Grammar Error Correction For Turkish

Göksu Güz

Introduction

Grammar Error Correction is one of the Natural Language Processing application areas. In this project, model training was emphasized to make an application that corrects Turkish grammar rules. Different models have been trained, and the training result statistics of a model that gives better results than other models in terms of laying the groundwork for future studies have been investigated.

Methods

Firstly, Turkish datasets were collected for model training, and some data were obtained from the platforms such as GitHub, Huggingface.

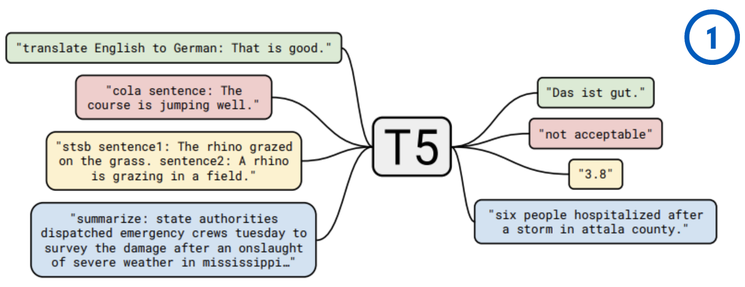
PREPARATION FOR FINE-TUNING

- Grammatically incorrect and corrected versions of sentences are collected.
- Datasets in which incorrect and corrected sentences were included were transformed into sequence-to-sequence labeled datasets.
- From datasets that include only grammatically correct sentences, grammatically incorrect sentences obtained by adding noise, and changes were made to these datasets to be used in model training.

USED MODEL (mT5-base Model)

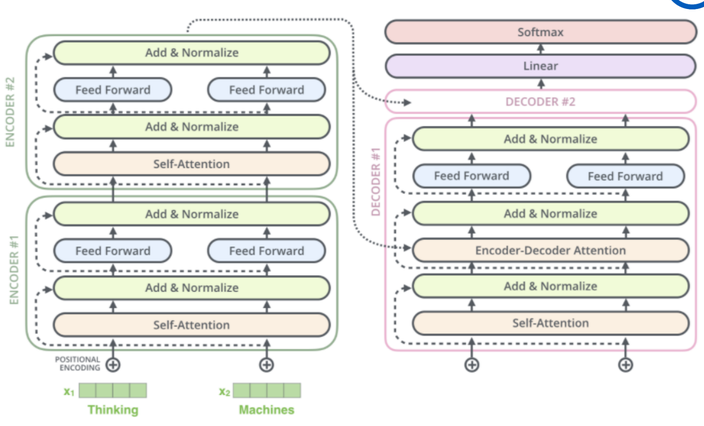
- **Multilingual Text-to-Text Transfer Transformer.**
- Pre-trained on masked language modeling.
- Text-to-text formatted.
- Sequence-to-sequence structured.
- Its vocabulary covers over 100 languages.

1



MT5 Multilingual Text-to-Text Transfer Transformer Representation [1]

2

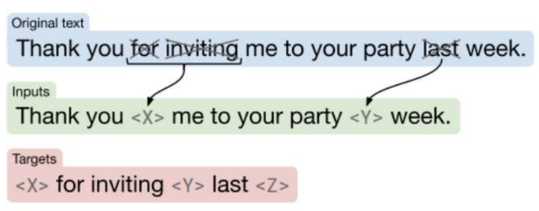


mT5 Model Structure [1]

UNSUPERVISED OBJECTIVE AND ARCHITECTURE

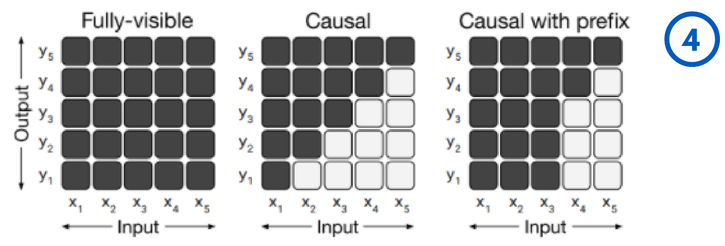
The basic logic behind the training of the model is to predict sentinel tokens to represent the dropped out text.

3

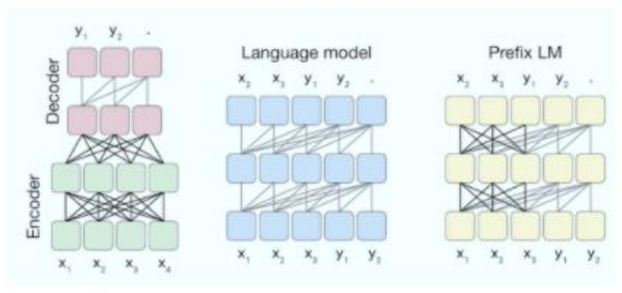


The unsupervised objective [1]

4



5



Different Attention Mask Patterns (upper) and Corresponding Models (bottom) [2]

FINE-TUNING

In order to collect data, it was necessary to categorize the common mistakes made in Turkish. In this sense, these errors were tried to be classified by using different sources. It was initially planned that there were six error categories [See Figure A]. However, on the similarities between these error categories, some errors were combined within themselves, and new categories were created [See Figure B]. Moreover, grammatically incorrect sentences are collected according to these categories, and sequence-to-sequence model compatible datasets are created. You can see a part of the sentences given to the datasets [See Figure C].

