

Deep Learning Project Final Presentation

Urdu Grammar Error Correction

اردو گرامر کی اصلاح

Presented By:

Ahmed Ali, Burhanuddin Aliasgher, Syed Ahad Ali

Research Question

تحقیقی سوال

How to generate synthetic (error + correct sentence pair) dataset for a low-resource language like Urdu?

How can human errors be mimicked in clean text?

Research Question

تحقيقی سوال

- Why not just substitute words randomly for generating such data? Why is it important to mimic human-made grammatical errors
 - There is a pattern to which we humans make grammatical errors. Capturing that pattern and structure makes it easier to train a DNN for such a task
 - A DNN would need to search an exponentially larger space which makes training infeasible

Set of all types of errors that can be inflicted on a sentence

Set of Errors
actually made by
humans

Required Result

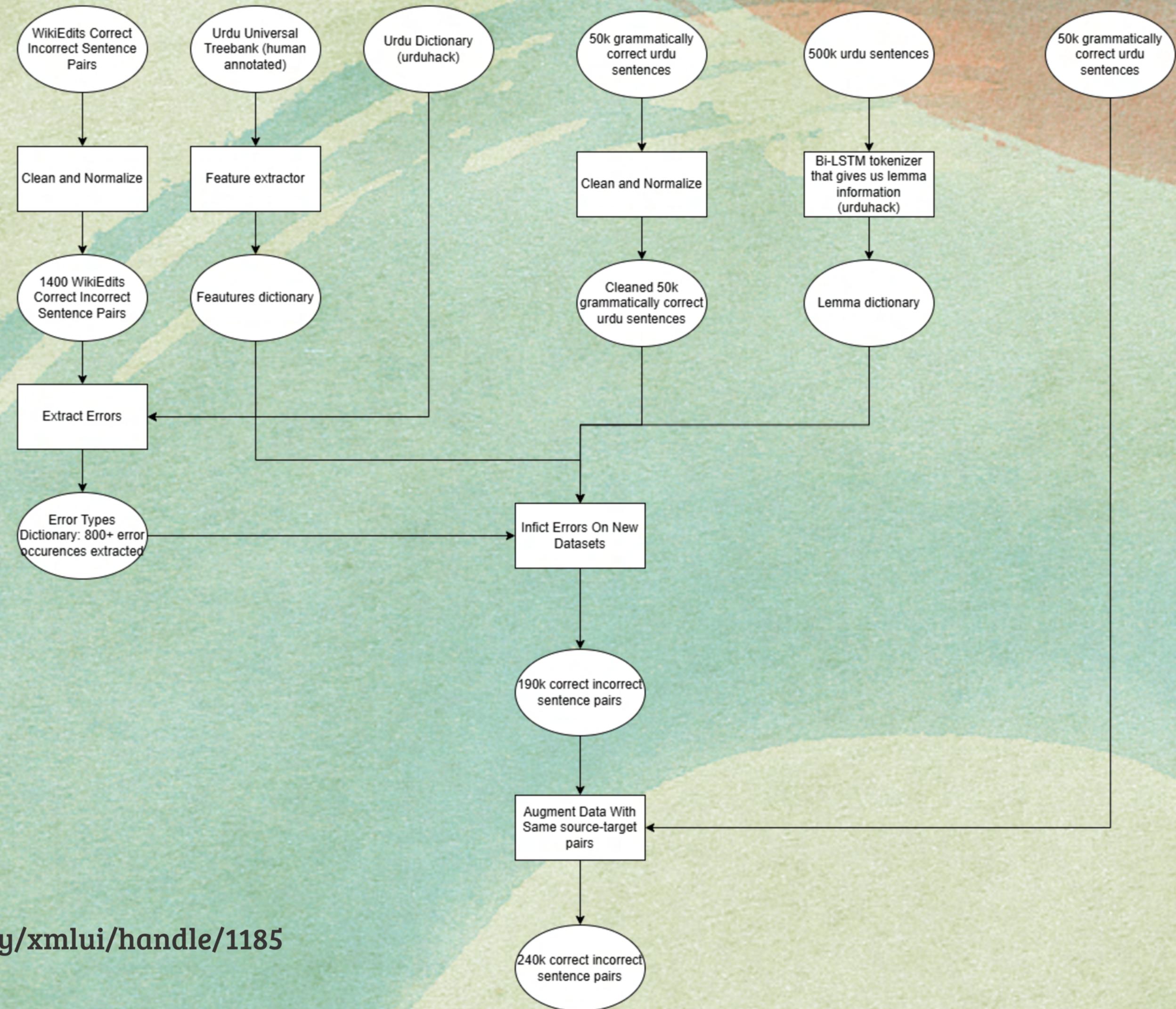
مطلوبہ نتیجہ

"وہ کتاب پڑھ ریا ہے اور میں بھی پڑھ ریا ہے"



