Towards grammatically-informed feedback comments

Diana Galvan-Sosa[1,2] Steven Coyne[1,2] Keisuke Sakaguchi[1,2] Kentaro Inui[1,2]
[1]Tohoku University [2]RIKEN
{dianags,keisuke.sakaguchi,kentaro.inui}@tohoku.ac.jp
coyne.steven.charles.q2@dc.tohoku.ac.jp

Abstract

Current writing assistants are good in error correction and in helping users to change ungrammatical sentences into their correct grammatical form. However, they still fall short on various dimensions, in particular error justification. While the current systems are useful when the main goal is expression, they are insufficient when the goal is the acquisition of a writing skill. It is clear that finding the root of an error is key for improvement. The question is how to do this automatically? We present here an approach that automatically aligns error annotations with grammatical-category annotations made on grammatical-ungrammatical sentence pairs. Our preliminary results suggest that such alignments provide a good hint concerning the specific grammar points a user should pay attention to.

1 Introduction

Writing is not an easy task. Whether we produce written text in our mother tongue (L1) or a foreign language (L2), the task is daunting because of the number of sub-tasks involved, and because of the lack of clear decision criteria. For example, when is a text optimal? What is a coherent text?, etc. Broadly speaking, writing requires three major steps: (1) idea generation (2) idea ordering and (3) linguistic expression. Assuming that the two first steps have been performed, we need to find the right words, put them in the right order and make the needed morphological adjustments. In other words, having decided what to talk about, and how to convey our ideas, we must make sure that the final output complies with the rules of language. Hence, a basic, yet very important aspect of this last step is to check the grammaticality of our sentences. Realizing the grammar-checking can be automated has motivated the development of a number of writing assistants, one of the most prominent ones being Grammarly[1)].

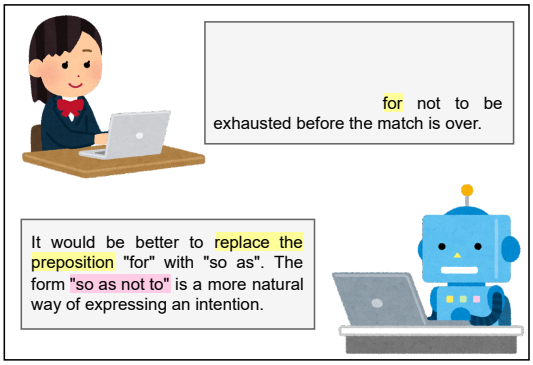

Figure 1 Example of a sentence generated by a L2 user and the expected feedback (output) from the writing assistant.

Ideally, the goal of modern writing assistants is not just to check grammar, but to check the quality of all the levels involved in writing. Put differently, ideally revision concerns the entire writing process. According to cognitive scientists, *revision* is a complex problem-solving process composed of various steps: (i) *problem detection*, (ii) *problem diagnosis* and (iii) *solution generation* [1]. If we are only interested in the correct form of a text, then it is acceptable for the writing assistant to only output the result of (i) and (iii). Yet, if the goal is the enhancement of the student's writing skills, the output of (ii) is very important. Note that composition, i.e., learning to write is a very common goal among L2 users. Consider the sentence shown in Figure 1, where the detected problem (hereafter referred to as *error*) is highlighted. The purpose of highlighting the preposition is not only to signal that it should be REPLACED (for → so as), but also to make the author aware of the fact that the *"so as not to"* form was used incorrectly. The goal of this diagnosis is not only to signal the error, but to explain it. Hence, the user's error is related to a *goal*, which includes the use of the preposition *"to"* and other GRAMMATICAL FORMS like *"so as to"*, *"in*

