Really good grammatical error correction, and how to evaluate it

Robert Östling
Department of Linguistics
Stockholm University
robert@ling.su.se

Murathan Kurfalı
Department of Linguistics
Stockholm University
murathan.kurfali@ling.su.se

Abstract

Grammatical error correction (GEC) has received relatively little attention for Swedish, with much of the important work published two decades ago. We perform a thorough evaluation using both automatic and manual methods of three very different types of models: the rule-based Granska system, a sequence-to-sequence model trained on artificial data, and the GPT-3 language model in a few-shot setup. We find that GPT-3 by far outperforms the other GEC systems for Swedish, a language comprising only 0.11% of its training data. When evaluating this heterogeneous set of GEC systems, we were also able to investigate how different evaluation methods are biased towards certain types of systems.

1 Background

Grammatical Error Correction (GEC) is typically used in an extended sense of correcting and improving language use at multiple levels, including spelling errors, grammatical errors, word choice and idiom usage. Traditionally, this has been approached by building systems that detect, categorize and correct individual errors, often with an eye towards applications in intelligent tutoring systems. We follow Sakaguchi et al. (2016) in rather taking native-level fluency as being the goal of GEC, beyond merely adhering to orthographic and grammatical rules of the language.

1.1 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 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 corupted data, generated in a fashion similar to that of Grundkiewicz et al. (2019).

1.2 Evaluating GEC

Traditional methods for GEC evaluation are reference-based, where either the GEC system output is compared to a human-created reference, e.g. GLEU (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, e.g. ERRANT (Bryant et al., 2017) which also allows a detailed automated analysis of error types.

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).

Often annotators are explicitly instructed to stay as close to the original text as possible (Volodina et al., 2019, Section 6.1). This has the effect of biasing existing automatic evaluations against systems that perform paraphrasing rather than conservatively fixing individual errors.

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 and outputs of GEC systems 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.

2 Evaluations

In this work we compare a total of five GEC systems and two baselines:

- **Uncorrected**: a dummy system that never changes anything.
- **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.
- **Nyberg MT**: the MT-based method of Nyberg (2022, Section 3.3).
- **Nyberg LM**: the LM-based method of Nyberg (2022, Section 3.4).
- **MT**: Similar to Nyberg MT, but with more data (3.2 billion words of cleaned and deduplicated modern Swedish text from Språkbanken[1]), a modified method for introducing synthetic errors and a different architecture. Among other things, word order shuffling is not limited to adjacent words. The model is based on the transformer architecture and trained using OpenNMT-py (Klein et al., 2017) following the suggested base configuration. The model is trained for 100K training steps with the effective batch size of 6400 sentences.
- **GPT-3**: OpenAI's (`text-davinci-002`) model (Brown et al., 2020) through their public API. We use a two-shot prompt, with authentic student sentences taken from the Cross-Check corpus[2] and manully corrected by us. The prompt is entirely in Swedish.
- **Reference**: human-normalized sentences from the SweLL project. Annotators were asked to perform minimal edits to reach a grammatically correct sentence, and had access to the full text for context.

[1] https://spraakbanken.gu.se/en/resources/corpus

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

2.1 Automatic evaluation

We use Swedish data from the SweLL project (Volodina et al., 2019), which consists of 502 student 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).

Table 1 shows the GLEU scores of the systems and baselines in our evaluation. We note that GPT-3 not only achieves higher scores than the other systems, but reaches a similar level for all three proficiency levels. The gap between the most easily corrected proficiency level (C) and the most difficult (B) is in the range 0.25–0.31 for all system except GPT-3, where it is only 0.13.