A

1- Grammatical	==>	Aileme çok özlüyorum.	==>	Aliemi çok özlüyorum.
2- Spelling	==>	Türksye çok güzel bir ülke.	==>	Türkiye çok güzel bir ülke.
3- Punctuation	==>	Eve mi gidiyorsun.	==>	Eve mi gidiyorsun?
4- Lexical	==>	Çok hoşuma geldi.	==>	Çok hoşuma gitti.
5- Syntactic	==>	Türkiye çok bir güzel ülke.	==>	Türkiye çok güzel bir ülke.
6- Capitalization	==>	eve geliyorum.	==>	Eve geliyorum.

B

1- (Sentence-Level) Grammatical, Syntactic, Punctuation, Clitic Errors

2- (Word-level) Spelling, Normalization, Capitilizaton

3- Lexical/synonym Replacement

C

Uncorrected Sentence	Corrected Sentence
Haklı enazından.	Haklı en azından.
Süper gç olmak böle bir şe işte.	Süper güç olmak böyle bir şey işte.
1 sene daha yajlaştık arkadaşlar.	1 sene daha yaklaştık arkadaşlar.
Milletvekillerine sira e zaman gelecek?	Milletvekillerine sıra ne zaman gelecek?
Çooooook temiz afiyet olsun.	Çok temiz afiyet olsun.
Evt ve hyr ne demek?	Evet ve hayır ne demek?

Fine-Tuned Model Results

Some of the model output examples are given below, there are three fine-tuned models, evaluations are made on these, and the outputs are taken from the last model, due to its larger training dataset.

Model Outputs

Evaluation Results

For Evaluation, Errant (Error Annotation Tool) [3] and Eryiğit &Torunoğlu's [4] publications, were searched. Errant is a rule-based evaluation tool written only in English, it was necessary to write the rule-based system for Turkish in order to evaluate with this tool, so Eryiğit & Torunoğlu's method was preferred.

- Evaluation is made by calculating Precision, Recall and F1-score values.
- Dataset taken from “de-da Takıntısı” [5] contains 100 sentences.

Data sets	Sentences	Tokens	Normalized tokens
Test $_{\text{small}}$	508	6,489	$1,275(19.6\%)$
BTS	23,526	283,312	$60,417(21.3\%)$
IWT	4828	38,760	$5,933(15.3\%)$

Statistics of the Datasets Used For Evaluation

Trained With	Tested With	Precision	Recall	F1-Score
”de-da” Clitics only Dataset	”de-da only dataset”	0.62	0.39	0.48
”de-da” Clitics only Dataset + Mixed Error Dataset	IWT	0.63	0.30	0.40
”de-da” Clitics only Dataset + Mixed Error Dataset	BTS	0.52	0.19	0.28
”””de-da” Clitics only Dataset + Mixed Error Dataset	Validation Dataset	0.71	0.32	0.44
”de-da” Clitics only Dataset + Mixed Error Dataset	Test Small Dataset	0.68	0.24	0.35
”de-da” Clitics only Dataset + Mixed Error Dataset + Diacritization Dataset + Vowel Restoration	IWT	0.69	0.35	0.46
”de-da” Clitics only Dataset + Mixed Error Dataset + Diacritization Dataset + Vowel Restoration	BTS	0.61	0.22	0.32
”de-da” Clitics only Dataset + Mixed Error Dataset + Diacritization Dataset + Vowel Restoration	Validation Dataset	0.74	0.40	0.52
”de-da” Clitics only Dataset + Mixed Error Dataset + Diacritization Dataset + Vowel Restoration	Test Small Dataset	0.71	0.28	0.40

Fine-Tuned Models on Different Datasets and Their Performances

olmali	olmalı
cekmiyor	çekmiyor
arkadas	arkadaş

Diacritization Error Examples

seviyrm	seviyorum
benm	benim
canm	canım

Vowelization Error Examples

Trained With	Tested With	Evaluated Error Type	Error Detection Rate	Error Correction Rate
”de-da” Clitics Only Dataset + Mixed Error Dataset + Diacritization Dataset	IWT	Diacritization	0.65	0.52
”de-da” Clitics Only Dataset + Mixed Error Dataset + Diacritization Dataset	IWT	Vowel Restoration	0.59	0.43
”de-da” Clitics Only Dataset + Mixed Error Dataset + Diacritization Dataset	BTS	Diacritization	0.55	0.39
”de-da” Clitics Only Dataset + Mixed Error Dataset + Diacritization Dataset	BTS	Vowel Restoration	0.60	0.47
”de-da” Clitics Only Dataset + Mixed Error Dataset + Diacritization Dataset	Test Small	Diacritization	0.73	0.62
”de-da” Clitics Only Dataset + Mixed Error Dataset + Diacritization Dataset	Test Small	Vowel Restoration	0.64	0.53

Fine-Tuned Models on Different Datasets and Their Performances On Vowelization and Diacritization Errors.