1) https://www.grammarly.com/

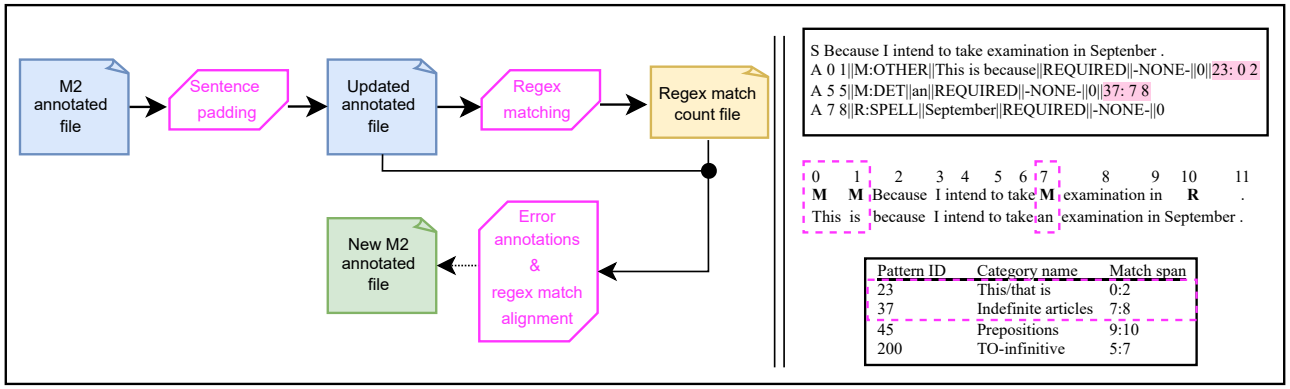

Figure 2 Overview of our sentence alignment pipeline (left). To the right, a sample sentence from the M2 file (top), the same sentence padded and aligned (middle) and the matched grammatical categories (bottom). The new annotations are highlighted at the top.

order to" and *"so that"*. This is what the user needs in order to avoid making the same mistake in the future.

The aforementioned diagnosis could easily be made by an English teacher. The question is: how could a writing assistant do the same thing? Thanks to the research done on Grammatical Error Correction (GEC), there is a fair amount of learner data with *error annotations* available. However, these annotations are limited to the identification of the *type* (e.g., replace a preposition) and the location of an error in the text. Our premise is that if the sentence were also annotated with its related grammatical categories, these annotations would provide a hint concerning the reasons causing the error.

We present here an algorithm that adds *grammatical categories* to existing error annotations. To do so, we rely on the English Grammar Profile defined by the CEFR-J [2] project.

2 Related work

Grammatical Error Correction. GEC is the task of detecting and correcting all kinds of errors in a sentence. Note that error diagnosis is neither expected nor required. There have been several attempts on tackling GEC, mostly led by the Building Educational Applications (BEA) 2019 Shared Task [3]. The BEA dataset[2], which includes annotations on an error's type, location and its correction, is the current benchmark measuring the performance of GEC systems. The dataset is composed of 4 corpora: the First Certificate in English (FCE) corpus [4], Lang-8 [5, 6], the National University of Singapore Corpus of Learner English (NUCLE) [7] and W&I+LOCNESS [3, 8].

2) Available at https://www.cl.cam.ac.uk/research/nl/bea2019st/#data

Feedback Comment Generation (FCG). Unlike GEC, the goal of FCG is to diagnose an error and output an explanatory note. This task was recently proposed by [9] as the GenChal 2022 shared task, together with a dataset[3] that pairs ungrammatical sentences with a feedback comment.

The ultimate goal of our work is directed towards FCG. We aim to generate feedback comments that not only explain the underlying grammatical rule of an error, but also suggest whether the misuse of other grammatical categories caused it. As a first step, we explore to what extent the grammatical categories identified in a grammatical sentence relate to the error annotations made on its ungrammatical counterpart.

3 Aligning errors with grammatical categories

3.1 The CEFR-J English Grammar Profile

The Common European Framework of Reference (CEFR) is a learning framework that describes the knowledge and skills needed by a learner to communicate in English. The CEFR includes an English Grammar Profile (EGP) that lists the grammatical forms and meanings a learner is expected to get familiar with as they progress along the learning curriculum. This framework has been widely adopted not only in Europe, but also in Latin America and Asia. Its adaptation in Japan (the CEFR-J) is of particular interest to us, as it defines a finer-grained EGP that includes 501 patterns across 263 grammatical categories. An example is shown in Table 1.

3) Available at https://fcg.sharedtask.org/data/

GramCat ID	GramCat name	Pattern	Pattern ID
8	This/That is	This/That is	23
		This/That is not	24
		Is this/that..?	25
		Isn't this/that..?	26

Table 1 Example of a grammatical category (GramCat) and its associated forms (patterns).

