Evaluation of Really Good Grammatical Error Correction

Robert Östling[1], Katarina Gillholm[2], Murathan Kurfalı[3], Marie Mattson[4], Mats Wirén[5]
[1]Department of Linguistics / Stockholm University / robert@ling.su.se
[2]Department of Linguistics / Stockholm University / gillholm.ina@gmail.com
[3]Psychology Department / Stockholm University / murathan.kurfali@su.se
[4]marie.mattson18@gmail.com
[5]Department of Linguistics / Stockholm University / mats.wiren@ling.su.se

Abstract

Traditional evaluation methods for Grammatical Error Correction (GEC) fail to fully capture the full range of system capabilities and objectives. The emergence of large language models (LLMs) has further highlighted the shortcomings of these evaluation strategies, emphasizing the need for a paradigm shift in evaluation methodology. In the current study, we perform a comprehensive evaluation of various GEC systems using a recently published dataset of Swedish learner texts. The evaluation is performed using established evaluation metrics as well as human judges. We find that GPT-3 in a few-shot setting by far outperforms previous grammatical error correction systems for Swedish, a language comprising only about 0.1% of its training data. We also found that current evaluation methods contain undesirable biases that a human evaluation is able to reveal. We suggest using human post-editing of GEC system outputs to analyze the amount of change required to reach native-level human performance on the task, and provide a dataset annotated with human post-edits and assessments of grammaticality, fluency and meaning preservation of GEC system outputs.

Keywords: grammatical error correction, computer-assisted language learning, natural language generation

1. Introduction

Grammatical Error Correction (GEC) is typically used in an extended sense of correcting language at multiple levels, including spelling errors, grammatical errors, word choice and idiom usage.

In the literature on evaluating GEC systems, one is rarely explicit about the purpose of the system. Following Sakaguchi et al. (2016), we see two somewhat different objectives:

1. **Error detection and correction**, where grammaticality has priority over fluency. The goal is to point out individual language errors, which could ideally be fixed one by one, resulting in an acceptable text that is as close as possible to the original.

2. **General text improvement**, where fluency is on equal footing with grammaticality. The goal is to produce a text which is as close as possible to what a highly proficient writer would have produced, assuming a perfect understanding of the intended message of the original text.

The distinction between the two objectives is less clear for writers at high proficiency levels, where changing an occasional spelling or grammar mistake typically results in a high-quality text. For a less proficient writer, a text may contain so many overlapping problems that it is difficult to identify local changes that together result in a high-quality text. If a GEC system is allowed to work directly at the level of general text improvement, its task may become significantly simpler.

The choice of objective has practical implications for how to evaluate the result. Traditional methods for GEC evaluation are reference-based, where either the GEC system output is compared to a human-created reference (e.g. Napoles et al., 2015), or the sets of edit operations produced by the GEC system is compared to those needed to transform the original text to the human reference (Bryant et al., 2017).

One important problem with reference-based evaluations is that there is typically a large and varied set of possible ways to express the same information. It is generally infeasible to approximate the full set of possibilities, although providing multiple references is a common approach in the machine translation community to alleviate this problem (e.g. Qin and Specia, 2015). Results are also highly dependent on the way the references were created. Freitag et al. (2020) show that biases due to "translationese" effects in the creation of references negatively affect the accuracy of the resulting evaluations, where interference from the source language may affect the translation to become less idiomatic in the target language. They obtained higher agreement between the automatic reference-based evaluations and human judgments by first asking human annotators to maximally paraphrase the reference sentences, to encourage diversity among the multiple references.

The references used for GEC evaluations suf-

LREC-COLING 2024, pages 6582–6593
20-25 May, 2024.

fer from the same bias, and often annotators are explicitly instructed to stay as close to the original text as possible (Volodina et al., 2019, Section 6.1). We are aware of no GEC evaluation data which, in the style of Freitag et al. (2020), aims for a high amount of diversity in the references. This has the effect of biasing existing automatic evaluations against systems that perform paraphrasing rather than conservatively fixing individual errors.

In the related field of text summarization, Goyal et al. (2022) found that automatic evaluation metrics severely underestimate the performance of large language models, further strengthening our suspicion that such powerful models necessitate a paradigm shift in evaluation methodology. In this work, we perform a comprehensive manual analysis of the output of multiple GEC systems, and point towards analysis of human post-edits as the most promising way of evaluating really good GEC systems.

2. Related work

2.1. Reference-free evaluation metrics

Yoshimura et al. (2020), building on earlier work by Asano et al. (2017), propose a reference-free metric named SOME that learns a weighting of grammatically, fluency, and semantic similarity scores obtained from three separate models. Interestingly, they find that tuning these weights on a dataset of human judgements results in 98% of the weight being put on the fluency score computed as the difference between language model cross-entropy for the system output and original text. This result aligns well with the argument of Sakaguchi et al. (2016) that fluency is what GEC system ought to aim for.

