Grammatical Error Correction

Petre Lorena, Mihalache Stefan

Topic

- Grammatical Error Correction (GEC) is the task of correcting different kinds of errors in text such as spelling, punctuation, grammatical, and word choice errors.
- GEC is typically formulated as a sentence correction task. A GEC system takes a potentially erroneous sentence as input and is expected to transform it to its corrected version.
- By the 1990s, grammar checkers started appearing in word processing software, like Microsoft Word. These were still basic but became popular tools for writers to catch typos and simple grammatical mistakes.
- Today, advanced grammar checkers like Grammarly, ProWritingAid, and Google Docs' grammar tool use artificial intelligence (AI) and machine learning to detect complex errors, including stylistic issues, context-based corrections, and even tone adjustments.

W&I + LOCNESS dataset

Write & Improve is an online web platform that assists non-native English students with their writing. Since W&I went live in 2014, W&I annotators have manually annotated some of these submissions and assigned them a CEFR level.

The LOCNESS corpus consists of essays written by native English students. It was originally compiled by researchers at the Centre for English Corpus Linguistics at the University of Louvain. Since native English students also sometimes make mistakes, we asked the W&I annotators to annotate a subsection of LOCNESS so researchers can test the effectiveness of their systems on the full range of English levels and abilities.

This dataset includes grammatical, lexical and orthographical errors.

https://huggingface.co/datasets/bea2019st/wi_locness

Split	Stats	A	B	C	N	Total
Train	Texts	1,300	1,000	700	0	3,000
	Sentences	10,493	13,032	10,783	0	34,308
	Tokens	183,684	238,112	206,924	0	628,720
Dev	Texts	130	100	70	50	350
	Sentences	1,037	1,290	1,069	988	4,384
	Tokens	18,691	23,725	21,440	23,117	86,973
Test	Texts	130	100	70	50	350
	Sentences	1,107	1,330	1,010	1,030	4,477
	Tokens	18,905	23,667	19,953	23,143	85,668
Total	Texts	1,560	1,200	840	100	3,700
	Sentences	12,637	15,652	12,862	2,018	43,169
	Tokens	221,280	285,504	248,317	46,260	801,361

Dataset Structure

Data Instances

An example from the `wi` configuration:

```
{
  'id': '1-140178',
  'userid': '21251',
  'cefr': 'A2.i',
  'text': 'My town is a medium size city with eighty thousand inhabitants. It has a high density population bec
  'edits': {
    'start': [13, 77, 104, 126, 134, 256, 306, 375, 396, 402, 476, 484, 579, 671, 774, 804, 808, 826, 838, 850,
    'end': [24, 78, 104, 133, 136, 262, 315, 379, 399, 411, 480, 498, 588, 671, 777, 807, 810, 835, 845, 856, 8
    'text': ['medium-sized', '-', ' of', 'Although', '', 'center', None, 'of', 'is', 'commercial', 'kinds', 'bu
  }
}
```

An example from the `locness` configuration:

```
{
  'id': '7-5819177',
  'cefr': 'N',
  'text': 'Boxing is a common, well known and well loved sport amongst most countries in the world howe
  'edits': {
    'start': [24, 39, 52, 87, 242, 371, 400, 528, 589, 713, 869, 992, 1058, 1169, 1209, 1219, 1255, 130
    'end': [25, 40, 59, 95, 249, 374, 400, 538, 595, 713, 869, 1001, 1063, 1169, 1209, 1219, 1255, 1315
    'text': ['-', '-', 'in', '. However,', '. There', 'their', ',', 'among', "there's", ' and', ',', 'u
  }
}
```

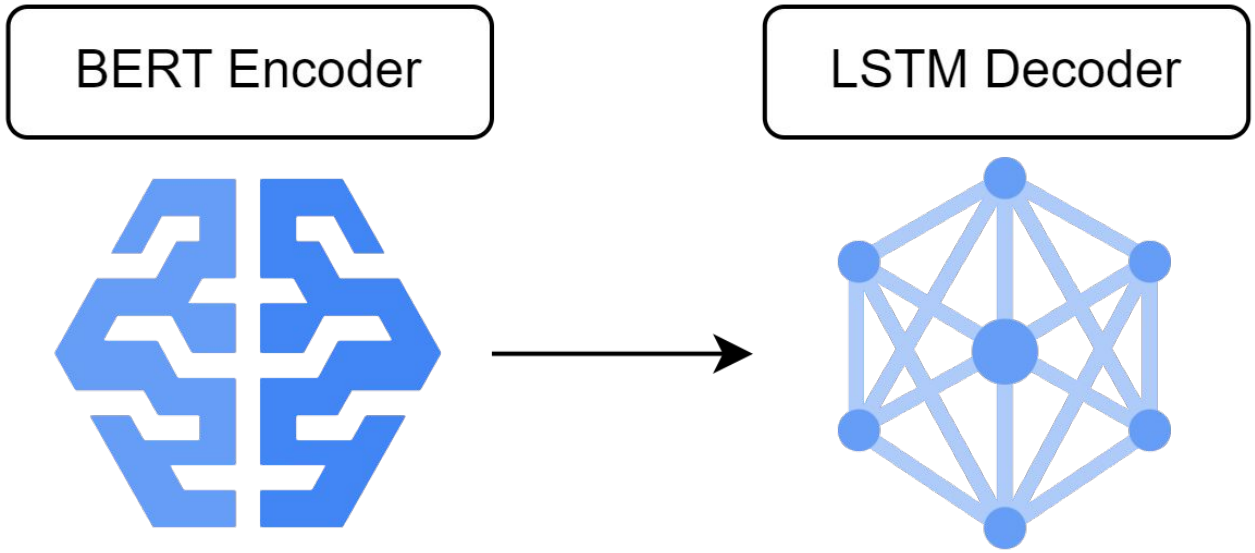
Similar solutions

Grammarly

Microsoft Editor

Ginger Software

Implementation



Any Questions?