Projeto Final 06
avanços 1a. semana

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

Rafael Ito
28/05/2020

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 1: 21/maio → 27/maio

(estudos iniciais)

○ Estudos direcionados.

○ Leitura dos artigos envolvidos.

○ Familiarização com datasets.

○ Primeira proposta de projeto em detalhes.

Semana 2: 28/maio → 03/junho

(experimentações)

○ 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).

O que foi realizado na semana

- Leitura de artigos
- Familiarização com datasets
- Códigos:
    - métricas de avaliação
    - primeiro modelo

HOO: Helping Our Own

HOO is an ongoing shared task concerned with the automated correction of errors in text.

HOO-2011:

- all error types
- almost all participating teams dealt with **article and preposition** errors only (besides spelling and punctuation errors)

HOO-2012:

- Article and preposition

2010 - [HOO] Helping Our Own: Text Massaging for Computational Linguistics as a New Shared Task (Dale & Kilgarriff, 2010)

2011 - [HOO-2011] Helping Our Own: The HOO 2011 Pilot Shared Task (Dale & Kilgarriff, 2011)

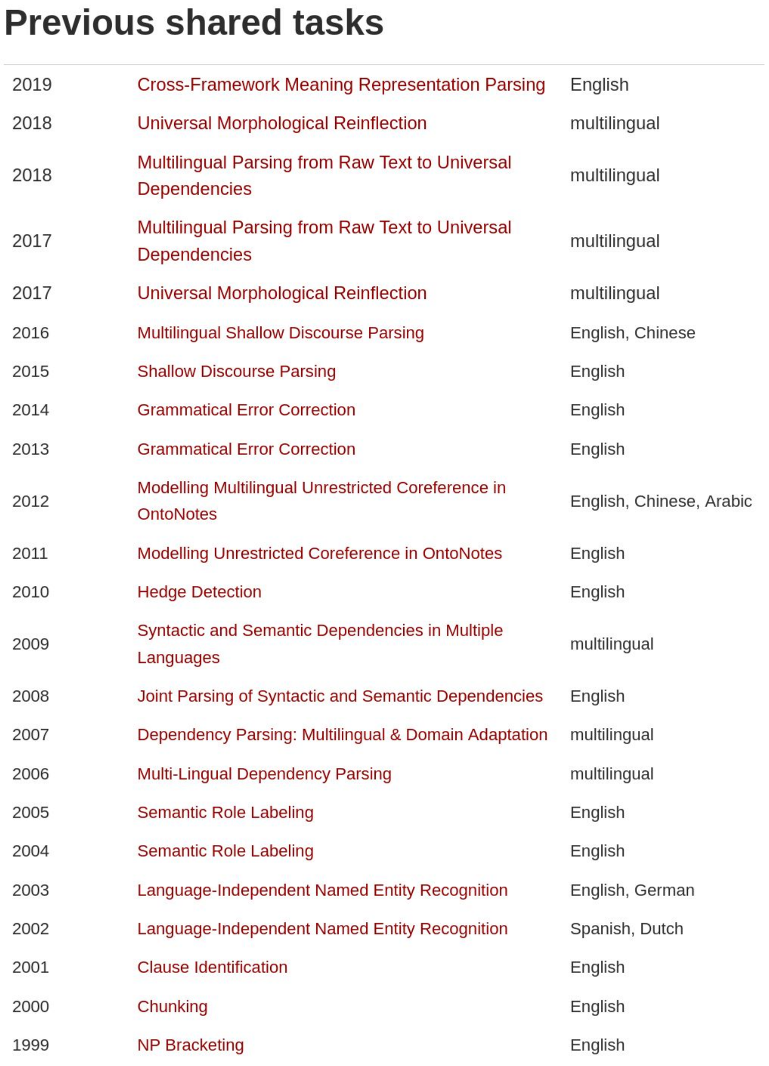
2012 - [HOO-2012] HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task (Dale et al., 2012)

CoNLL

The SIGNLL Conference on Computational Natural Language Learning

Tasks:

- NER (named entity recognition)
- SRL (semantic role labeling)
- dependency parsing
- coreference resolution
- etc