Islam and Magnani (2021) go one step further and dispose of the grammaticality and semantic similarity models, and simple use language model scores combined with a filter based on string similarity measures to reject "corrections" that look too different from the original.[1] System-level Pearson correlations with human scores are very high for all these systems: from about 0.88 (Yoshimura et al., 2020) to 0.98 (Asano et al., 2017). However, metrics that rely heavily on language model scores do not handle semantic changes well. As Islam and Magnani (2021) point out, the SOME metric assigns a very positive score when "He is going school." is corrected into "He He He He He He."

More recent work on reference-free metrics includes that of Maeda et al. (2022), who generate partially corrected sentences from parallel data of original and corrected sentences, which they then train a neural scorer on.

2.2. Human evaluation

Grundkiewicz et al. (2015) performed a comprehensive ranking-based human evaluation of current (at the time) GEC systems, and compared the human rankings to a number of automatic metrics. In their study, the edit distance-based MaxMatch (M^2) score (Dahlmeier and Ng, 2012) with a bias towards precision ($\beta = 0.18$) achieved the highest correlation with the human rankings. Pure machine translation metrics (BLEU, METEOR) were found to have *negative* correlations with human rankings. Napoles et al. (2015) also performed a ranking-based human evaluation, and obtained very similar results to Grundkiewicz et al. (2015). However, in this case their proposed GLEU metric achieved a higher correlation with the human rankings than did the M^2 score. In both studies, correlations were relatively modest, with Spearman ρ and Pearson r in the order of 0.7–0.75 (Grundkiewicz et al., 2015) and 0.55 (Napoles et al., 2015) for the highest-correlated metrics. Náplava et al. (2022) performed a similar human evaluation of more recent Czech GEC systems, and found the agreement between human and GEC rankings to be very high, with Pearson's r in the range 0.95–0.98 for GLEU, M^2, I-measure and their Czech adaptation of ERRANT (Felice et al., 2016; Bryant et al., 2017).

Rather than ranking system outputs akin to machine translation evaluation, Yoshimura et al. (2020) applied a range of different GEC systems to the same set of sentences and obtained human absolute scores in the dimensions of grammaticality, fluency, and meaning preservation. This allows GEC systems to be compared using each dimension individually or in combination.

In work concurrent to ours, Wu et al. (2023) performed a human evaluation of ChatGPT (based at the time on GPT-3.5) for English GEC. Their evaluation found that ChatGPT had overall better results than the other systems evaluated, and was characterized by few under-corrected parts (high recall) but a tendency to rewrite text more often, resulting in over-corrections (low precision). However, unlike our work, their evaluation did not investigate what effect this had on text meaning.

2.3. Co-evolution of systems and metrics

As Yoshimura et al. (2020) demonstrated, it is possible to obtain a metric with an extremely strong correlation (up to $\rho = 0.98$) to human ratings, by optimizing the parameters of the metric with respect to human ratings. However, their best result was obtained by relying almost exclusively on the fluency score. This indicates that their data does not

[1] The paper somewhat unusually refers to the string similarity measures as "syntactic similarity".

contain enough examples of GEC-“corrected” sentences that are scored high by a language model, but suffer from problems like a low degree of meaning preservation. Since their data is based on the outputs of existing GEC systems, it inherits their biases towards certain types of mistakes. Most of these systems have been developed against metrics such as GLEU and ERRANT, who tend to reward a conservative approach where precision is prioritized over recall and more substantial rewriting is discouraged. A GEC system optimized for high scores according to GLEU, ERRANT, or similar metrics, is thus less likely to suggest major changes that (when incorrect) may significantly alter the meaning of a text. The main dimension along which the performance of such systems vary is to what extent they can find and correct simple spelling, grammar or word choice errors – the presence of which correlates strongly with a poor language model score.

We hypothesize that the co-evolution of GEC systems and their evaluation metrics has resulted in reinforcing the bias towards certain types of properties, namely a conservative approach which avoids paraphrasing. Traditionally this has not been much of an issue, since we did not have particularly good models for paraphrasing non-standard text. With the advent of large language models that excel at this task, we argue that it is time to break this circle of GEC system development and metric development.

2.4. Swedish GEC

We now briefly review published GEC systems for Swedish, focusing on general-coverage methods that perform automated correction. We do not cover methods specializing in specific error types, like spelling or collocations, or those that are not able to automatically suggest corrections.

Granska (Domeij et al., 2000) is a mostly rule-based system for grammatical error detection and correction, which has later been combined with a probabilistic model (Bigert and Knutsson, 2002) that uses a language model to score variants of the input sentence. Another contemporary rule-based system based on Constraint Grammar (Karlsson et al., 1995) was developed by Birn (2000). More recently, Nyberg (2022) implemented Swedish versions of the following two methods. First, the model of Bryant and Briscoe (2018), which is based on generating variants of the input sentence and using a language model to choose the highest-scoring one. Second, using a neural machine translation model trained on artificially corrupted data, generated in a fashion similar to that of Grundkiewicz et al. (2019).