The CEFR-J EGP is available online, together with a set of scripts that identify all the patterns in a text using regular expressions (regex).[4)]

3.2 Error spans and regex matches

As a source of learner's data, we use the BEA and GenChal 2022 datasets introduced in Section 2. Data in the BEA dataset has already been standardized with ERRANT [10], the most common error type framework used in GEC. The top-right of Figure 2 shows an example of ERRANT's M2 annotation format, which identifies a token-based *error span*, an *error type* and an *edit text* (i.e., the text to correct the identified error). As a FCG dataset, GenChal 2022 does not use the M2 format, but it is possible to standardize as long as we input pairs of (UNGRAMMATICAL,GRAMMATICAL) sentences. Since this dataset **does not include** the GRAMMATICAL counterpart of an UNGRAMMATICAL sentence, it needs to be created. For the scope of this paper, we annotated an initial subset of 500 sentences out of the total 5000.

With both datasets in M2 format, we rely on the CEFR-J scripts for adding grammatical category information to the M2 annotations. These scripts take as input a grammatically correct text and output a *count* file with two columns: one with the 501 Pattern IDs and another with the count of the sentences that matched each pattern. The scripts were modified to also output *which* sentences matched a pattern and the span of the match.

3.3 Alignment process

Padding. Apart from *error categories* like "DET" (Determiner), ERRANT classifies errors depending on whether tokens need to be *inserted*, *deleted* or *substituted*. These operations are referred to as Missing ("M"), Unnecessary ("U") and Replacement ("R"), respectively. As shown at the top-right of Figure 2, an UNGRAMMATICAL sentence usually needs more than one edit operation to become GRAMMATICAL. As a result, most of the time these sentences have different lengths. In order to align the error spans in the UNGRAMMATICAL sentence with the regex matches found on the GRAMMATICAL one, we need both to have the same length. As a general rule, "M" and "U" operations lead to padding the ungrammatical and grammatical sentence, respectively. In the case of "R" operations, no padding will be needed if the token-length of the *edit span* and the *edit text* are the same. The ungrammatical sentence will need padding if the length of the *edit text* >*edit span*. The grammatical sentence will be padded otherwise.

Alignment. We define as "alignment" when an *edit span* and a *match span* overlap. The bottom-right of Figure 2 shows the four patterns found in the GRAMMATICAL sentence by the CEFR-J scripts. However, there are only two alignments: "M:OTHER" → "This/that is" and "M:DET" → "Indefinite articles". In the case of the second alignment, the grammatical category matched is consistent with the error category that was already annotated. The first alignment, though, adds new information. The "M:OTHER" error type identifies that some tokens are missing, but is not able to point out what exactly the error is (therefore, the use of the generic "OTHER" category). The grammatical category identified in the alignment is more informative, suggesting that the root problem is related to *how to point to something*. In this particular case, using the *"This/that is"* form is a better than *"because"* to detail a *reason*.

When an alignment is found, a new token with the format PATTERNID: OVERLAP SPAN is appended at the end of the corresponding *error annotation* line. Note that the sample sentence in Figure 2 has modifications to its first two annotations, while the third one remains unchanged. This is because an *edit span* can either be aligned with one, several or no *match span* at all.

4 Data analysis

In order to get a better understanding of the potential benefits of aligning an error's *edit span* and a grammatical category *match span*, we first look at the number of alignments found in the BEA dataset and our subset from GenChal2022. The results are summarized in Table 2. Considering that a sentence can have one or more error annotations, the table shows both the total number of sentences in a dataset and the error annotations. Note that

4) http://cefr-j.org/download.html#cefrj_grammar

Dataset	Sentences	Error annotations	Alignments	1 alignment	>1 alignment
FCE	33,236	64,638 (noop: 11,967, edits: 52,671)	22,496	18,380 (81.7%)	4,116 (18.3%)
LANG-8	1,037,561	1,940,760 (noop: 539,858, edits: 1,400,902)	520,708	384,519 (73.8%)	136,189 (26.2%)
NUCLE	57,151	79,798 (noop: 35,316, edits: 44,482)	17,330	13,792 (79.6%)	3,538 (20.4%)
W&I+LOCNESS	37,704	82,684 (noop: 12,319, edits: 70,365)	27,090	22,397 (82.7%)	4,693 (17.3%)
GenChal2022	543	1,844 (noop: 1, edits: 1,843)	943	839 (89.0%)	104 (11.0%)

Table 2 Number of alignments by dataset. Sentences can have 1 or more error annotations. Here, we take into account ALL the error annotations in a dataset. A "noop" annotation indicates that a sentence has no errors.

sentences in a dataset can be GRAMMATICAL or UNGRAMMATICAL. A sentence is annotated even if it is judged as grammatical, in which case the *error type* is "noop". The coverage of alignments found on the error annotations (i.e., edits) ranges from 37% to 42% in the BEA benchmark and 50% in our subset from GenChal2022. Taking a closer look at the total number of alignments, we observe that most of the times, our algorithm found one alignment between an *error span* and a *match span*. It is interesting to see that there were some cases were a *error span* overlapped with more than one *match span*. An example of both situations is shown in Figure 3.