Previous shared tasks

2019	Cross-Framework Meaning Representation Parsing	English
2018	Universal Morphological Reinflection	multilingual
2018	Multilingual Parsing from Raw Text to Universal Dependencies	multilingual
2017	Multilingual Parsing from Raw Text to Universal Dependencies	multilingual
2017	Universal Morphological Reinflection	multilingual
2016	Multilingual Shallow Discourse Parsing	English, Chinese
2015	Shallow Discourse Parsing	English
2014	Grammatical Error Correction	English
2013	Grammatical Error Correction	English
2012	Modelling Multilingual Unrestricted Coreference in OntoNotes	English, Chinese, Arabic
2011	Modelling Unrestricted Coreference in OntoNotes	English
2010	Hedge Detection	English
2009	Syntactic and Semantic Dependencies in Multiple Languages	multilingual
2008	Joint Parsing of Syntactic and Semantic Dependencies	English
2007	Dependency Parsing: Multilingual & Domain Adaptation	multilingual
2006	Multi-Lingual Dependency Parsing	multilingual
2005	Semantic Role Labeling	English
2004	Semantic Role Labeling	English
2003	Language-Independent Named Entity Recognition	English, German
2002	Language-Independent Named Entity Recognition	Spanish, Dutch
2001	Clause Identification	English
2000	Chunking	English
1999	NP Bracketing	English

CoNLL-2013

Task: GEC (Grammatical Error Correction)

Correct only 5 error types:

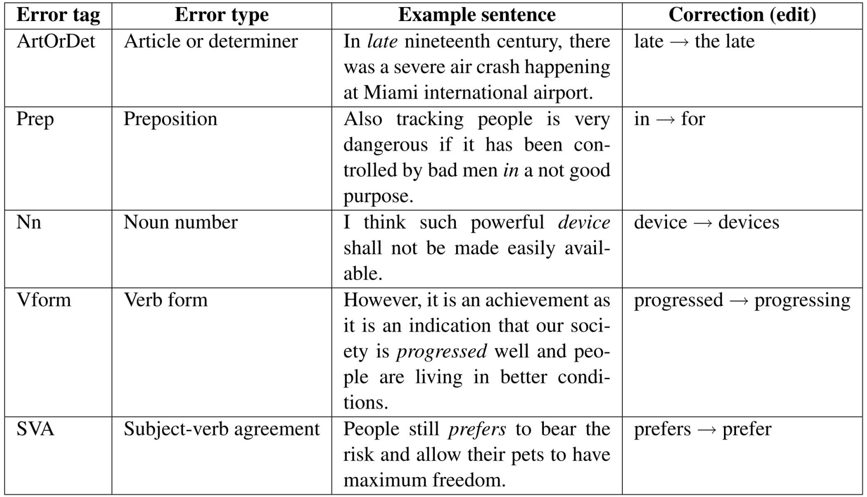
- Article or determiner
- Preposition
- Noun number
- Verb form
- Subject-verb agreement

Evaluation metric:

- F_1 score
- MaxMatch (M^2) score: score **minimal edits**

1 human annotator

2013 - [CoNLL-2013] The CoNLL-2013 Shared Task on Grammatical Error Correction (Ng et al., 2013)



Error tag	**Error type**	**Example sentence**	**Correction (edit)**
ArtOrDet	Article or determiner	In *late* nineteenth century, there was a severe air crash happening at Miami international airport.	late → the late
Prep	Preposition	Also tracking people is very dangerous if it has been controlled by bad men *in* a not good purpose.	in → for
Nn	Noun number	I think such powerful *device* shall not be made easily available.	device → devices
Vform	Verb form	However, it is an achievement as it is an indication that our society is *progressed* well and people are living in better conditions.	progressed → progressing
SVA	Subject-verb agreement	People still *prefers* to bear the risk and allow their pets to have maximum freedom.	prefers → prefer

CoNLL-2014

Task: GEC (Grammatical Error Correction)

Correct 28 error types

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).

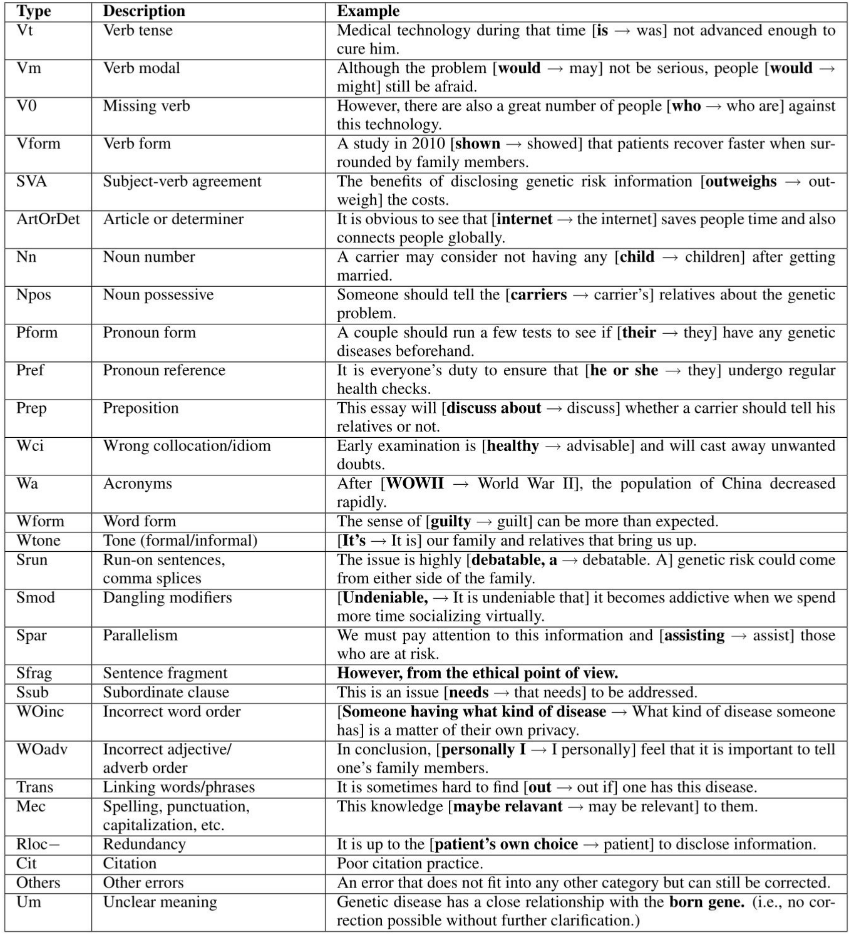
Evaluation metric:

