	If I [**have** → **had**] a job, I could buy more things.
Dummy Subject	It is important [**that**] university students [**have**] a part time job.
Derivation	Due to the time, we lived in a [**peace** → **peaceful**] world.
Hyphenation	It is important for students to have a [**part time** → **part-time**] job.
Modal/Auxiliary	Students [**would better** → **should**] have part-time jobs.
Nominalization	[**Breathe** → **Breathing**] fresh air is important.
Noun Countability	Also, they can buy other [**stuffs** → **stuff**].
Parallel Structure	...hanging out with my best friend, [**buy** → **buying**] cosmetics, or shopping
Participle	In some restaurant, we can see students [**works** → **working**] as waiters.
Passive Voice	As a result, their performance in school may be [**get**] influenced.
Possessive	Studying is the main task [**to** → **of**] students.
Preposition + Transitivity	I completely agree [**with**] this opinion.
Purpose Clause	They should earn money [**for** → **to**] spend in the daily life by themselves.
Quantifier	Almost [**all**] non-smokers hate the cigarette smoke.
Question Formation	Why [**students must** → **must students**] do part time job[**.** → **?**]
Redundancy	I did part-time jobs last summer vacation to [**go travel**] to a foreign land.
Relative Clause	College students [**who**] jump in part-time job have a variety of reasons.
Run-on Sentence	In a word, I'll try[**,** → **.**] if I find a job fit me, I'll do that!
Subject-Verb Agreement	The [**students works**] part time job
Transitions	[**But** → **However,**] it costs a lot to go to the university.
Word Order	**What more serious is...** → **What is more serious...**

Table 2: Annotation System for Feedback Comment Topics, Abstract Tags and Grammatical Pattern Tags.

sometimes be necessary to refer to other words in the sentence which are not necessarily erroneous in order to explain the relative position of a suggested operation such as insertion. We will call such words "reference words."

Creating templates also allows us a chance to rewrite their contents to be more suitable to the task. For example, it may be ideal to limit the amount of direct citation which takes place, particularly for the meaning of collocations, which may be difficult to extract in a reliable manner. In addition to this, we find that many of the comments in the commented ICNALE dataset have fairly advanced grammatical explanations, which can be simplified to help learners understand them. An example of a modular feedback comment template with such revision can be seen in Figure 5.

3.4 Generation Experiments

After tagging, grouping, and template composition is complete, we move on to experiments with generation models. The experiments performed by teams in GenChal 2022 show that it is possible to enhance generation performance using a variety of supplemental features obtained from the data. Systems of interest include GECToR, which tags sequences with edit operations, as well as parsing trees which use recent strategies to specialize on erroneous text. These include the SynGEC system (Zhang et al., 2022), which can output special tags for words which are missing or which should be rewritten, as well as a parser trained on the Tenbusu Treebank (Morgado da Costa et al., 2022), which incorporates "mal-rules", specialized rules which match ungrammatical structures, allowing the parser to describe erroneous text.

There are additional systems to consider as well, with the caveat that they require corrected versions of the sentences. These include the aforementioned ERRANT as well as SERRANT (Choshen et al., 2021), a more recent addition which incorporates additional tags focused on syntax errors,