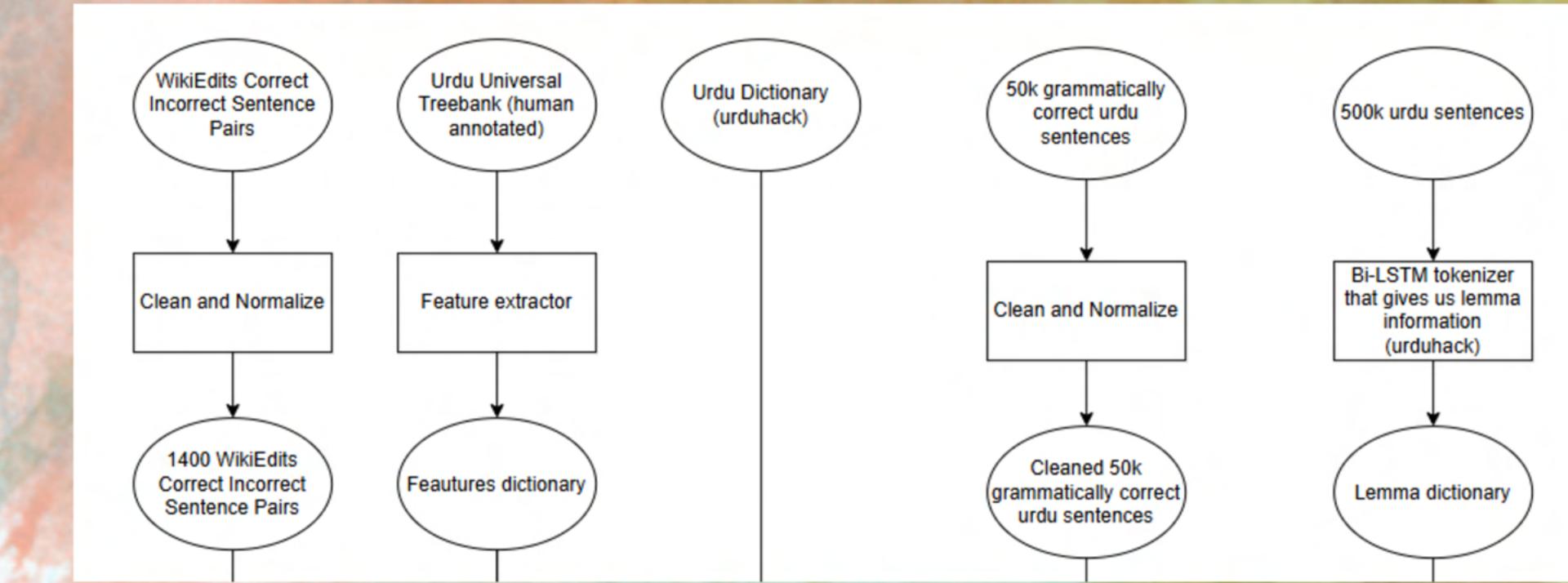
"وہ کتاب پڑھ ریا ہے اور میں بھی پڑھ ریا ہوں"

Synthetic Data Generation Pipeline



Main Data Corpus:
<https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-65A9-5>

SYNTHETIC ERROR GENERATION PIPELINE: PREPROCESSING



"کتاب": [
"کتابوں",
"کتاب",
"کتابیں"]
،
"والا": [
"والے",
"والا",
"الوں",
"الو"
]،

"آؤ": [
{"
"id": "13",
"text": "آؤ",
"lemma": "آٹا",
"upos": "VERB",
"xpos": "VAUX",
"feats": "Mood=Imp|Number=Plur|Person=3",
"head": "12",
"deprel": "compound",
"deps": "_",
"misc": "ChunkId=VGF2|chunkType=child|LT"}
]

Feature Dictionary

Lemma Dictionary

داخلے میں کوئی مسئلہ درپیش ہے دوبارہ اندرج کریں
بننے والی برقی ڈاک کے پتے کیلیے بھیج دیا گیا ہے
مول بو جا تو براہ کرم اسکے ذریعے دوبارہ داخل ہوں
ترمیم و تدوین کیلیے آپ کا لا زمی ہے
کلم شناخت بھیج دیا گیا

Wikiedits correct sequence

داخلے میں کوئی مسئلہ درپیش ہے دوبارہ اندرج کیجیے
کے نام سے بننے والی برقی ڈاک کے پتے کیلی بھیج دیا گیا ہے
جب وہ مومول بو جا تو براہ کرم اسکے ذریعے دوبارہ داخل ہوں
ترمیم و تدوین کے لی آپکا لا زمی ہے
کلم شناخت بھیج دیا گیا

