Return_Zero at Bhashabhrom: Bangla Grammatical Error Detection Leveraging Transformer-based Token Classification

Shayekh Bin Islam Ridwanul Hasan Tanvir Sihat Afnan

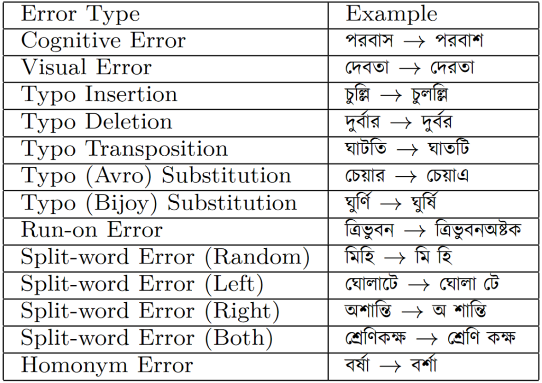
Department of CSE, BUET

March 18, 2023

Errors In Bangla Text

Errors In Bangla Text

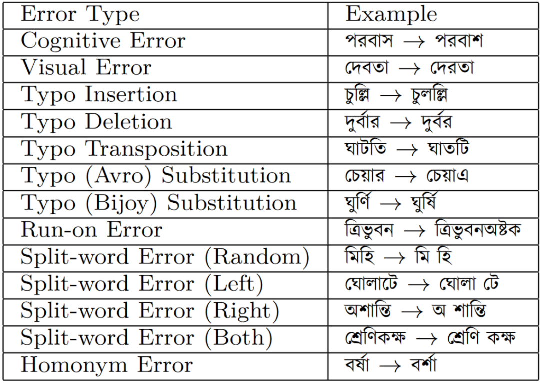
TABLE I
Types of Spelling Errors



Error Type	Example
Cognitive Error	পরবাস → পরবাশ
Visual Error	দেবতা → দেরতা
Typo Insertion	চুল্লি → চুলল্লি
Typo Deletion	দুর্বার → দুর্বর
Typo Transposition	ঘাটতি → ঘাতটি
Typo (Avro) Substitution	চেয়ার → চেয়াএ
Typo (Bijoy) Substitution	ঘুর্ণি → ঘুর্ষি
Run-on Error	ত্রিভুবন → ত্রিভুবনঅষ্টক
Split-word Error (Random)	মিহি → মি হি
Split-word Error (Left)	ঘোলাটে → ঘোলা টে
Split-word Error (Right)	অশান্তি → অ শান্তি
Split-word Error (Both)	শ্রেণিকক্ষ → শ্রেণি কক্ষ
Homonym Error	বর্ষা → বর্শা

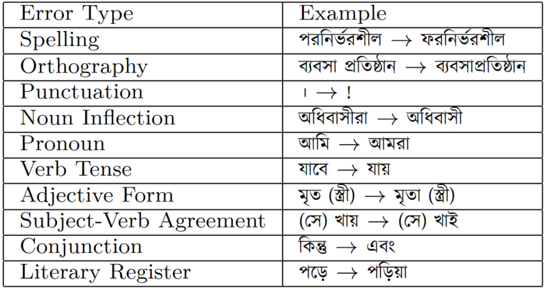
Errors In Bangla Text

TABLE I
Types of Spelling Errors



Error Type	Example
Cognitive Error	পরবাস → পরবাশ
Visual Error	দেবতা → দেরতা
Typo Insertion	চুল্লি → চুলল্লি
Typo Deletion	দুর্বার → দুর্বর
Typo Transposition	ঘাটতি → ঘাতটি
Typo (Avro) Substitution	চেয়ার → চেয়াএ
Typo (Bijoy) Substitution	ঘুর্ণি → ঘুর্ষি
Run-on Error	ত্রিভুবন → ত্রিভুবনঅষ্টক
Split-word Error (Random)	মিহি → মি হি
Split-word Error (Left)	ঘোলাটে → ঘোলা টে
Split-word Error (Right)	অশান্তি → অ শান্তি
Split-word Error (Both)	শ্রেণিকক্ষ → শ্রেণি কক্ষ
Homonym Error	বর্ষা → বর্শা