Table 2 shows the normalized Scribendi score (Islam and Magnani, 2021), a reference-free metric that ranges from -1 (the GEC system always makes the text *worse*) to 1 (the system always improves the text). Not changing the text results in a value of 0 by definition. We have adapted it to Swedish by using the GPT-SW3 language model (Ekgren et al., 2022). We lack figures from Nyberg (2022), but the general trend is similar to the GLEU results: GPT-3 at the top, followed by the MT model and then Granska. The difference between the two neural systems (MT and GPT-3) is however much smaller than in the GLEU evaluation, while the advantage over Granska increases. Note also that GPT-3 scores better than the reference. This is not surprising, since the reference sentences are aimed at reaching grammaticality rather than fluency, while GPT-3 is excellent at producing fluent text and the similar GPT-SW3 model is excellent at detecting fluency. As long as the meaning of the corrections produced by GPT-3 does not diverge sufficiently to be detected by the Scribendi scoring algorithm, this almost guarantees a higher score than the reference.

2.2 Manual evaluation

We have two native speakers annotating the output of each GEC system for grammaticality, fluency and meaning preservation, with their respective 4-level scales as used by Yoshimura et al. (2020). In addition, the annotators provide a minimal edit of

		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**
Reference	1.0	1.0	1.0	1.0

Table 1: GLEU scores on the test set of Nyberg (2022). The Nyberg MT and LM scores are from that paper, the rest computed by us.

		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
Reference	0.68	0.67	0.77	0.65

Table 2: Normalized Scribendi scores on the test set of Nyberg (2022).

the system output to reach full scores on all three dimensions. The amount of editing done, here measured using normalized Levenshtein distance (NLD), captures how close the system's output is to the *closest* native-sounding alternative.

The result is shown in Table 3.[3] We see here that for both grammaticality and fluency, the GEC systems rankings agree with the GLEU and Scribendi evaluations. For meaning preservation, however, the order is reversed. This reflects the tendency of GPT-3 in particular to occasionally produce a perfectly grammatical and fluent sentence that means something completely different.

3 Conclusions

The GPT-3 language model turns out to be the best GEC system for Swedish, by a comfortable margin, according to all of our evaluation methods. This is particularly interesting given that it has only seen 221 million words of Swedish data during training, 0.11% of the total amount of training words, and indicates excellent cross-linguistic generalization abilities.

[3]This is on-going work. The preliminary values in the table are based on only 20 sentences from the Nyberg (2022) development set annotated during annotator training, and should be taken with a grain of salt. We note however that the system rankings by all metrics are identical for both annotators.

	Granska	MT	GPT-3
Grammaticality	2.6	3.1	**3.6**
Fluency	2.0	2.7	**3.4**
Meaning	**3.8**	3.6	3.2
NLD	0.17	0.14	**0.11**

Table 3: Human ratings of GEC system outputs. Scales ranges from 1 (worst) to 4 (best), except for normalized Levenshtein distance (NLD) which ranges from 0 (best) to 1 (worst).

By evaluating multiple GEC models of very different types, we also learn about evaluation methodology. Reference-based metrics are mainly aimed at evaluating how well a GEC system performs minimal changes towards grammaticality. This biases them against systems that tend to paraphrase the input sentences to sound more native-like. This is consistent with what we see when comparing our Table 1 with Table 3, where the reference-based GLEU metric scores GPT-3 only slightly higher than the MT system, but the human evaluation has a rather large margin.

Previous GEC systems rarely alter semantics significantly (Yoshimura et al., 2020, Figure 1), this has also been confirmed by our manual evaluation (Table 3). However, semantic corruption turns out to be a particular failure mode of GPT-3. Reference-free metrics have been shown in previous work to achieve very high agreement with human judgements, but optimizing the correlation between human judgement and previous GEC system outputs results in semantic corruption being overlooked.

By manually correcting the system outputs and measuring the amount of corrections, we can measure how much work the system left undone. We have simply used the normalized Levenshtein distance here, but one can also analyze in greater detail which types of corrections were made. This is left for future work.

Acknowledgements

Annotation work was performed by the first author and Katarina Gillholm, who also participated in the development of the annotation guidelines.

The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

This work was funded in part by the Swedish Research council through grant agreement no. 2019-04129.

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.

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.

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.

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.

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.

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.

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. 6:67–104.

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.