3. Purpose and aims

Most previous work on GEC evaluation has been performed using relatively limited and conservative systems, and we see a need to extend this line of work to systems based on large language models (LLMs) that are able to perform more substantial corrections than previous methods. Given the problems pointed out above with both reference-based and reference-free automated evaluation metrics, we think it is important to consider manual evaluation methods in addition to automated ones. Finally, since LLM-based systems are approaching human-level performance, we also want to include text versions created by humans in the evaluation, on equal footing with the automatic GEC systems in order to ensure a fair comparison.

Our main contributions in this work are the following:

- We evaluate a set of GEC systems, belonging to diverse paradigms ranging from rule-based systems to large language models. In addition to automatic GEC systems, we also include in the evaluation paraphrases from humans following two different guidelines.
- We investigate how different evaluation methods, automatic and manual, compare across this range of different GEC paradigms, and point out relative strengths and weaknesses of each paradigm.
- We use human post-edits of GEC system outputs and human paraphrases, both to evaluate the systems and to quantify the differences between the final versions of each text after correction/paraphrasing and post-editing.
- We make several new resources publicly available: Human evaluations and post-edits of GEC system outputs, detailed annotation guidelines, a novel annotation tool that was used to produce the above, and a baseline GEC system for Swedish.[2]

4. GEC systems

In this work we compare a total of five Swedish GEC systems:

1. **Granska**: the web API version of the rule-based system Granska (Domeij et al., 2000). We always accept its top suggestion for changes, but multiple suggestions that change the same span are rejected.

[2] Our code and data is available at: `https://github.com/robertostling/gec-evaluation`

2. **Nyberg MT**: the MT-based method of Nyberg (2022, Section 3.3), training a neural machine translation system to translate artificially corrupted text back into its original form.

3. **Nyberg LM**: the LM-based method of Nyberg (2022, Section 3.4), using a language model to iteratively score local edits suggested by a heuristic procedure, until no further changes sufficiently improve the score.

4. **MT**: The MT-based method of Kurfalı and Östling (2023), which in turn is a development of Nyberg MT with more data, a modified method for introducing synthetic errors and a different architecture.

5. **GPT-3**: OpenAI's (`text-davinci-002`) model (Brown et al., 2020) through their public API. We use a two-shot prompt, with authentic learner sentences taken from the CrossCheck corpus[3] and manually corrected by us. The prompt is entirely in Swedish, and is identical for all processed sentences.

We also consider the dummy baseline **Uncorrected**, which simply leaves the text unchanged, and the following three human-corrected versions:

1. **Human minimal**: human-normalized sentences from the SweLL project (Volodina et al., 2019). Annotators were asked to perform minimal edits in order to produce a grammatically correct sentence while trying to preserve the meaning; for normalization guidelines, see Rudebeck et al. (2021). The annotators had access to the full context for each sentence.

2. **Human fluent**: human-normalized sentences produced by having a native Swedish annotator, different from the other annotators in the project, edit the **Human minimal** sentences to achieve native-like fluency while staying as close as possible to the original.

3. **Human free**: human-normalized sentences produced by having a native Swedish annotator, different from the other annotators in the project, edit the **Human minimal** sentences to achieve native-like fluency while being encouraged to change the sentence as much as needed to achieve the most idiomatic way of expressing the given meaning.

5. Data and annotation

We use Swedish data from the SweLL project (Volodina et al., 2019), which consists of 502 learner texts collected from different levels of L2 Swedish education. The texts are annotated with an approximate CEFR level, and have been manually normalized by minimally editing them into a grammatically correct version. We use the sentence segmentations and the division into a test and a development set from Nyberg (2022). For the systems **Nyberg MT** and **Nyberg LM**, we report evaluation results from Nyberg (2022).

	Gramm.	Fluency	Meaning
Round 1	0.45	0.69	0.51
Round 2	0.84	0.81	0.71

Table 1: Quadratically weighted kappa (QWK) between the two annotators during the two-round pilot phase, for grammaticality, fluency and meaning preservation.

Two independent annotators, both co-authors of this paper and native Swedish speakers, were tasked with performing the following procedure for each corrected sentence in two pilot datasets, containing ten sentences each:

1. In a text box, read the system's output and if necessary modify it to reach the level of a native writer, considering both fluency and grammaticality. The annotators are instructed to perform the minimum amount of editing to reach this goal.

2. When the (possibly) edited system output is submitted, the existing human-normalized reference from the SweLL data is shown (**Human minimal**), and the annotator is asked to confirm whether the meaning of the edited sentence matches the reference. If the annotator thinks otherwise, the human reference is hidden again and the tool returns to step 1.