Operational Tags

Tag Name	Example
Capitalization	In [**korea** → **Korea**], it is common.
Incorrect/Double Negative	If smoking [**not be** → **is not**] banned, a lot of people will smoke.
Missing Adjective	Almost [**all**] restaurant in Japan have smoking seat.
Missing Adverb	And [**when**] they can get right answer, I feel very happy.
Missing Determiner	They will relax after having [**a**] meal.
Missing Noun	For students who don't have money, [**jobs**] are very necessary.
Missing Preposition	70% [**of**] men in this country is smoking
Missing Pronoun	Try to tell them what [**they**] should do, and what [**they**] should not to do.
Missing Verb	Some of them can not [**pay**] their education fees.
Noun Number	College students have a lot of [**times** → **time**].
Other	(Miscellaneous Topics)
Punctuation	They can learn the value of money[**,**] they use, too.
Replace Adjective	It 's [**interested** → **interesting**] to me .
Replace Adverb	I have [**ever** → **never**] been in this situation.
Replace Determiner	Second, they can know [**an** → **the**] importance of money.
Replace Noun	I will talk about my [**opinion** → **reason**] why.
Replace Preposition	I have three reasons [**about** → **for**] it.
Replace Pronoun	They need work for them or [**they** → **their**] family .
Replace Verb	It [**does** → **is**] important and helpful when taking a job.
Spacing	Customers [**may be** → **maybe**] don't want to go that restaurant again.
Spelling	[**The** → **They**] will make good use of the money.
Unnecessary Adjective	And it will be very [**important**] worthwhile in life.
Unnecessary Adverb	I feel bored every time [**when**] someone smokes near me.
Unnecessary Determiner	Nowadays it is [**a**] common for college students to have a part-time job.
Unnecessary Noun	Students have burden on a lot of assignments and expensive tuition [**fee**].
Unnecessary Preposition	Many students had a part-time job because they need [**to**] money.
Unnecessary Pronoun	I have acquaintances that [**he**] died from smoking.
Unnecessary Verb	Many of people [**are**] get a part time job for many reasons.
Verb Conjugation	Smoking [**are** → **is**] very popular these days.
Verb Form	How about [**give** → **giving**] sometime to think yourself.
Verb Tense	Most students [**are** → **were**] isolated from society before.

Table 3: Annotation System for Feedback Comment Topics, Operational Tags

outlined in Choshen et al. (2020). CEFR-J (Ishii and Tono, 2018), a framework for describing the features and proficiency level of learner text, offers a set of scripts to find grammatical items with regular expressions. Some of these are quite complex, and may be able to indicate broader grammatical structures in the dataset sentences. For example, CEFR-J scripts can recognize both "not as large" and "larger than" as part of a comparative group of tags starting with "COMP." These are promising as predictive features for a "comparative" topic tag.

We hypothesize that the tags output by these systems, especially those found to correlate with particular feedback topics, can provide useful information to language models when incorporated into input sequences. We can obtain these features by first creating a corrected version of the input sentence via an automatic tool such as GECToR or a generative model. We then apply SERRANT to the sentence-correction pair to obtain error annotations, and apply CEFR-J scripts to the corrected sentence to obtain grammatical item matches.

To assess whether the above strategies are effective for feedback comment generation, we will repeat the experiments from Hanawa et al. (2021), using simple generation, neural retrieval, and retrieve and edit models and assessing the differences associated with our changes. Given more time, we will move on to larger and more recent language models suited for generation tasks, such as GPT-2, T5, and BART to examine whether and how performance improves given the additional features.

4 Future Directions

This paper presents details about the first two years (the master's component) of a research project for a five-year combined master's and Ph.D program. There are many additional research directions and concepts which can be incorporated before the final thesis is complete.

For generation, given the mixed output issue observed in previous studies, and the overlapping nature of the tagging system, it may be prudent to separate generation models. Different models could be used for abstract comments versus the operational and grammatical pattern ones. If a more abstract comment is generated, the grammatical

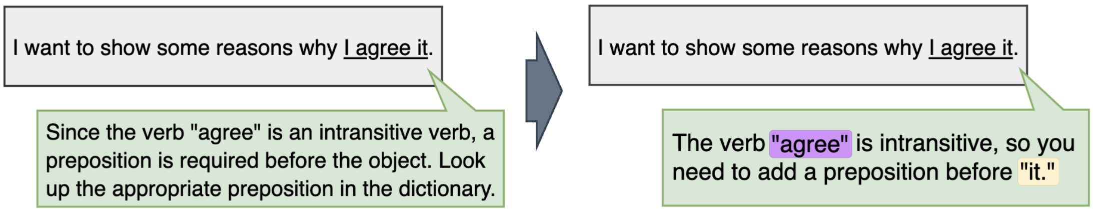

I want to show some reasons why I agree it.
Since the verb "agree" is an intransitive verb, a preposition is required before the object. Look up the appropriate preposition in the dictionary.
I want to show some reasons why I agree it.
The verb "agree" is intransitive, so you need to add a preposition before "it."

