GECko+: a Grammatical and Discourse Error Correction Tool

Eduardo Calò, Léo Jacqmin, Thibo Rosemplatt, Maxime Amblard, Miguel Couceiro, Ajinkya Kulkarni

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

GECko+: a Grammatical and Discourse Error Correction Tool

Eduardo Calò[1*] Léo Jacqmin[1*] Thibo Rosemplatt[1*]
Maxime Amblard[12] Miguel Couceiro[12] Ajinkya Kulkarni[2]
(1) IDMC, Université de Lorraine, F-54000 Nancy, France
(2) Université de Lorraine, CNRS, Inria N.G.E., LORIA, F-54000, France
{eduardo.calo6,leo.jacqmin8,thibo.rosemplatt3}@etu.univ-lorraine.fr,
{maxime.amblard,miguel.couceiro,ajinkya.kulkarni}@loria.fr

RÉSUMÉ

Nous présentons GECko+, un assistant d'écriture pour l'anglais qui corrige des erreurs au niveau de la phrase et du discours. Il se base sur deux modèles état de l'art pour la correction grammaticale et pour la réorganisation de phrases. GECko+ est disponible en ligne sous la forme d'une application web qui implémente une chaîne de traitement assemblant ces deux modèles.

ABSTRACT

GECko+ : a Grammatical and Discourse Error Correction Tool

We introduce GECko+, a web-based writing assistance tool for English that corrects errors both at the sentence and at the discourse level. It is based on two state-of-the-art models for grammar error correction and sentence ordering. GECko+ is available online as a web application that implements a pipeline combining the two models.

MOTS-CLÉS : assistant d'écriture, correction grammaticale, analyse de discours.

KEYWORDS: writing assistant tool, grammatical error correction, discourse analysis.

1 Introduction

While most people can write, few would boast they never produce spelling and grammar mistakes, let alone systematically write coherent prose and express ideas clearly. Natural language processing (NLP) techniques have the potential to help in that regard. In particular, such technologies can have a beneficial impact on two issues related to the way we write.

First, NLP techniques can help us alleviate language-related discrimination (Papakyriakopoulos *et al.*, 2020), that occurs, e.g., in the professional world where job applications are rejected simply due to the quality of one's writing. Additionally, errorful writing is poorly perceived in social contexts and is often synonymous with barriers.

Second, those who are already proficient in writing can benefit from these techniques to improve the quality of their prose. This aspect applies to journalists, business-persons, college students, and teachers alike. These individuals are often required to write lengthy reports. The frequency with which these reports are produced is such that topological or consistency errors can occur. As a result,

*. These authors contributed equally to this work.

their message may not be delivered as intended.

To address these issues, we propose a digital writing assistance tool for English that we call GECko+ that uses existing state-of-the-art models to tackle both sentence-level mistakes and discourse incoherence. To correct spelling and grammar mistakes, we use GECToR (Omelianchuk *et al.*, 2020), a grammatical error correction (GEC) model developed by the well-known Grammarly [1]. For tackling discourse incoherence, we make use of a sentence ordering model [2] (Prabhumoye *et al.*, 2020) based on Google's BERT (Devlin *et al.*, 2019). We created a web interface that users can access to correct paragraphs of text in English [3]. The code is publicly available on GitHub [4].

2 Background

GEC in NLP encompasses any sort of modifications made to automatically correct an errorful sentence. This includes spelling, punctuation, grammar, and word choice errors. Given a potentially errorful sentence or short piece of text as input, a GEC system is expected to output a corrected version of that text. We have reviewed several approaches to correct sentences individually (Chollampatt & Ng, 2018; Junczys-Dowmunt *et al.*, 2018).

However, language does not simply consist of individual, independent sentences that are added one after the other, but rather forms a coherent whole composed of interconnected sentences. This coherent whole is commonly referred to as discourse. The area of NLP concerned with how sentences fit together is called discourse coherence or discourse analysis (Jurafsky & Martin, 2009). Discourse analysis encompasses many different aspects and can be very fine-grained. One of these aspects is sentence ordering, whose goal is to arrange sentences of a given text in the correct order, i.e., in a coherent manner.