3. When the edited system output is accepted, the annotator is shown the learner sentence, the non-edited system output, and the SweLL reference. Then the annotator chooses a score on a 4-level Likert scale (or "other") for the three dimensions of grammaticality, fluency and meaning preservation. We follow Yoshimura et al. (2020) for the definition of the scales for each of these three dimensions. Our annotation tool and full guidelines are publicly available. [4]

Sentences were randomized from a pool containing the outputs of all systems under evaluation, and a custom-made annotation tool was used for the task. After discussions between the annotators and

[3] https://www.csc.kth.se/tcs/projects/xcheck/korpus.html

[4] https://github.com/robertostling/gec-evaluation/tree/main/annotator

after the two pilot datasets, the level of agreement was high enough (QWK in the range 0.71–0.84) to allow one annotator to continue annotating the full data. Details are presented in Table 1. The full dataset was subsampled from the development set of Nyberg (2022) and contains 64 sentences each from CEFR proficiency levels A, B and C, for a total of 192 sentences. Each sentence has been processed by three GEC systems and two human paraphrasers, for a total of $192 \times 5 = 960$ output sentences with post-edits and scores.

As the final result, we have for each sentence produced by a GEC system: (a) scores for grammaticality, fluency, and meaning preservation; (b) a post-edited version of the output with the minimal edits required to obtain maximum scores on grammaticality, fluency and meaning preservation.

In order to achieve comparability with previous work on Swedish GEC, we adapted the sentence-level train/test split of Nyberg (2022), and continued performing all analysis on the sentence level. The only instance where a wider context is used, is for the minimal normalization from the SweLL project data. These normalized versions are used as references, but since GEC systems have access to less context than the reference was based on, models that are able to take longer contexts into account are put at a disadvantage during evaluation. In several cases, it is even impossible to correctly and unambiguously interpret the sentence without further context. Only the human corrections, which are directly or indirectly based on the full context, do not suffer from this problem. We have manually inspected all cases of "moderate" or "substantial" differences in meaning for the annotations of the best-performing GEC system's (GPT-3) output, and found that 8 out of 28 (29%) such sentences require further context. While this is a methodological problem to be addressed in future work, we see that the impact of this problem is limited.

An example of this problem can be seen in Table 2, which contains all twelve versions of one specific sentence and also serves to illustrate how our annotated data looks. The student's original sentence could be interpreted in two ways with regards to the purpose of leaving the house: either to be "left alone" or to "be free". The SweLL project annotator (Minimal) has access to the full student text and chose the former interpretation, so according to our annotation guidelines this is the interpretation to which meaning preservation should be compared. In this case, the student's ambiguity and lack of context unfairly penalizes the neural systems (MT, GPT-3) with respect to meaning preservation, since they both produce the other interpretation ("be free"), which is perfectly reasonable. Again, this problem could be remedied by giving all systems and annotators access to the same (wide) context, and to the original student sentence. The rule-based Granska system is correctly penalized for producing a nonsensical version of this phrase that substantially changes the meaning of the sentence. The humans (Human free, Human fluent) follow the interpretation in the Minimal version, and so obtain perfect meaning preservation scores. The fluent version contains a paraphrase which, according to the annotator, makes the sentence more native-like and this version is the only one with a perfect score in all dimensions.

6. Results

6.1. Automatic evaluation metrics

Table 3 shows the performance of each system using the reference-based GLEU metric, while Table 4 contains the corresponding evaluation using the reference-free Scribendi score (Islam and Magnani, 2021).[5] Both metrics yield the same ranking of the systems: GPT-3 scores best, followed by the NMT systems, followed in turn by the rule-based system. However, the relative differences between the systems differ considerably between the metrics. In particular, for the Scribendi score (Table 4) we see a very sharp divide between the neural and the non-neural systems. For all different levels, GPT-3 in fact scores higher than the human reference, even though its output contains a substantial amount of errors (as shown in the human evaluation, see Table 5). This is not very surprising, since the Scribendi score mainly represents the number of sentences where the Swedish GPT-SW3 model (Ekgren et al., 2022) assigns a higher score to the system output than to the original sentence, and the GPT-3 output was obtained from the same family of language models.

6.2. Human evaluation

The result of the human evaluation is summarized in Table 5. As for grammaticality and fluency, the ranking of the GEC systems is identical to that of the automatic metrics. In both cases, GPT-3 performs at or near human levels. The MT-based system follows, at a considerable distance, and the rule-based system scores last. The trend is identical across CEFR proficiency levels. For meaning preservation, the situation is different. Here, there are no major differences between systems, all of them consistently perform below human levels. The gap to the human paraphrases is highest for the B level sentences, which contain the most

[5] Since figures for **Nyberg MT** and **Nyberg LM** are taken from Nyberg (2022), which only reports GLEU, they are missing from Table 4.