Wikiedits incorrect sequence

SYNTHETIC ERROR GENERATION PIPELINE: ERROR EXTRACTION

ALIGNMENT

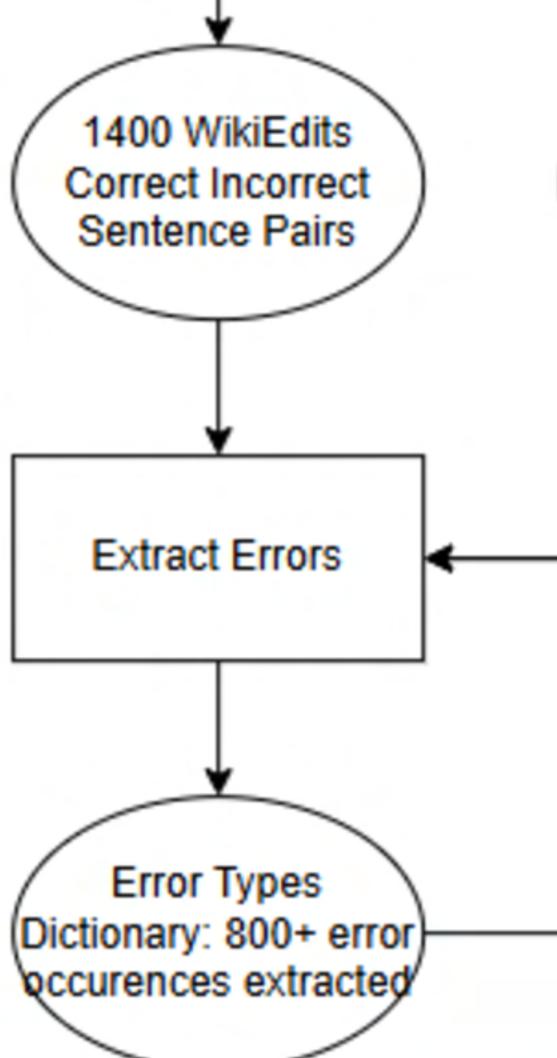
وہ کتاب پڑھ رہا ہیں۔

وہ کتاب پڑھ رہا ہے۔

کتاب پڑھ رہا ہیں۔

وہ کتاب پڑھ رہا ہے۔

```
[('I', 0, 0, 0, 1), ('M', 0, 1, 1, 2), ('M', 1, 2, 2, 3), ('M', 2, 3, 3, 4), ('S', 3, 4, 4, 5)]
```

