GECKo+: a Grammatical and Discourse Error Correction Tool
Eduardo Calò, Léo Jacqmin, Thibo Rosemplatt, Maxime Amblard, Miguel Couceiro, Ajinkya Kulkarni

To cite this version:
Eduardo Calò, Léo Jacqmin, Thibo Rosemplatt, Maxime Amblard, Miguel Couceiro, et al.. GECko+: a Grammatical and Discourse Error Correction Tool. TALN 2021 - 28e Conférence sur le Traitement Automatique des Langues Naturelles, Jun 2021, Lille / Virtual, France. pp.8-11. hal-03265914v2

HAL Id: hal-03265914
https://hal.archives-ouvertes.fr/hal-03265914v2
Submitted on 23 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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,
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 sentence, GECToR is applied to the entire list $S$ to correct discourse incoherence. The final output is a coherent and error-free paragraph.

1. https://github.com/grammarly/gector
2. https://github.com/shrimai/Topological-Sort-for-Sentence-Ordering
3. https://gecko-app.azurewebsites.net/
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 1 – GECko+’s pipeline.

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.

Figure 2 – GECko+’s interface.
Références