- $F_{0.5}$ score: emphasizes precision over recall

2 human annotators (independently)

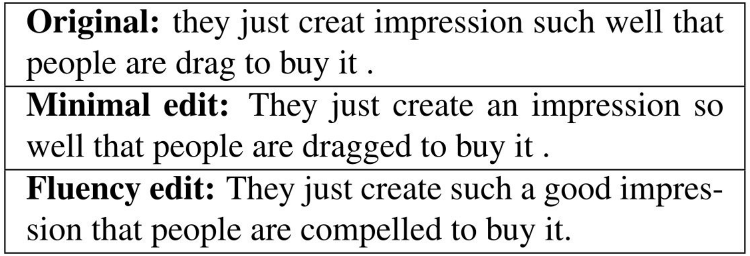
SGML format (Standard Generalized Markup Language)

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



Type	Description	Example
Vt	Verb tense	Medical technology during that time [**is** → was] not advanced enough to cure him.
Vm	Verb modal	Although the problem [**would** → may] not be serious, people [**would** → might] still be afraid.
V0	Missing verb	However, there are also a great number of people [**who** → who are] against this technology.
Vform	Verb form	A study in 2010 [**shown** → showed] that patients recover faster when surrounded by family members.
SVA	Subject-verb agreement	The benefits of disclosing genetic risk information [**outweighs** → outweigh] the costs.
ArtOrDet	Article or determiner	It is obvious to see that [**internet** → the internet] saves people time and also connects people globally.
Nn	Noun number	A carrier may consider not having any [**child** → children] after getting married.
Npos	Noun possessive	Someone should tell the [**carriers** → carrier's] relatives about the genetic problem.
Pform	Pronoun form	A couple should run a few tests to see if [**their** → they] have any genetic diseases beforehand.
Pref	Pronoun reference	It is everyone's duty to ensure that [**he or she** → they] undergo regular health checks.
Prep	Preposition	This essay will [**discuss about** → discuss] whether a carrier should tell his relatives or not.
Wci	Wrong collocation/idiom	Early examination is [**healthy** → advisable] and will cast away unwanted doubts.
Wa	Acronyms	After [**WOWII** → World War II], the population of China decreased rapidly.
Wform	Word form	The sense of [**guilty** → guilt] can be more than expected.
Wtone	Tone (formal/informal)	[**It's** → It is] our family and relatives that bring us up.
Srun	Run-on sentences, comma splices	The issue is highly [**debatable, a** → debatable. A] genetic risk could come from either side of the family.
Smod	Dangling modifiers	[**Undeniable,** → It is undeniable that] it becomes addictive when we spend more time socializing virtually.
Spar	Parallelism	We must pay attention to this information and [**assisting** → assist] those who are at risk.
Sfrag	Sentence fragment	**However, from the ethical point of view.**
Ssub	Subordinate clause	This is an issue [**needs** → that needs] to be addressed.
WOinc	Incorrect word order	[**Someone having what kind of disease** → What kind of disease someone has] is a matter of their own privacy.
WOadv	Incorrect adjective/ adverb order	In conclusion, [**personally I** → I personally] feel that it is important to tell one's family members.
Trans	Linking words/phrases	It is sometimes hard to find [**out** → out if] one has this disease.
Mec	Spelling, punctuation, capitalization, etc.	This knowledge [**maybe relavant** → may be relevant] to them.
Rloc−	Redundancy	It is up to the [**patient's own choice** → patient] to disclose information.
Cit	Citation	Poor citation practice.
Others	Other errors	An error that does not fit into any other category but can still be corrected.
Um	Unclear meaning	Genetic disease has a close relationship with the **born gene.** (i.e., no correction possible without further clarification.)