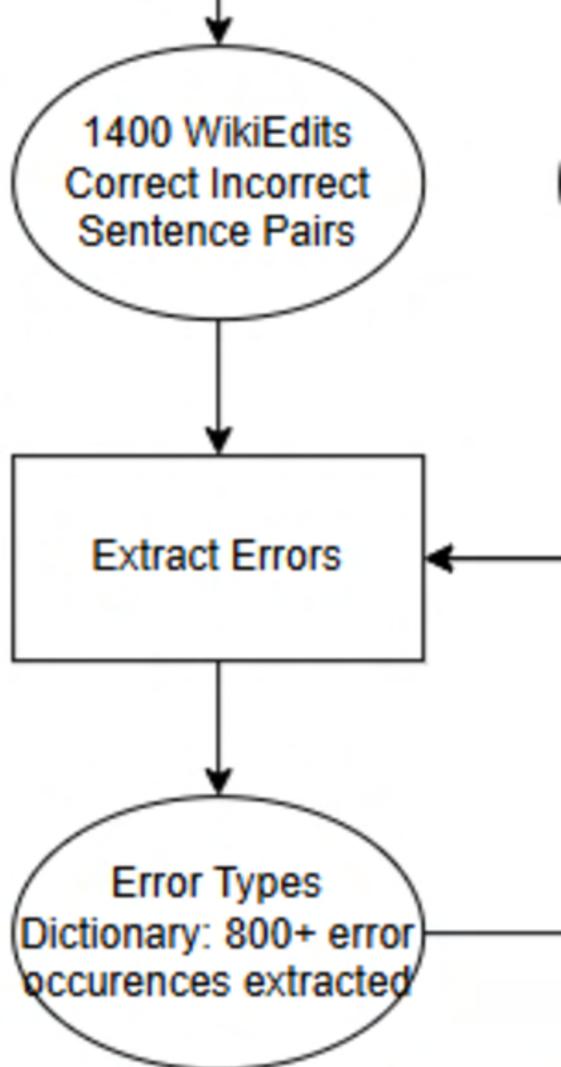


SYNTHETIC ERROR GENERATION PIPELINE: ERROR EXTRACTION

Motivation:
ERRANT

Code	Meaning	Description / Example
ADJ	Adjective	<i>big</i> → <i>wide</i>
ADJ:FORM	Adjective Form	Comparative or superlative adjective errors. <i>goodest</i> → <i>best</i> , <i>bigger</i> → <i>biggest</i> , <i>more easy</i> → <i>easier</i>
ADV	Adverb	<i>speedily</i> → <i>quickly</i>
CONJ	Conjunction	<i>and</i> → <i>but</i>
CONTR	Contraction	<i>n't</i> → <i>not</i>
DET	Determiner	<i>the</i> → <i>a</i>
MORPH	Morphology	Tokens have the same lemma but nothing else in common. <i>quick (adj)</i> → <i>quickly (adv)</i>
NOUN	Noun	<i>person</i> → <i>people</i>
NOUN:INFL	Noun Inflection	Count-mass noun errors. <i>informations</i> → <i>information</i>
NOUN:NUM	Noun Number	<i>cat</i> → <i>cats</i>
NOUN:POSS	Noun Possessive	<i>friends</i> → <i>friend's</i>
ORTH	Orthography	Case and/or whitespace errors. <i>Bestfriend</i> → <i>best friend</i>
OTHER	Other	Errors that do not fall into any other category (e.g. paraphrasing). <i>at his best</i> → <i>well</i> , <i>job</i> → <i>professional</i>
PART	Particle	<i>(look) in</i> → <i>(look) at</i>
PREP	Preposition	<i>of</i> → <i>at</i>
PRON	Pronoun	<i>ours</i> → <i>ourselves</i>
PUNCT	Punctuation	<i>! → .</i>
SPELL	Spelling	<i>genetic</i> → <i>genetic</i> , <i>color</i> → <i>colour</i>
UNK	Unknown	The annotator detected an error but was unable to correct it.
VERB	Verb	<i>ambulate</i> → <i>walk</i>
VERB:FORM	Verb Form	Infinitives (with or without "to"), gerunds (-ing) and participles. <i>to eat</i> → <i>eating</i> , <i>dancing</i> → <i>danced</i>
VERB:INFL	Verb Inflection	Misapplication of tense morphology. <i>getted</i> → <i>got</i> , <i>fliped</i> → <i>flipped</i>
VERB:SVA	Subject-Verb Agreement	<i>(He) have</i> → <i>(He) has</i>
VERB:TENSE	Verb Tense	Includes inflectional and periphrastic tense, modal verbs and passivization. <i>eats</i> → <i>ate</i> , <i>eats</i> → <i>has eaten</i> , <i>eats</i> → <i>can eat</i> , <i>eats</i> → <i>was eaten</i>
WO	Word Order	<i>only can</i> → <i>can only</i>

Table 2: The list of 25 main error categories in our new framework with examples and explanations.