Version	Text	G	F	M
Translation	Suddenly I think that I must leave this house in order to be [*left alone/free*] and live in peace			
Student	Plötslig tencke jag måste gå ifrån det har huset jag vill bli fre [unclear meaning] och leva i fred .			
Minimal	Plötsligt tänker jag att jag måste lämna det här huset , jag vill vara i fred [*be left alone*] och leva i fred .			
Granska	Plötslig tencke jag måste gå ifrån det har huset jag vill bli före [*become before*] och leva i fred .	1	1	0
Granska+post	Plötslig tänker jag att jag måste lämna det här huset , jag vill vara i fred [*be left alone*] och leva i fred .			
MT	Plötsligt måste jag tencke gå ifrån det här huset , jag vill bli fri [*be free*] och leva i fred .	2	2	1
MT+post	Plötsligt tänker jag att jag måste lämna det här huset för jag vara i fred [*be left alone*] och leva i fred .			
GPT-3	Plötsligt tänkte jag att jag måste gå ifrån det här huset . Jag vill bli fri [*be free*] och leva i fred .	4	3	1
GPT-3+post	Plötsligt tänkte jag att jag måste lämna det här huset . Jag vill vara i fred [*be left alone*] och leva i fred .			
Human fluent	Plötsligt tänker jag att jag måste lämna det här huset . Jag vill vara i fred och leva i fred .	4	3	4
Fluent+post	Plötsligt tänker jag att jag måste lämna det här huset . Jag vill vara i fred och leva i fred .			
Human free	Plötsligt tänker jag att jag måste komma iväg från [*get away from*] det här huset , jag vill vara i fred och leva i fred .	4	4	4
Free+post	Plötsligt tänker jag att jag måste komma iväg från det här huset , jag vill vara i fred och leva i fred .			

Table 2: Scores for (G)rammatical, (F)luency and (M)eaning preservation for selected examples from the annotated data. Underlined words in the student sentence contain word-level errors.

		CEFR level		
System	All	A	B	C
Uncorrected	0.44	0.29	0.17	0.53
Granska	0.47	0.35	0.24	0.55
Nyberg MT	0.51	0.42	0.30	0.58
Nyberg LM	0.52	0.42	0.32	0.58
MT	0.57	0.48	0.38	0.63
GPT-3	0.63	0.60	0.52	0.65
Human minimal	1.0	1.0	1.0	1.0

Table 3: Reference based evaluation: GLEU scores on the test set of Nyberg (2022).

		CEFR level		
System	All	A	B	C
Uncorrected	0	0	0	0
Granska	0.03	0.08	0.11	-0.01
MT	0.51	0.57	0.68	0.43
GPT-3	0.69	0.70	0.83	0.65
Human minimal	0.68	0.67	0.77	0.65

Table 4: Reference-free evaluation: normalized scribendi scores on the test set of Nyberg (2022).

errors and are generally the most difficult to correct. This is presented in more detail in Table 6, where the frequency of each individual score is given. While the human corrections nearly always are classified as having no or minor differences, all automatic systems have significant numbers of moderate and substantial differences. Even given the fact that about 30% of these divergences are due to insufficient context (see Section 5), the difference is large enough to indicate that all GEC systems have problems producing corrections with adequate semantics.

6.3. A tree of corrections

In this project, we have produced a total of ten different versions of each sentence (three GEC systems and two humans, each with a post-edit), in addition to the original and its minimal correction from the SweLL project data. We visualize this by computing the normalized character-level Levenshtein distance between each pair among the twelve versions of each sentence, then using multidimensional scaling (Kruskal, 1964) as implemented by Pedregosa et al. (2011).

In Figure 1, we show the result as a tree starting at the original learner text (“original”), through its

