Projeto Final 06
avanços 2a. semana

NSGC - Neural Spell & Grammar Checker (en/pt)

Rafael Ito
03/06/2020

Plano da Apresentação

- Planejamento/Cronograma original
- O que foi realizado na semana
- O que será feito na próxima
- Planejamento/Cronograma atualizado

Planejamento/Cronograma original

Semana 2: 28/maio → 03/junho

○ Leitura do último artigo (GECToR).

○ Teste de modelos com diferentes números de parâmetros (base, large, etc).

○ **Comparação em termos de qualidade e custo computacional de corretores baseados em BERT e T5.**

○ Uso de diferentes métricas para calcular palavras/sentenças mais próximas da que está em análise (edit distance, SBERT).

○ Proposta da arquitetura final.

○ Compilação dos diferentes testes produzidos na semana anterior.

O que foi realizado na semana

- Estudo sobre distância e similaridade entre strings
- Código:
  - esqueleto da arquitetura do corretor
  - métrica de similaridade entre palavras
  - avaliação nos datasets
  - adaptação do código da métrica GLEU

JFLEG dataset and GLEU metric

Original: they just creat impression such well that people are drag to buy it .
Minimal edit: They just create an impression so well that people are dragged to buy it .
Fluency edit: They just create such a good impression that people are compelled to buy it.

GUG corpus (Grammatical/Ungrammatical)

- 3.1k sentences written by English language learners for the TOEFL exam
- **GUG score:** (1–4, where 4 is perfect or native sounding, and 1 incomprehensible)
- **Evaluation metric:** GLEU (Generalized Language Understanding Evaluation)

GLEU:

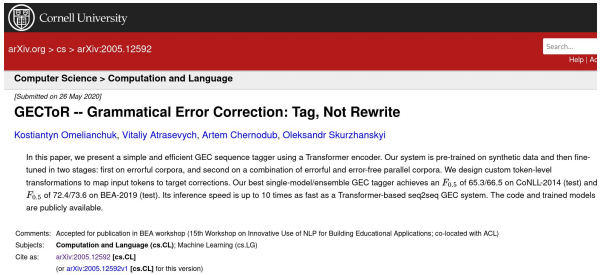
- based on BLEU
- score **fluency** in addition to minimal edits
- penalize n-grams that should have been changed in the system output but were left unchanged

2015 - [GLEU] Ground Truth for Grammatical Error Correction Metrics (Napoles et al., 2015)

2016 - [GLEU] GLEU Without Tuning (Napoles et al., 2016)

2017 - [JFLEG] JFLEG: A Fluency Corpus and Benchmark for Grammatical Error (Napoles et al., 2017)

Late paper



Cornell University

arXiv.org > cs > arXiv:2005.12592

Computer Science > Computation and Language

[Submitted on 26 May 2020]

GECToR -- Grammatical Error Correction: Tag, Not Rewrite

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, Oleksandr Skurzhanskyi

In this paper, we present a simple and efficient GEC sequence tagger using a Transformer encoder. Our system is pre-trained on synthetic data and then fine-tuned in two stages: first on errorful corpora, and second on a combination of errorful and error-free parallel corpora. We design custom token-level transformations to map input tokens to target corrections. Our best single-model/ensemble GEC tagger achieves an $F_{0.5}$ of 65.3/66.5 on CoNLL-2014 (test) and $F_{0.5}$ of 72.4/73.6 on BEA-2019 (test). Its inference speed is up to 10 times as fast as a Transformer-based seq2seq GEC system. The code and trained models are publicly available.

Comments: Accepted for publication in BEA workshop (15th Workshop on Innovative Use of NLP for Building Educational Applications; co-located with ACL)

Subjects: **Computation and Language (cs.CL)**; Machine Learning (cs.LG)

Cite as: arXiv:2005.12592 **[cs.CL]**

(or arXiv:2005.12592v1 **[cs.CL]** for this version)

~~Neural Machine Translation (NMT)-based~~ → pre-trained Transformer-NMT-based ~~sequence generation~~ → sequence tagging

Evaluation:

- CoNLL-2014
- BEA

Encoder	CoNLL-2014 (test)			BEA-2019 (dev)		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$
LSTM	51.6	15.3	35.0	-	-	-
ALBERT	59.5	31.0	50.3	43.8	22.3	36.7
BERT	65.6	36.9	56.8	48.3	29.0	42.6
GPT-2	61.0	6.3	22.2	44.5	5.0	17.2
RoBERTa	**67.5**	38.3	**58.6**	**50.3**	30.5	**44.5**
XLNet	64.6	**42.6**	58.5	47.1	**34.2**	43.8

Table 6: Varying encoders from pretrained Transformers in our sequence labeling system. Training was done on data from training stage II only.