TABLE II
Errors by ERRANT Classification



Error Type	Example
Spelling	পরনির্ভরশীল → ফরনির্ভরশীল
Orthography	ব্যবসা প্রতিষ্ঠান → ব্যবসাপ্রতিষ্ঠান
Punctuation	। → !
Noun Inflection	অধিবাসীরা → অধিবাসী
Pronoun	আমি → আমরা
Verb Tense	যাবে → যায়
Adjective Form	মৃত (স্ত্রী) → মৃতা (স্ত্রী)
Subject-Verb Agreement	(সে) খায় → (সে) খাই
Conjunction	কিন্তু → এবং
Literary Register	পড়ে → পড়িয়া

The Task

- To detect sub-strings of a Bangla text that contain grammatical, punctuation, or spelling errors.

The Task

- To detect sub-strings of a Bangla text that contain grammatical, punctuation, or spelling errors.

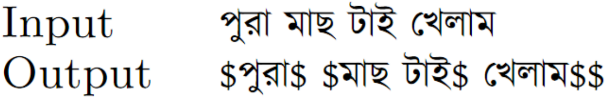
  For example:

Input	পুরা মাছ টাই খেলাম
Output	$পুরা$ $মাছ টাই$ খেলাম$$

The Task

- To detect sub-strings of a Bangla text that contain grammatical, punctuation, or spelling errors.

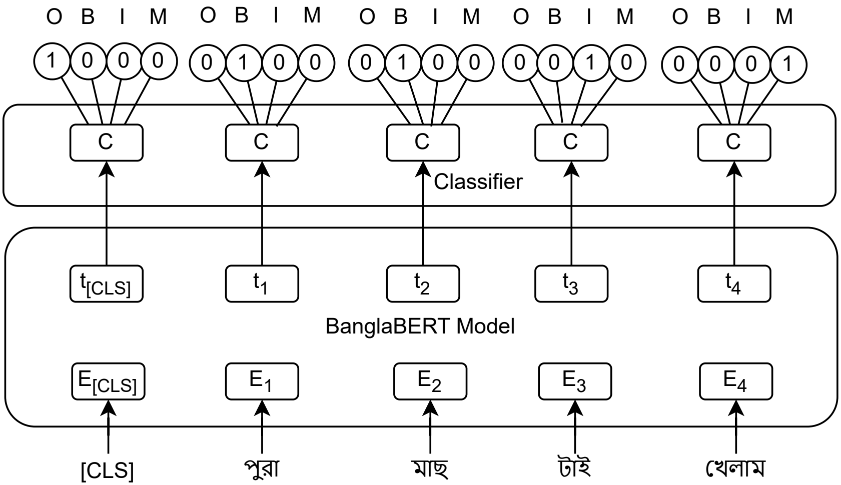
  For example:



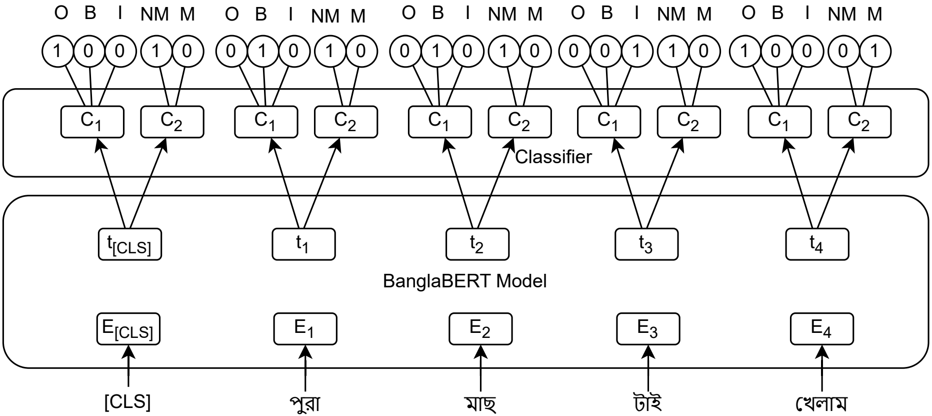
Input	পুরা মাছ টাই খেলাম
Output	$পুরা$ $মাছ টাই$ খেলাম$$

- Data
  - Training data: Around 20,000 texts with 7500 errors
  - Test data: 5,000 texts