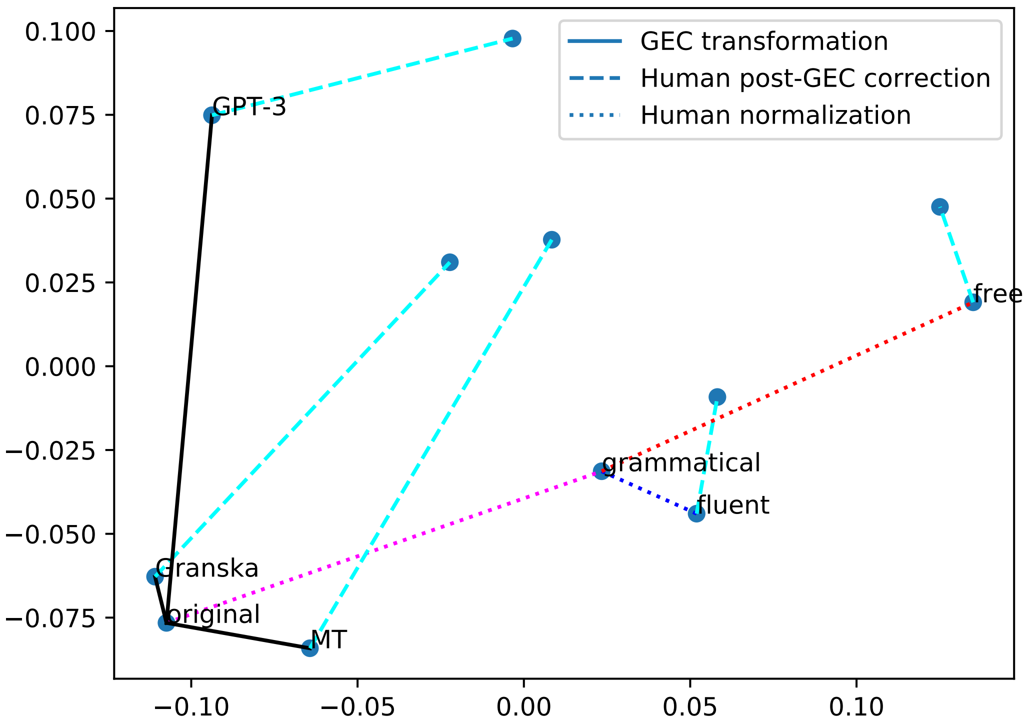

Figure 1: Multidimensional scaling view of all text versions, using normalized Levenshtein distance. The resulting graph is a tree, with its root at the “original” node that represents the original text of the learners. Edges represent transformations made by either computers (solid black lines) or humans (all other line types). Unlabeled leaf nodes represent human post-edits to their parent nodes in order to achieve full grammaticality, fluency and meaning preservation. Each human annotator is represented by a different color. The node "grammatical" represents the human-normalized references from the SweLL data, and "fluent" and "free" represent the Human fluent and Human free sentences as produced by the annotators, respectively.

immediate transformations by each of the GEC systems (solid lines) and the minimal human correction (“grammatical”), followed by the human rewrites for fluency (“fluent” and “free”) to the post-edited versions (dashed lines). This figure complements the bottom part of Table 5, which gives the distances of the post-editing edges.

It is clear that GPT-3 by far performs the most extensive changes to the text among the GEC systems. Still, the amount of post-editing required is much higher than the human-corrected versions. To some extent this can be explained by the fact that the “fluent” and “free” human corrections indirectly have access to a wider context through the SweLL “grammatical” correction, but as was shown in Section 5 the size of this effect is limited.

We can also see in Figure 1 that although the post-edits (leaf nodes) converge somewhat, they still produce rather different versions. According to the human annotator, all of the leaf nodes represent perfect corrections and none of them should be penalized in an evaluation. From the figure, we get an impression of the space of acceptable versions of the text in relation to the space of unacceptable versions.

7. Discussion

We have compared a diverse set of GEC systems (rule-based, MT-based, LLM-based, human), using a diverse set of metrics (reference-based, reference-free, human scoring, human post-editing). This allows us to make some general observations on the problem of GEC evaluation.

7.1. Metrics

The reference-free metric (Table 4) shows that the neural methods (MT, GPT-3) achieve very high scores, with GPT-3 scoring above or on par with

System	CEFR level All	 A	 B	 C
Grammaticality				
Granska	3.0	3.1	2.5	3.3
MT	3.3	3.4	3.0	3.5
GPT-3	3.7	3.8	3.6	3.8
Human fluent	3.9	3.9	3.8	3.9
Human free	3.9	3.9	3.9	3.8
Fluency				
Granska	2.8	3.0	2.3	3.2
MT	3.1	3.2	2.7	3.3
GPT-3	3.6	3.7	3.4	3.7
Human fluent	3.8	3.8	3.7	3.8
Human free	3.8	3.8	3.9	3.8
Meaning preservation				
Granska	3.5	3.5	3.2	3.7
MT	3.4	3.5	3.1	3.6
GPT-3	3.4	3.6	3.1	3.6
Human fluent	3.9	4.0	3.9	3.8
Human free	3.8	3.8	3.8	3.8
Normalized Levenshtein distance (NLD)				
Granska	0.126	0.119	0.180	0.079
MT	0.113	0.095	0.158	0.087
GPT-3	0.076	0.068	0.112	0.050
Human fluent	0.034	0.034	0.045	0.022
Human free	0.029	0.030	0.034	0.025

Table 5: Human evaluation: mean score per system and assessment dimension. Higher is better for all assessments (range: 1–4), while lower is better for NLD.

the minimal human correction. The rule-based Granska system, however, achieves scores comparable to the baseline of leaving the text uncorrected. By comparing with the human evaluation (Table 5), we see considerable differences. To begin with, the human evaluation demonstrates that all GEC systems are clearly below human-level preformance. We also see that while MT is clearly better than Granska with respect to grammaticality, fluency and post-edit distance, the gap is by no means as large as suggested by the Scribendi score.

As seen in Table 3, the reference-based GLEU suffers from ceiling effects, in particular for data from the most proficient (level C) learner group. The GLEU scores are nearly identical (0.63 and 0.65) for the MT and GPT-3 systems, but in terms of grammaticality, fluency and post-edit distance there is a clear difference.