2020 - [GECToR] GECToR – Grammatical Error Correction: Tag, Not Rewrite (Omelianchuk et al., 2020)

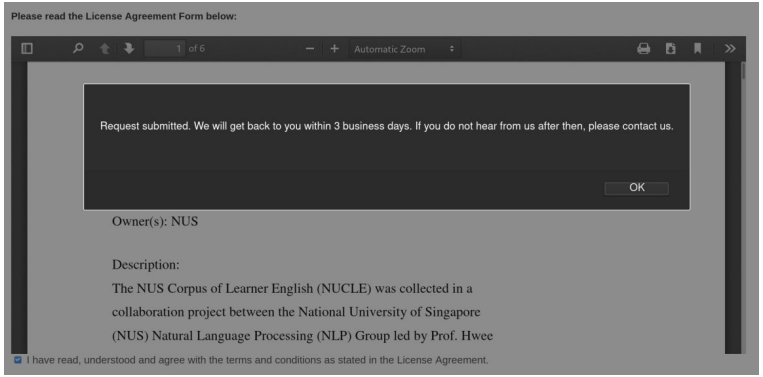
CoNLL-2014 dataset

Task: GEC (Grammatical Error Correction)

NUCLE corpus (the NUS Corpus of Learner English)

Collection of 1,414 essays written by students that have English as a 2nd language at the National University of Singapore (NUS).

- **Only test set available publicly!**
- **Fulfilled the form signing the license agreement**
- **Sent an email to the staff**



NUCLE Release 3.3 Posta in arrivo ×

Lin Qian <linq@comp.nus.edu.sg>
a me

inglese › italiano Traduci messaggio

Dear Rafael,

Please find attached the NUCLE Release 3.3 and the preprocessing script.

Regards,
Lin Qian

2014 - [CoNLL-2014] The CoNLL-2014 Shared Task on Grammatical Error Correction (Ng et al., 2014)

BEA dataset

Building Educational Applications 2019 Shared Task: Grammatical Error Correction (GEC)

- **Cambridge English Write & Improve (W&I)** corpus (Yannakoudakis et al., 2018)
  - online web platform that assists non-native English students with their writing
  - letters, stories, articles and essays
  - CEFR level
    - A (beginner)
    - B (intermediate)
    - C (advanced)

- **LOCNESS** corpus (Granger, 1998)
  - essays written by native English students
    - N (native)

Already tokenised with spaCy

Metric: ERRANT

2017 - [ERRANT] Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction [University of Cambridge] (Bryant et al., 2017)

2019 - [BEA] The BEA-2019 Shared Task on Grammatical Error Correction [University of Cambridge] (Bryant et al., 2019)

String similarity and distance

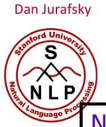
Edit distance:

- Levenshtein distance (insertions, deletions or substitutions)
- Damerau–Levenshtein distance (insertions, deletions, substitutions or **transposition**)
- Jaro distance (only transposition)

$$D(i,j) = \min \begin{cases} D(i-1,j) + del[x(i)] \\ D(i,j-1) + ins[y(j)] \\ D(i-1,j-1) + sub[x(i),y(j)] \end{cases}$$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	`	1	2	3	4	5	6	7	8	9	0	-	=	
1		q	w	e	r	t	y	u	i	o	p	[	]	\
2		a	s	d	f	g	h	j	k	l	;	'		
3		z	x	c	v	b	n	m	,	.	/			

Dan Jurafsky



The Edit Distance Table

$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	**0**	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

https://web.stanford.edu/class/cs124/lec/med.pdf

Code

```
[349]  1 # mask tokens
       2 for i in range(len(input_ids)):
       3     input_ids[i][i+1] = tokenizer.mask_token_id
       4 input_ids
```

```
tensor([[ 101,  103,  117, 1142, 1331, 1110, 1515,  170, 1992, 1849,  119,  102],
        [ 101, 8094,  103, 1142, 1331, 1110, 1515,  170, 1992, 1849,  119,  102],
        [ 101, 8094,  117,  103, 1331, 1110, 1515,  170, 1992, 1849,  119,  102],
        [ 101, 8094,  117, 1142,  103, 1110, 1515,  170, 1992, 1849,  119,  102],
        [ 101, 8094,  117, 1142, 1331,  103, 1515,  170, 1992, 1849,  119,  102],
        [ 101, 8094,  117, 1142, 1331, 1110,  103,  170, 1992, 1849,  119,  102],
        [ 101, 8094,  117, 1142, 1331, 1110, 1515,  103, 1992, 1849,  119,  102],
        [ 101, 8094,  117, 1142, 1331, 1110, 1515,  170,  103, 1849,  119,  102],
        [ 101, 8094,  117, 1142, 1331, 1110, 1515,  170, 1992,  103,  119,  102],
        [ 101, 8094,  117, 1142, 1331, 1110, 1515,  170, 1992, 1849,  103,  102]])
```

https://colab.research.google.com/drive/194LQ5UyrmFJOKUPL7qyAcDFkcfWF3qV1?authuser=1#scrollTo=gTGvw969QXqO

O que será feito na próxima semana

- Avaliação e definição do modelo a ser usado (BERT e/ou T5)
- Organização do notebook
- Entrega do corretor neural em inglês

Planejamento/Cronograma atualizado

Semana 3: 04/junho → 10/junho

(entrega do corretor em inglês)

- Decidir o modelo a ser usado
- Começar o corretor em inglês
- Finalização do corretor para inglês.
- **Avaliação de performance entre BERT e T5.**

Semana 4: 11/junho → 17/junho

(entrega do corretor em português)

- Aplicação do procedimento que melhor funcionou para o inglês, mas agora com textos em português.
- Possíveis pequenos ajustes devido a mudança de idioma.
- Finalização do corretor para português.
- **Criação de um dataset artificial em português com erros ortográficos para avaliação**