Figure 5: Example of a feedback comment rewritten as a template and simplified. This example template has two slots to fill with citations, a "problem word" whose mistaken combination caused the error ("agree"), and a "reference word" for indicating a position for a suggested operation ("it").

pattern and operational models can withhold suggestions. It may also be effective to separate the generators in a more horizontal manner, with some focusing on prepositions, others on verb errors, and so on, likely informed by the results of clustering analysis. A separate model would classify the errors and call the relevant generator.

There is also a general lack of corpora for the task of feedback comment generation. Given that each teacher has their own idiosyncrasies in correcting learner text, it is highly desirable to collect more data from a variety of writers. Furthermore, the ICNALE Learner Essays with Feedback Comments dataset contains only essays written by learners from China, India, Japan, Korea, Thailand, and Taiwan, and the number of comment writers is limited. The learner's CEFR levels range from A2 to B2, the ages range from 15 to 37, and all writing is in the context of a single-draft essay. Working with corpora with different learner first languages, age groups, or writing tasks may further affect the annotation sets and clusters discussed in this work, as well as provide valuable training data in the form of new and unique sentence-comment pairs, particularly for categories such as language transfer. We have preliminary plans to construct a new corpus of learner sentences, feedback comments, and comment topic labels, which will be informed by the insights gained during this research.

5 Summary

To assist in the task of feedback comment generation, we add manual labels to the feedback-enhanced ICNALE dataset which consider broad-scope errors, explore grouping comments using these manual labels as a reference, craft modular templates for highly diverse categories of feedback comments, and perform modeling experiments with a variety of architectures and using features obtained by parsers and GEC tools, reporting on the best combinations.

Limitations

Only sentence-level errors and comments are considered in this proposal. A separate body of work, automated essay scoring, addresses paragraph and document level writing issues. Extending feedback comment generation to that scope is left for future research. Both are useful for the intended settings of online essay grading and intelligent tutoring systems, so it would be ideal to see them connected.

The proposed tags are ultimately manual, so data from any new corpora must be tagged by hand as well if it is to align with this work.

There are some cases where the new tags offer little more than existing automatic tools, particularly for the operational annotations. Furthermore, some may question whether we need another tagging system for learner essays and their issues, especially after ERRANT was introduced to unify disparate systems such as the Cambridge Learner Corpus and NUCLE. Again, this is because this data, and thus the proposed tags, are focused on learner support, not grammatical error correction or writing support, and are meant to describe the topic of a comment and its link to an error rather than the local features of the error itself. Additionally, they are meant to complement existing error-focused systems such as ERRANT or GECToR, and therefore provide information from a slightly broader context which can be used to identify additional kinds of issues.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP22H00524 and JP21K21343. We thank Kentaro Inui, Diana Galvan, Keisuke Sakaguchi, Michael Zock, and Ryo Nagata, who shared their thoughts at various stages of this project. We also thank the reviewers of the 2023 EACL Student Research Workshop, whose insightful criticisms led to several improvements to this paper.

References

Wendy Baker and Rachel Hansen Bricker. 2010. The effects of direct and indirect speech acts on native english and esl speakers' perception of teacher written feedback. *System*, 38(1):75–84.

John Bitchener. 2008. Evidence in support of written corrective feedback. *Journal of Second Language Writing*, 17(2):102–118.

John Bitchener and Ute Knoch. 2010. Raising the linguistic accuracy level of advanced l2 writers with written corrective feedback. *Journal of Second Language Writing*, 19(4):207–217.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. 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)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Leshem Choshen, Dmitry Nikolaev, Yevgeni Berzak, and Omri Abend. 2020. Classifying syntactic errors in learner language. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 97–107, Online. Association for Computational Linguistics.

Leshem Choshen, Matanel Oren, Dmitry Nikolaev, and Omri Abend. 2021. SERRANT: a syntactic classifier for english grammatical error types. *CoRR*, abs/2104.02310.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Saadiyah Darus and Kaladevi Subramaniam. 2009. Error analysis of the written english essays of secondary school students in malaysia: A case study. *European Journal of Social Sciences*, 8:483–495.