A further illustration of this is found in Table 7, which shows system-level correlations for GLEU, Scribendi score, and the human assessments with respect to each of the human assessment dimensions. Here, the only negative correlations are precisely between the GLEU and Scribendi scores with respect to meaning preservation.

We argue that post-edit distance, when affordable, is perhaps the fairest single-dimensional GEC evaluation metric. In this work we quantify the edit distance using normalized character-level Levenshtein distance, but other edit distances that better model moved text segments may be more appropriate. Table 7 shows that NLD has the overall strongest correlation with other assessment dimensions, whereas there is a split between grammaticality and fluency on one hand, and meaning preservation on the other hand, which are only moderately correlated.

7.2. System objectives

As we discussed in the introduction, GEC systems can roughly be divided those that focus on providing error detection and correction, and those that aim to perform overall improvement of the text. Figure 1 clearly shows how this divide is reflected in post-edit distance. The Granska system is based on rules for specific error types, and provides human-readable feedback and suggested corrections for the error types covered, but overall makes very small edits. Although GPT-3 has no explicit objective in its design, due to its language model foundation it is particularly well-suited for holistic text normalization. This reflected in much larger edit distances. The neural MT model falls somewhere in between, as a neural model that is much smaller and less capable of language modeling than GPT-3, as well as being trained on normalizing synthetic errors. It is difficult to classify which objective this model really has. For the human versions, we see that the *free* version (with a general text improvement objective) brings the *grammatical* version (with a correction objective) further from the original learner text.

We note in Figure 1 that after the human post-edits, all systems end up in different locations in the space of possible corrections, regardless of what the objective of the system is. The systems with pure correction objectives (primarily *Granska* and *grammatical*) end up in a somewhat tighter region after post-editing, compared to those with a writing improvement objectives (*GPT-3* and *free*) that span a larger area.

7.3. Methodological limitations

An important lesson for future work concerns the importance of producing test sets with sufficiently long context, preferably whole documents. This would allow models with long context windows to demonstrate their full potential, and give a fair comparison to humans and to other computational models.

System	Identical	Minor	Moderate	Substantial	Other
Granska	125	34	11	13	9
MT	122	35	19	13	3
GPT-3	126	36	14	14	2
Human fluent	176	11	3	1	1
Human free	160	26	5	0	1

Table 6: Human evaluation: actual distribution of meaning preservation scores. The mean over each row in this table is summarized in the *All* column under *Meaning preservation* in Table 5.

	Gram.	Fluency	Meaning	−NLD
GLEU	0.97	0.96	-0.93	0.92
Scribendi	0.94	0.92	-0.96	0.86
Gram.	—	1.00	0.67	0.96
Fluency	1.00	—	0.66	0.96
Meaning	0.67	0.66	—	0.83
−NLD	0.96	0.96	0.83	—

Table 7: System-level Pearson correlation between each metric (automatic as well as human assesment dimensions) and each human assesment dimensions. Note that −NLD (higher is better) is used instead of NLD (lower is better) to make the correlations more intuitively interpretable. Uncertainty is very high due to the small number of systems (3), and any attempt to quantify this uncertainty would depend heavily on the prior distribution. We have chosen to present only maximum-likelihood estimates here.

We also note that our choice of relying on the minimal corrections from the SweLL data as gold standard is sometimes problematic, since multiple corrections with different semantics can be plausible. In our work, we used these minimal corrections as a basis for the other ("fluent" and "free") human corrections, which has the effect of reducing the diversity among the human corrections. If multiple human corrections from the original text were to be performed, we would also recommend annotating which cases are truly ambiguous even to a human. In addition, it would be helpful to include annotations of the grammaticality and fluency of the original sentence, for reference.

7.4. Advances in large language models

Since the initial annotations performed in this work, the state of the art in LLMs has advanced very rapidly. Compared to the GPT-3.5 `text-davinci-002` model used here, numerous very capable models have been published. Penteado and Perez (2023) evaluate one of the most capable, GPT-4, for Brazilian Portuguese GEC and finds that it is superior to GPT-3.5 in correcting the mostly orthographic and word choice errors present in their evaluation data. Yancey et al. (2023) similarly compare GPT-3.5 and GPT-4 in the more complex task of automatic writing evaluation, and again find that GPT-4 shows higher agreement with human raters. It is reasonable to expect that GPT-4 would have performed better than GPT-3.5 in our case as well. An important question for future work is to what extent GPT-4 and other recent models close the gap between GPT-3.5 and human performance that we identified.

7.5. Summary and future work

In conclusion, we show that with the advent of large language models, Swedish GEC has made enormous progress compared to early work. One of these models, GPT-3, produces corrections with human-like grammaticality and fluency. However, in the critical aspect of semantic accuracy, we see little improvment compared to other types of models.