Preparation For Future Work

In this project, not only the mT5-base model was investigated, but also an attempt was made to create a base study for future studies. As can bee seen in the table below, Turkish equivalents of the M2 format used in English were prepared. Also, Turkish dataset research that are available for GEC applications is made and table created for future use.

Code	Category	Description	Example
ADJ	Adjective	Wrong choice of adjective	büyük → geniş
ADJ:FORM	Adjective Form	Wrong usage of comparative or superlative adjective	çok en büyük → çok daha büyük
ADV	Adverb	Wrong choice of adverb	-
CONJ	Conjunction	Wrong choice of conjunction	ama → ve
CONTR	Contraction	Wrong usage of contraction	arayor → arıyor
DET	Determiner	Wrong choice of determiner	bir ev → bu ev
MORPH	Morphology	Tokens have the same lemmabut nothing else in common	kolay → kolayca
NOUN	Noun	Wrong choice of noun	sayesinde → yüzünden
NOUN:INFL	Noun Inflection	Count-mass noun errors	sular → su
NOUN:NUM	Noun Number	Wrong usage of noun number	ders → dersler
NOUN:POSS	Noun Possessive	Wrong usage of noun possessive	arkadaş → arkadaşın
ORTH	Orthography	Case and/or whitespace errors	herşey → her şey
OTHER	Other	Errors that do not fall into any other category	–
PART	Participle	Wrong choice of participle	ödev çalıştı → ödev yaptı
PREP	Preposition	Wrong choice of preposition	aşağısında → altında
PRON	Pronoun	Wrong usage of pronoun	kendimin → benim
PUNCT	Punctuation	Wrong usage of punctuation	,→ .
SPELL	Spell	Misspelling	herkez → herkes
VERB	Verb	Wrong choice of verb	görmek → bakmak
VERB:FORM	Verb Form	Infinitives, gerunds and participles	etkileyici → etkili
VERB:INFL	Verb Inflection	Wrong usage of tense morphology	gidiyormuştum → gidiyormuşum
VERB:SVA	Subject-Verb Agreement	Confliction between subject and verb	Biz yaptı. → Biz yaptık.
VERB:TENSE	Verb Tense	Inflectional and periphrastic tense, modal verbs and passivization	yaptı → yapmıştı, yapar → yapabilir, yaptı → yapıldı
WO	Word Order	scrambled word order	çok bir güzel ülke → çok güzel bir ülke

M2 Format Error Correspondings in Turkish

Conclusion

The model is open to development; its attention and encoder-decoder layers have not been changed in the model used; by increasing the epoch number, using large datasets and different configurations, the model can achieve better results. It also does not make unnecessary and incorrect changing of input words without the need for a copy mechanism because it is pretrained with a large dataset. As a result of the limited opportunities available, and because the training is carried out on platforms such as Google Colab, the number of epochs could not be higher than the latest trained model. Results that were obtained show that the mT5 model has the potential to be developed and used for Grammar Error Correction applications.

References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." J. Mach. Learn. Res. 21.140 (2020): 1-67.
2. Xue, Linting, et al. "mT5: A massively multilingual pre-trained text-to-text transformer." arXiv preprint arXiv:2010.11934 (2020).
3. Bryant, Christopher, Mariano Felice, and Edward Briscoe. "Automatic annotation and evaluation of error types for grammatical error correction." Association for Computational Linguistics, 2017.
4. ERYİĞİT, GÜLŞEN, and D. İ. L. A. R. A. TORUNOĞLU-SELAMET. "Social media text normalization for Turkish." Natural Language Engineering 23.6 (2017): 835-875.
5. @misc{-de/-da takıntısı, title={-DE/-Da Yazım Hatası bulucu}, url={http://dedatakintisi.derlem.com/deda/bul/}, journal={-de/-da Takıntısı}}

İTÜ
ISTANBUL TECHNICAL UNIVERSITY
Department of Computer Engineering
Faculty of Computer and Informatics Engineering
İTÜ Ayazaga Campus, 34469, Maslak/Istanbul
bbf@itu.edu.tr
http://bm.itu.edu.tr