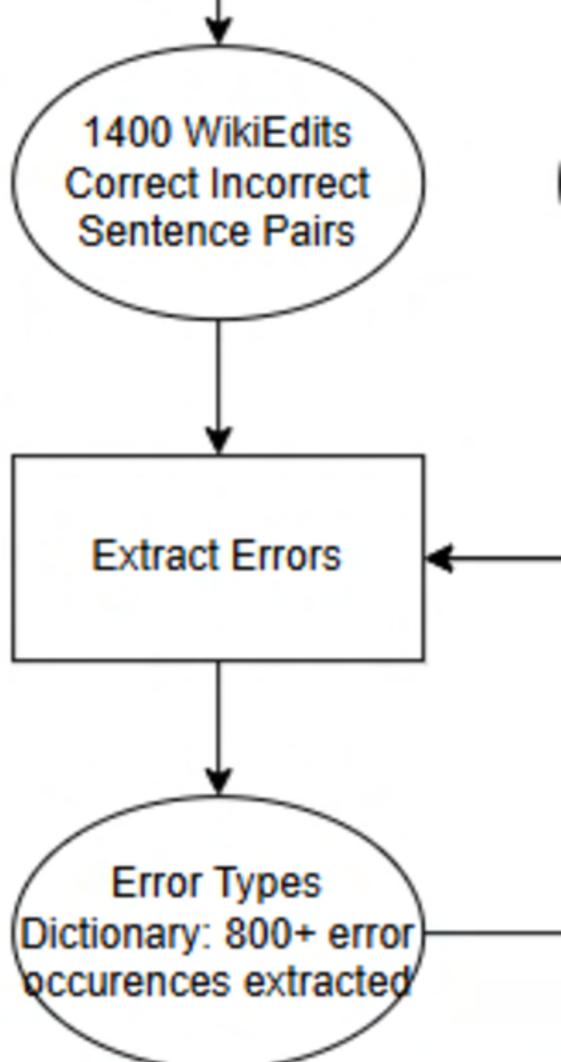


SYNTHETIC ERROR GENERATION PIPELINE: ERROR EXTRACTION

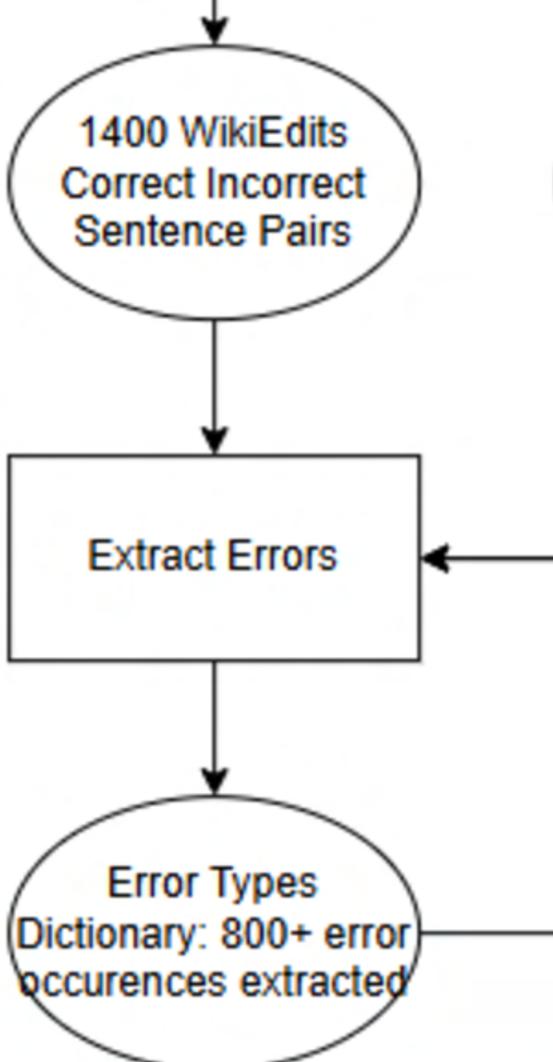
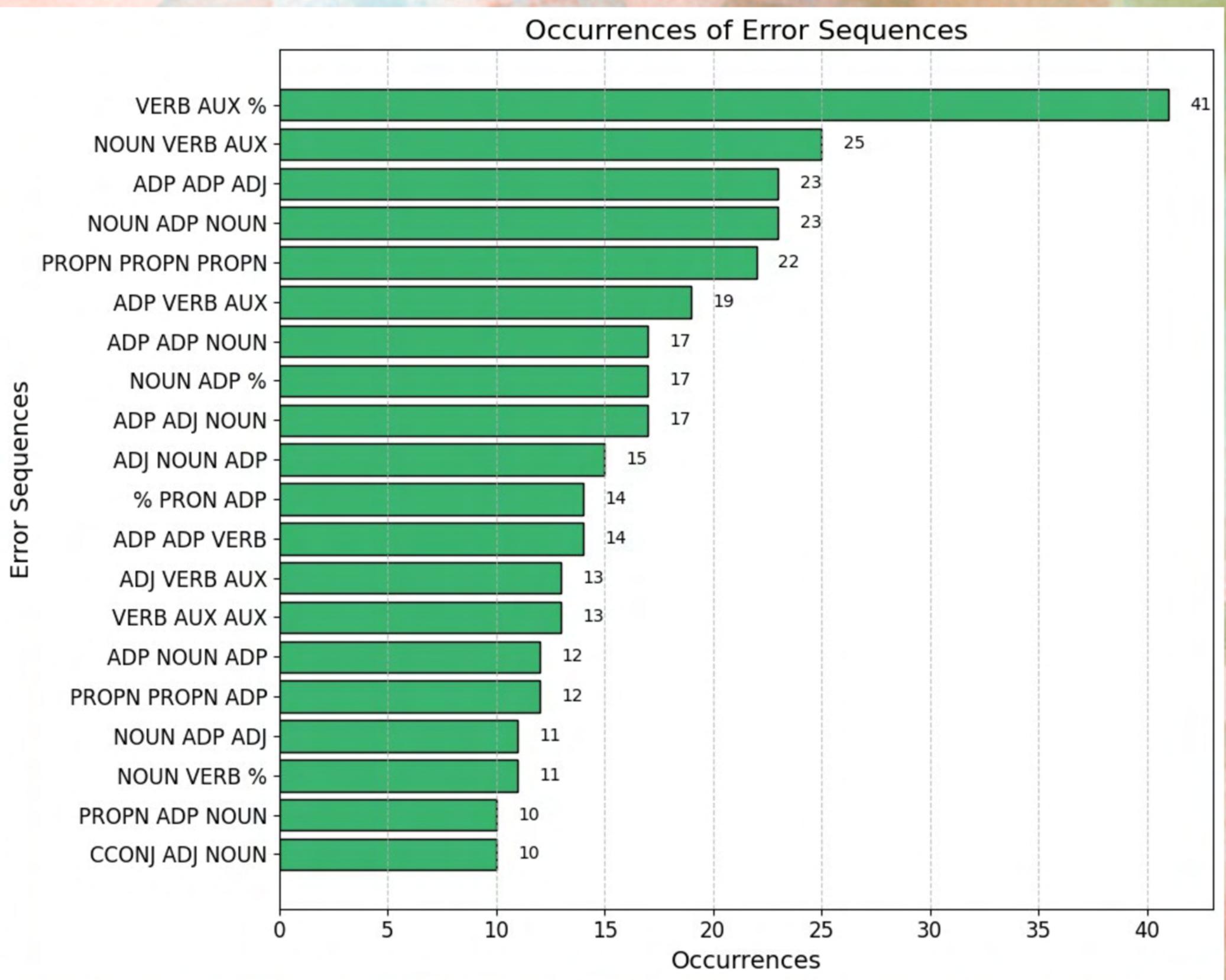
OUR TYPE ANNOTATION

1. Sentence structure
2. Features like:
 - a. Gender
 - b. Tense
 - c. Form
 - d. Norm

```
"ADP_ADP_ADJ-1-15e04819": {  
    "type": "S",  
    "incorrect_word_upos": "ADP",  
    "correct_word_upos": "ADP",  
    "incorrect_feats": "Gender=Masc|AdpType=Post|VerbForm=Part|Number=Sing|Case=Acc|As",  
    "correct_feats": "Gender=Masc|AdpType=Post|VerbForm=Part|Polite=Form|Number=Sing|V",  
    "occurrence": 18,  
    "id": "ADP_ADP_ADJ-1-15e04819",  
    "percentage": 2.2004889975550124  
},
```