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)

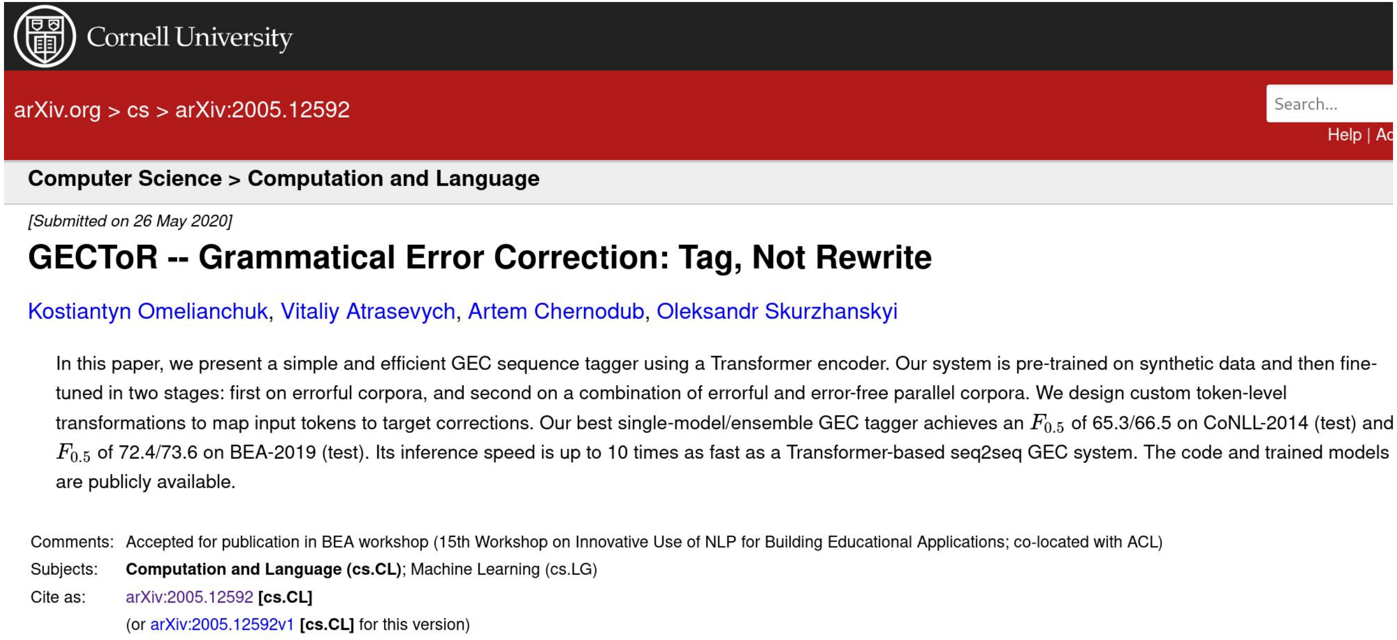
2016 - [GLEU] GLEU Without Tuning (Napoles et al., 2016)

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

Other datasets

- CoNLL-2013
- **CoNLL-2014**
- **JFLEG**
- FCE
- ICNALE
- KJ
- **BEA:** grammatical, lexical and orthographical errors

Late paper



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)

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

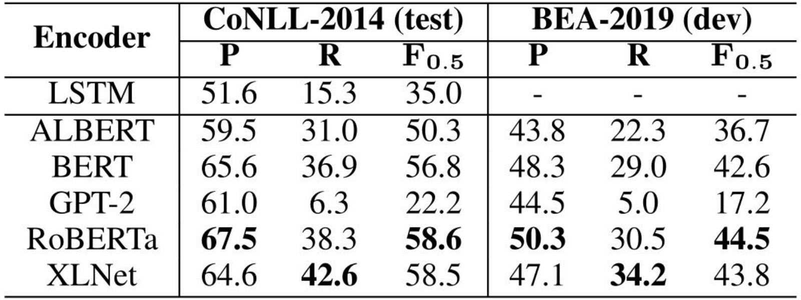
Late paper

~~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)

Code

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

O que será feito na próxima semana

- Experimentações
  - BERT (base / large)
  - T5 (small, base, large, 3b, ~~11b~~)
- Tabela com resultados
  - Qualidade
  - Custo computacional
- Proposta de arquitetura final (RoBERTa?)

Planejamento/Cronograma atualizado

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.**

Semana 3: 04/junho → 10/junho

(entrega do corretor em inglês)

~~○ Compilação dos diferentes testes produzidos na semana anterior.~~

~~○ Proposta da arquitetura final.~~

○ **Decidir o modelo a ser usado**

○ **Começar o corretor em inglês**

○ Finalização do corretor para inglês.