Estela Ene and Thomas A. Upton. 2014. Learner uptake of teacher electronic feedback in esl composition. *System*, 46:80–95.

Dana Ferris and Barrie Roberts. 2001. Error feedback in l2 writing classes: How explicit does it need to be? *Journal of Second Language Writing*, 10(3):161–184.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. 2021. Exploring methods for generating feedback comments for writing learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9719–9730, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tatsunori B. Hashimoto, Kelvin Guu, Yonatan Oren, and Percy Liang. 2018. A retrieve-and-edit framework for predicting structured outputs. In *Neural Information Processing Systems*.

Yasutake Ishii and Yukio Tono. 2018. Investigating japanese efl learners' overuse/underuse of english grammar categories and their relevance to cefr levels. In *In Proceedings of Asia Pacific Corpus Linguistics Conference 2018*, pages 160–165.

Shin'ichiro Ishikawa. 2013. The icnale and sophisticated contrastive interlanguage analysis of asian learners of english. *Learner corpus studies in Asia and the world*, pages 91–118.

Eun Young Kang and ZhaoHong Han. 2015. The efficacy of written corrective feedback in improving l2 written accuracy: A meta-analysis. *The Modern Language Journal*, 99:1–18.

Yi-Huei Lai and Jason Chang. 2019. TellMeWhy: Learning to explain corrective feedback for second language learners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 235–240, Hong Kong, China. Association for Computational Linguistics.

Icy Lee. 2013. Research into practice: Written corrective feedback. *Language Teaching*, 46(1):108–119.

John Sie Yuen Lee, Chak Yan Yeung, Amir Zeldes, Marc Reznicek, Anke Lüdeling, and Jonathan James Webster. 2015. Cityu corpus of essay drafts of english language learners: a corpus of textual revision in second language writing. *Language Resources and Evaluation*, 49:659–683.

Luís Morgado da Costa, Francis Bond, and Roger V. P. Winder. 2022. The tembusu treebank: An English learner treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4817–4826, Marseille, France. European Language Resources Association.

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215, Hong Kong, China. Association for Computational Linguistics.

Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. 2021. Shared task on feedback comment generation for language learners. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 320–324, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Ryo Nagata and Kazuaki Hanawa. 2020. Bunpo ayamari kaisetsubun seisei to wa dono yona tasuku na no ka? [what kind of task is grammatical error commentary generation?]. In *Proceedings of the 26th Annual Conference of the Association for Natural Language Processing*, pages 513–516.

Ryo Nagata and Kazuaki Hanawa. 2021. Kasotekina ayamari taipu no wariate ni yoru kaisetsubun seisei no seino kojo [improving the performance of commentary generation by assigning virtual error types]. In *Proceedings of the 27th Annual Conference of the Association for Natural Language Processing*, pages 679–684.

Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. 2020. Creating corpora for research in feedback comment generation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 340–345, Marseille, France. European Language Resources Association.

Diane Nicholls. 2003. The cambridge learner corpus: Error coding and analysis for lexicography and elt. *Proceedings of the Corpus Linguistics 2003 conference*, page 572–581.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Ildiko Pilan, John Lee, Chak Yan Yeung, and Jonathan Webster. 2020. A dataset for investigating the impact of feedback on student revision outcome. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 332–339, Marseille, France. European Language Resources Association.

Younghee Sheen. 2007. The effect of focused written corrective feedback and language aptitude on esl learners' acquisition of articles. *TESOL Quarterly*, 41(2):255–283.

John Truscott. 1996. The case against grammar correction in l2 writing classes. *Language learning*, 46(2):327–369.

UCI Writing Center. Correction symbols for uci writing programs [online]. 2008.

Somchai Watcharapunyawong and Siriluck Usaha. 2013. Thai efl students' writing errors in different text types: The interference of the first language. *English Language Teaching*, 6:67–78.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. 2022. Syngec: Syntax-enhanced grammatical error correction with a tailored gec-oriented parser. In *Proceedings of EMNLP*.