SYNTHETIC ERROR GENERATION PIPELINE: ERROR EXTRACTION



SYNTHETIC ERROR GENERATION PIPELINE: ERROR INFILCTION EXAMPLE

+ 4 additions

5 lines Copy

1 نہیں نوید ان میں سے کوئی بھی لطیفہ میں ے یوٹ نہیں کیا اس لئے اس لئے اینی قلمی صلاحیت کا مظاہرہ کریں۔

2 انٹرنیٹ اکسیلوور اس کے مقابلے میں بہت نیزی سے جل رہا ہے۔

3 اس کا بہتر جو اب تو قاعر صاحب ہی دے سکتے ہیں

4 فیٹہ لگانا، جوڑن، اجز بندی کرنا، م جلد، انتظام، سلسلہ سوچل نیٹ ورکنگ سائٹ ٹوٹر یہ امیتابھ کے جانے والوں کی تعداد ۷۰ لاکھ سے تجاوز کر جکی ہے۔

5 اگر نمسائینسی مادہ کی بات کر رہے ہو تو غیر مادہ، ور نہ؟

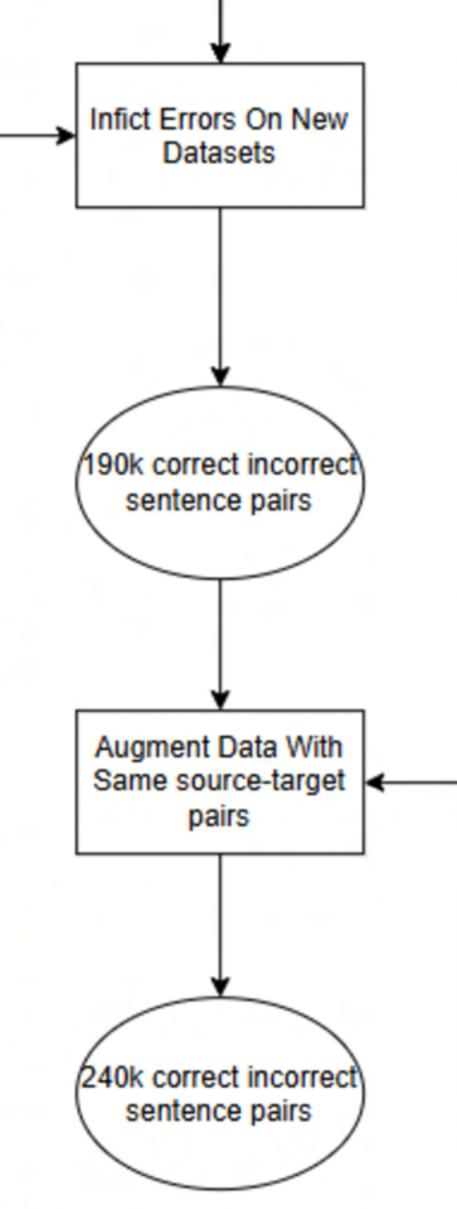
1 نہیں نوید ان میں سے کوئی بھی لطیفہ میں ے یوٹ نہیں کیا اس لئے اس لئے اینی قلمی صلاحیت کا مظاہرہ کریں۔

2 انٹرنیٹ اکسیلوور اس کے مقابلے میں بہت نیزی سے جل رہا ہے۔

3 وہ کا بہتر جو اب تو قاعر صاحب ہی دے سکتے ہیں

4 فیٹہ لگانا، جوڑن، اجز بندی کرنا، م جلد، انتظام، سلسلہ سوچل نیٹ ورکنگ سائٹ ٹوٹر یہ امیتابھ کی جانے والوں کی تعداد ۷۰ لاکھ سے تجاوز کر جکی ہے۔

5 اگر نمسائینسی مادہ کی بات کر رہے ہو تو غیر مادہ، ور نہ؟



Results

Performance Metrics

GLEU Score

GLEU score is simply the minimum of recall and precision. This GLEU score's range is always between 0 (no matches) and 1 (all match) and it is symmetrical when switching output and target.

$$\text{GLEU} = \frac{\sum_{n=1}^N \min(\text{Count}_{\text{prediction}}(n), \text{Count}_{\text{true}}(n))}{\sum_{n=1}^N \text{Count}_{\text{true}}(n)}$$

F0.5

The F0.5 score is the weighted harmonic mean of the precision and recall (given a threshold value). Unlike the F1 score, which gives equal weight to precision and recall, the F0.5 score gives more weight to precision than to recall.

$$F0.5 = (1 + 0.5^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{0.5^2 \cdot \text{Precision} + \text{Recall}}$$

Finetuning - Model

bigscience/mt0-base:

Summary from Hugging Face - We present BLOOMZ & mT0, a family of models capable of following human instructions in dozens of languages zero-shot. We finetune BLOOM & mT5 pretrained multilingual language models on our crosslingual task mixture (xP3) and find our resulting models capable of crosslingual generalization to unseen tasks & languages.