3 Description of the Tool

GECko+ combines two state-of-the-art models into a single pipeline. To tackle sentence-wise errors, it employs GECToR (Omelianchuk *et al.*, 2020), which treats GEC as a sequence tagging task, relying on a Transformer-based encoder. To address discourse coherence, it utilizes a sentence ordering model (Prabhumoye *et al.*, 2020), which predicts the relative ordering between pairs of sentences from an input list of sentences. The reordering task is treated as a constraint learning task. The pipeline is shown in Figure 1.

As the diagram shows, the text given as input by the user gets segmented into sentences. After the segmentation, we obtain a list S of sentences, whose length ranges from one to n. Then, GECToR is applied to each sentence s_i in S, in order to perform sentence-wise error correction. Each sentence is iteratively processed by the model to ensure that all interdependent errors get corrected. As a result, the n sentences that constitute S are now free of grammatical errors. Subsequently, if $n = 1$, the single corrected sentence is directly output to the user. Conversely, if S contains more than one

1. https://github.com/grammarly/gector
2. https://github.com/shrimai/Topological-Sort-for-Sentence-Ordering
3. https://gecko-app.azurewebsites.net/
4. https://github.com/psawa/gecko-app

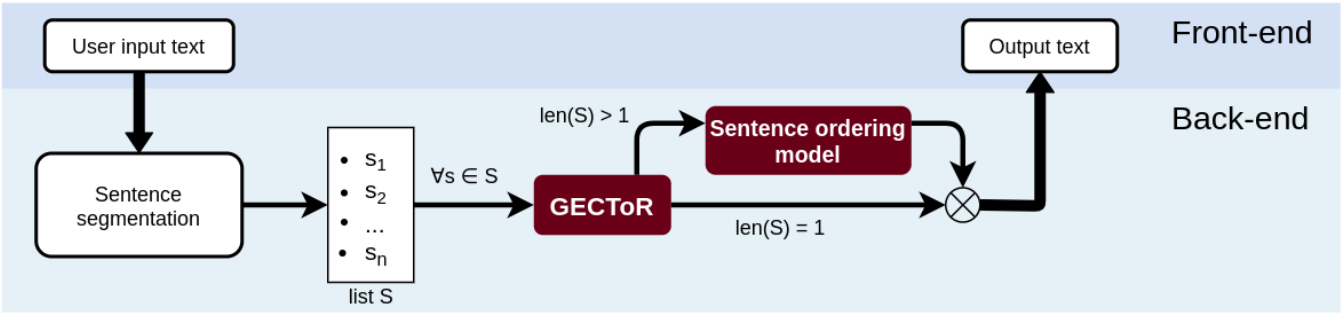

FIGURE 1 – GECko+'s pipeline.

element, the potentially unordered list of sentences will be given as input to the sentence ordering model. Once the sentences are ordered, the output is displayed to the user.

GECko+ employs a simple but effective color code to highlight mistakes. Changes are highlighted token-wise : deletions are underlined in red, modifications in blue, and additions in green. Currently there is no explicit indication of how sentences have been reordered. Ideally, a user should be able to visualize which sentences were swapped. We leave it for future work. Refer to Figure 2 for GECko+'s interface with an example sentence containing various spelling, grammar, and discourse mistakes.