For evaluating GEC systems, we demonstrate that different types of automatic evaluation metrics display different biases with respect to different types of GEC systems. Reference-free metrics favor neural systems, even over human corrections, while reference-based metrics struggle to differentiate systems at high proficiency levels.

In this work, we used simple Normalized Levenshtein distance to quantify the differences between post-edited corrections. For future work, we believe that a more thorough analysis of these differences would provide valuable insights into the weaknesses remaining even in very strong GEC systems. This could be done manually, or in some cases automatically similar to Felice et al. (2016).

Given the strong and rapidly improving ability of LLMs to handle extended contexts, we also see a need to perform future evaluations on longer segments of texts, including entire documents. However, working at the document level requires considerable adaptations of most existing evaluation methods, which would be another interesting direction of future work.

Acknowledgments

This work was funded by the Swedish Research Council (2019-04129) and in part by the Swedish national research infrastructure Språkbanken and Swe-Clarin, funded jointly by the Swedish Research Council (2018–2024, grant agreement 2017-00626) and the 10 participating partner institutions.

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at C3SE partially funded by the Swedish Research Council through grant agreements 2022-06725 and 2018-05973.

We also want to thank the anonymous reviewers for their comments which have significantly improved the final version of this paper.

8. Bibliographical References

Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. 2017. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Johnny Bigert and Ola Knutsson. 2002. Robust error detection: A hybrid approach combining unsupervised error detection and linguistic knowledge. In *Proc. 2nd Workshop Robust Methods in Analysis of Natural language Data (ROMAND'02)*, pages 10–19, Frascati, Italy.

Juhani Birn. 2000. Detecting grammar errors with lingsoft's Swedish grammar checker. In *Proceedings of the 12th Nordic Conference of Computational Linguistics (NODALIDA 1999)*, pages 28–40, Trondheim, Norway. Department of Linguistics, Norwegian University of Science and Technology, Norway.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Christopher Bryant and Ted Briscoe. 2018. Language model based grammatical error correction without annotated training data. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 247–253, New Orleans, Louisiana. Association for Computational Linguistics.

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.

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*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Rickard Domeij, Ola Knutsson, Johan Carlberger, and Viggo Kann. 2000. Granska–an efficient hybrid system for Swedish grammar checking. In *Proceedings of the 12th Nordic Conference of Computational Linguistics (NODALIDA 1999)*, pages 49–56, Trondheim, Norway. Department of Linguistics, Norwegian University of Science and Technology, Norway.

Ariel Ekgren, Amaru Cuba Gyllensten, Evangelia Gogoulou, Alice Heiman, Severine Verlinden, Joey Öhman, Fredrik Carlsson, and Magnus Sahlgren. 2022. Lessons learned from GPT-SW3: Building the first large-scale generative language model for Swedish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3509–3518, Marseille, France. European Language Resources Association.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, Florence, Italy. Association for Computational Linguistics.

Md Asadul Islam and Enrico Magnani. 2021. Is this the end of the gold standard? a straightforward reference-less grammatical error correction metric. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3009–3015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fred Karlsson, Atro Voutilainen, Juha Heikkila, and Arto Anttila. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Walter de Gruyter & Co., USA.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Joseph Bernard Kruskal. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Murathan Kurfalı and Robert Östling. 2023. A distantly supervised grammatical error detection/correction system for Swedish. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 35–39, Tórshavn, Faroe Islands. LiU Electronic Press.

Koki Maeda, Masahiro Kaneko, and Naoaki Okazaki. 2022. IMPARA: Impact-based metric for GEC using parallel data. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3578–3588, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China. Association for Computational Linguistics.

Martina Nyberg. 2022. Grammatical error correction for learners of swedish as a second language. Master's thesis, Uppsala University, Department of Linguistics and Philology.

Jakub Náplava, Milan Straka, Jana Straková, and Alexandr Rosen. 2022. Czech Grammar Error Correction with a Large and Diverse Corpus. *Transactions of the Association for Computational Linguistics*, 10:452–467.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maria Carolina Penteado and Fábio Perez. 2023. Evaluating GPT-3.5 and GPT-4 on grammatical error correction for brazilian portuguese. In *LatinX in AI Workshop at ICML 2023 (Regular Deadline)*.

Ying Qin and Lucia Specia. 2015. Truly exploring multiple references for machine translation evaluation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, Antalya, Turkey. European Association for Machine Translation.

Lisa Rudebeck, Gunlög Sundberg, and Mats Wirén. 2021. Swell normalization guidelines. Technical Report GU-ISS-2021-03, Department of Swedish, University of Gothenburg, Gothenburg.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4:169–182.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2019. The swell language learner corpus: From design to annotation. *Northern European Journal of Language Technology (NEJLT)*, 6:67–104.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark.

Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584, Toronto, Canada. Association for Computational Linguistics.

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. SOME: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522, Barcelona, Spain (Online). International Committee on Computational Linguistics.