Our Dataset

Synthentic Dataset
240K Pairs

Gold Dataset
12K Pairs

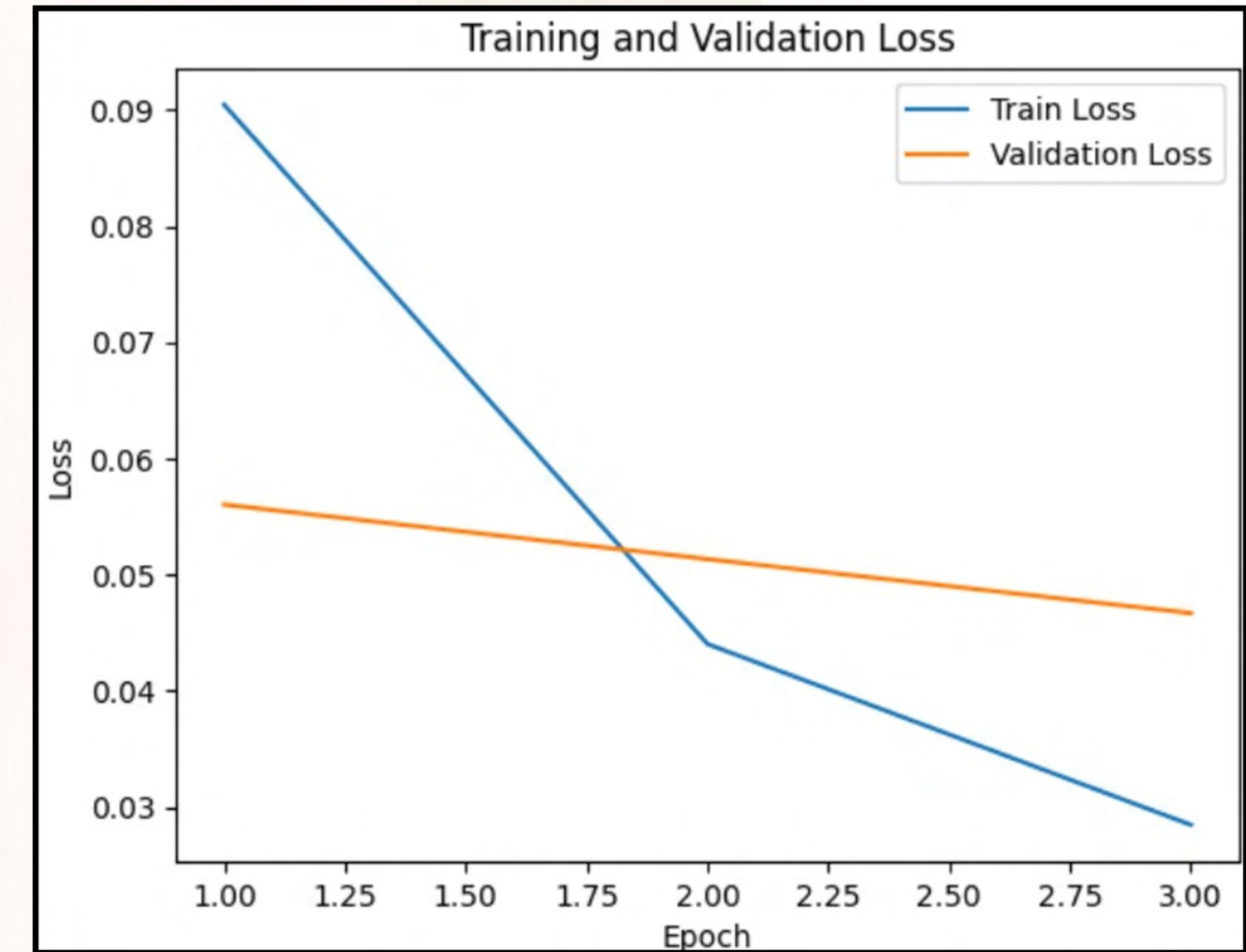
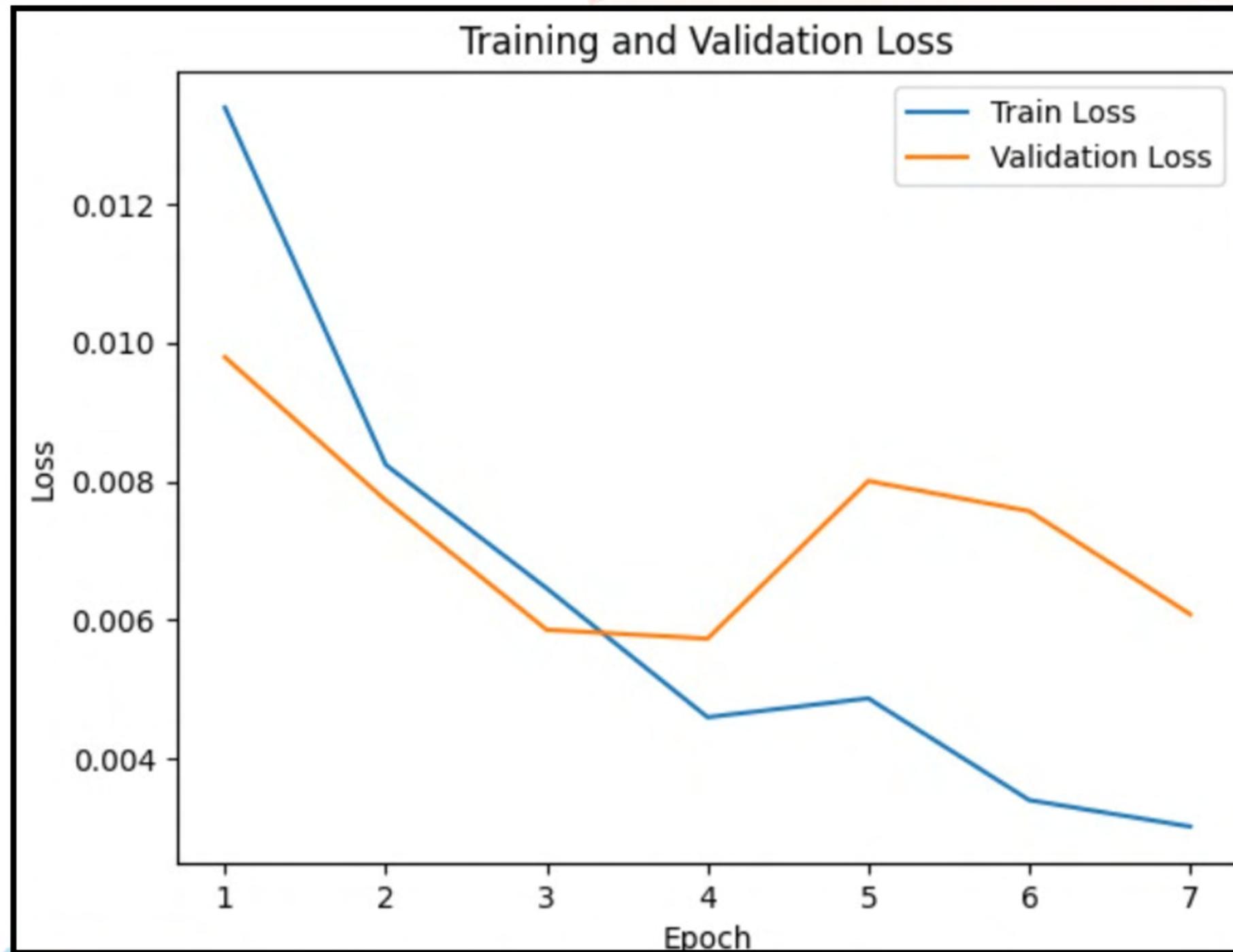
Results - Last Year