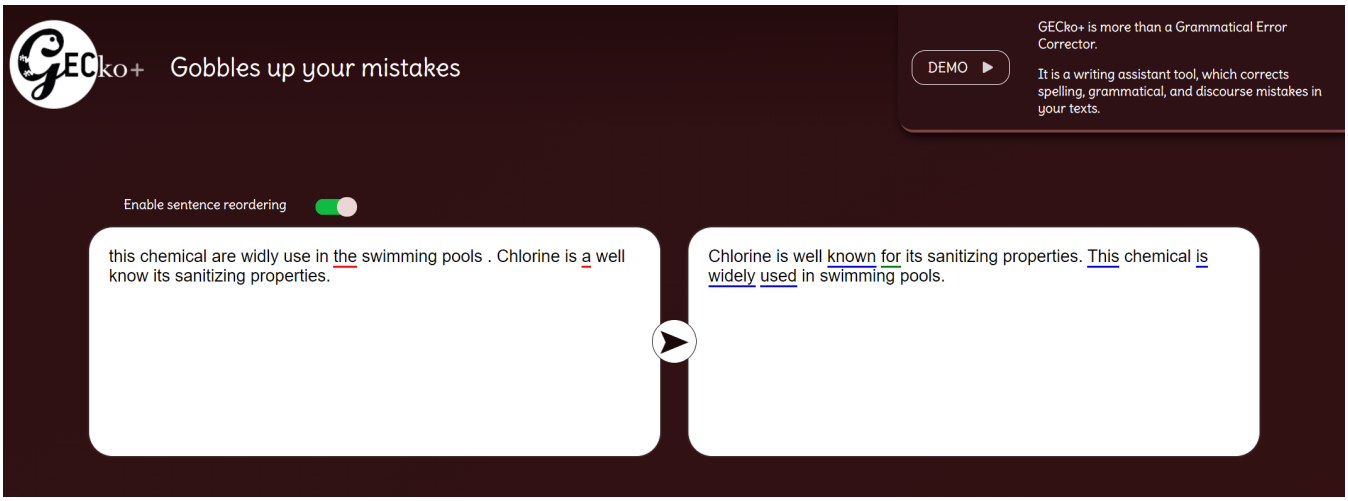

FIGURE 2 – GECko+'s interface.

4 Evaluation

The results for GECToR have been reported on CoNLL-2014 test set (Ng *et al.*, 2014) with the M^2 Scorer (Dahlmeier & Ng, 2012) and on BEA-2019 development and test sets (Bryant *et al.*, 2019) with ERRANT (Bryant *et al.*, 2017). For single models, they achieved state-of-the-art performance with an XLNeT-based model, which we use for our application, obtaining $F_{0.5} = 65.3$ on CoNLL-2014 (test) and $F_{0.5} = 72.4$ on BEA-2019 (test). The sentence ordering model was evaluated across several datasets using multiple metrics along with a human evaluation. The BERT-based approach scored higher than the previous state-of-the-art method on all metrics, obtaining a Sentence Accuracy of 61.48 on the NIPS dataset. Refer to (Prabhumoye *et al.*, 2020) for a detailed description of the results.

Références

BRYANT C., FELICE M., ANDERSEN Ø. E. & BRISCOE T. (2019). The bea-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 52–75 : Association for Computational Linguistics. DOI : 10.18653/v1/W19-4406.

BRYANT C., FELICE M. & BRISCOE T. (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)*, p. 793–805, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1074.

CHOLLAMPATT S. & NG H. T. (2018). A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, p. 5755–5762, New Orleans, Louisiana USA.

DAHLMEIER D. & NG H. T. (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*, p. 568–572, Montréal, Canada : Association for Computational Linguistics.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

JUNCZYS-DOWMUNT M., GRUNDKIEWICZ R., GUHA S. & HEAFIELD K. (2018). Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 595–606, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1055.

JURAFSKY D. & MARTIN J. H. (2009). *Speech and Language Processing (2nd Edition)*. USA : Prentice-Hall, Inc.

NG H. T., WU S. M., BRISCOE T., HADIWINOTO C., SUSANTO R. H. & BRYANT C. (2014). The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning : Shared Task*, p. 1–14, Baltimore, Maryland : Association for Computational Linguistics. DOI : 10.3115/v1/W14-1701.

OMELIANCHUK K., ATRASEVYCH V., CHERNODUB A. & SKURZHANSKYI O. (2020). GECToR – grammatical error correction : Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 163–170, Seattle, WA, USA. Online : Association for Computational Linguistics.

PAPAKYRIAKOPOULOS O., HEGELICH S., SERRANO J. C. M. & MARCO F. (2020). Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, p. 446–457, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3351095.3372843.

PRABHUMOYE S., SALAKHUTDINOV R. & BLACK A. W. (2020). Topological sort for sentence ordering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 2783–2792, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.248.