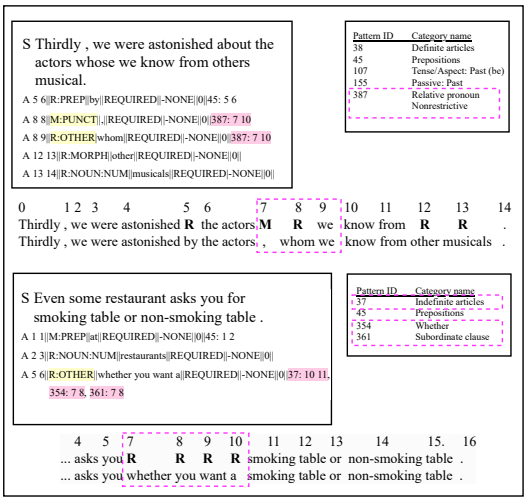

Figure 3 Example of two sentences and their alignments (top: 1 alignment, bottom: >1) found in their error annotations.

Although we expected that the greater the number of alignments, the more informative a *match span* would be, our preliminary observations suggest that single alignments can also be revealing. The example shown at the top of Figure 3 showcases a situation where two edits are aligned with the same grammatical category. The edits simply point out that the sentence is missing a comma and that *"whose"* should be replaced with *"whom"*. The reasoning behind these edits is that a removable idea, known as a NONRESTRICTIVE CLAUSE, is being introduced by the RELATIVE PRONOUN. Therefore, the comma is necessary. The alignments made on the bottom example explain a more elaborated error. The main problem with the sentence is the use of the "for" preposition. While the use of the collocation *"ask for"* to introduce a *request* is correct, the request in question concerns two possibilities, in which case it is better to use the SUBORDINATE CONJUNCTION "whether". This is related to the use of a SUBORDINATE CLAUSE, which is in fact, being suggested in the *edit text* (i.e., *whether you want a*). Besides the use of "whether", the suggested subordinate clause involves the use of INDEFINITE ARTICLES. In summary, the *error type* is helpful for identifying and correcting the error. The alignments found by our algorithm complement the error annotation with information related to the *error diagnosis*.

So far, it seems that our alignment approach is a good direction towards a "teacher-like" error diagnosis, which in turn is a step towards the generation of grammatically-informed feedback comments. The ongoing qualitative analysis presented here is being performed drawing on the second author's experience in English education.

5 Conclusion and Future Work

AI is evolving at an amazing pace. The development of AI-powered writing assistants is, beyond any doubt, one of the most notable contributions to make the writing process less painful for authors. However, there is still work to be done in order to bridge the gap between human and computer-based writing revision. In this work, we explore the plausibility of automatically identifying the grammar forms that a user needs to know in order to avoid making the same grammatical error in the future. Since the BEA dataset does not include feedback comments, our future work only considers the GenChal 2022 dataset. Our immediate next step is to design an experimental setup that will quantitatively evaluate the usefulness of error's *edit spans* and grammatical categories' *match spans* on feedback comment generation.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP22H00524 and JP21K21343.

References

[1] Kwangsu Cho and Charles MacArthur. Learning by reviewing. **Journal of educational psychology**, Vol. 103, No. 1, p. 73, 2011.

[2] Yasutake Ishii and Yukio Tono. 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**, pp. 160–165, 2018.

[3] Christopher Bryant, Mariano Felice, Øistein E Andersen, and Ted Briscoe. The bea-2019 shared task on grammatical error correction. In **Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 52–75, 2019.

[4] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 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**, pp. 180–189, 2011.

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

[6] Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. Tense and aspect error correction for esl learners using global context. In **Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 198–202, 2012.

[7] Daniel Dahlmeier and Hwee Tou Ng. 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**, pp. 568–572, 2012.

[8] Sylviane Granger. The computer learner corpus: a versatile new source of data for sla research. In **Learner English on computer**, pp. 3–18. Routledge, 2014.

[9] Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. Shared task on feedback comment generation for language learners. In **Proceedings of the 14th International Conference on Natural Language Generation**, pp. 320–324, 2021.

[10] Christopher Bryant, Mariano Felice, and Edward Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. Association for Computational Linguistics, 2017.