Manual GLEU (160 epochs)	Manual F0.5 (160 epochs)	Manual + Synthetic GLEU (160 + 80 epochs)	Manual + Synthetic F0.5 (160 + 80 epochs)
0.61	0.47	0.72	0.73

Our Results

Epochs	Data Length	Train Split	Validation Split	Test Split	Testing Done On	GLEU	F0.5	Precision	Recall
Baseline						0.243	0.228	0.522	0.322
1	240163	90%	10%	N/A	Gold	0.762	0.668	0.765	0.782
7	240163	70%	10%	20%	Synthetic	0.993	0.987	0.991	0.992
7	240163	90%	10%	N/A	Gold	0.749	0.653	0.745	0.765
3	20000	70%	10%	20%	Synthetic	0.948	0.951	0.963	0.964
3	20000	90%	10%	N/A	Gold	0.791	0.716	0.803	0.813

Loss Graphs



Next Steps

- Do research on training and architectures used by grammar correction models
- Increase the dataset size and improve the data generation pipeline
- Improve test dataset

Citations:

- [Corpora Generation for Grammatical Error Correction] (<https://aclanthology.org/N19-1333>) (Lichtarge et al., NAACL 2019)
- [GenERRate: Generating Errors for Use in Grammatical Error Detection] (<https://aclanthology.org/W09-2112>) (Foster & Andersen, BEA 2009)
- [Using Wikipedia Edits in Low Resource Grammatical Error Correction] (<https://aclanthology.org/W18-6111>) (Boyd, WNUT 2018)
- [WikiAtomicEdits: A Multilingual Corpus of Wikipedia Edits for Modeling Language and Discourse] (<https://aclanthology.org/D18-1028>) (Faruqui et al., EMNLP 2018)
- [A Low-Resource Approach to the Grammatical Error Correction of Ukrainian] (<https://aclanthology.org/2023.unlp-1.14>) (Palma Gomez et al., UNLP 2023)
- A. Solyman, Z. Wang and Q. Tao, "Proposed Model for Arabic Grammar Error Correction Based on Convolutional Neural Network," 2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), Khartoum, Sudan, 2019, pp. 1-6, doi: [10.1109/ICCCEEE46830.2019.9071310](https://doi.org/10.1109/ICCCEEE46830.2019.9071310).
- [Generating Inflectional Errors for Grammatical Error Correction in Hindi] (<https://aclanthology.org/2020.aacl-srw.24>) (Sonawane et al., ACL 2020)