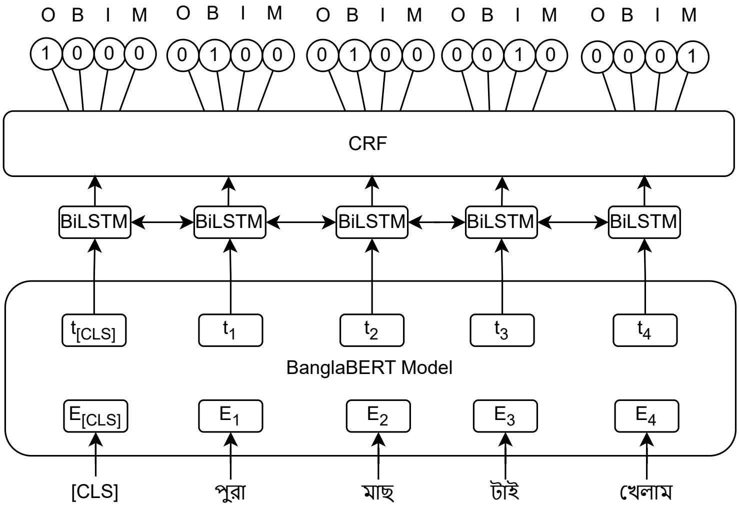
Token Classification Model I: 4 Classes



Token Classification Model II: 3+2 Classes



Token Classification Model III: LSTM-CRF



Token Classification Models

- We observe no performance gain
  - By adding BiLSTM and CRF
  - By modelling missing error separately

Table: Comparison of Token Class. Models on Dev Set

Model	Levenstein Distance
BanglaBERT-base 4 Classes	**1.0239**
BanglaBERT-base 3+2 Classes	1.0743
BanglaBERT-base+BiLSTM+CRF	1.0534

Pre-Trained Models

- We find BanglaBERT-base and BanglaBERT-large to perform best.

Table: Comparison of Transformers Models on Private Test Set

Model	Leventstein Distance
XLM-RoBERTa-base	1.3940
DeBERTa-V3-large	1.3552
BanglaBERT-base	1.2120
BanglaBERT-large	**1.1844**

Label Smoothing

- Mitigates overfitting and noise-modelling.
- 0.1 for BanglaBERT-base and 0.2 for BanglaBERT-large.

Table: Private test set results for label smoothing

Type	Levenstein Distance
BanglaBERT-large+standard CE	1.1640
BanglaBERT-large+smoothing factor 0.2	**1.1588**

Unicode Normazation and De-normalization

- Unicode characters may have multiple representations

য় VS য ◌়

- De-normalized output using minimum edit distance alignment

Table: Results of Unicode normalization on the private test set

Type	Levenstein Distance
BanglaBERT-large without normalization	1.130
BanglaBERT-large with normalization	**1.084**

Rule-based Error Detection

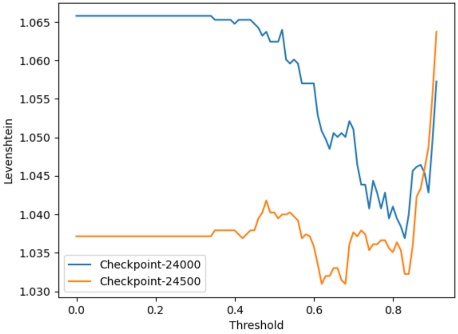
- We deterministically fix
  - Extra spaces before punctuation (space fix)
  - Missing punctuation at the end (end fix)

Model	**Levenstein Distance**
BanglaBERT-base	1.2948
BanglaBERT-base+space fix	1.246
BanglaBERT-base+space fix+end fix	1.212

Rule-based Error Detection: Spelling Errors

- Steps
  - Collect the database of common Bangla Spelling Error from DPCSpell Paper.
  - Remove error in online Bangla dictionary.
  - Remove Bangla Wikipedia title words.
  - Finally, only apply if the word is not a named-entity as per the BNLP NER model.
- Result: Small gain (0.0008) in performance at Phase 1 test set.

Confidence Thresholding



BanglaBERT-large+threshold 0.0	1.1892
BanglaBERT-large+threshold 0.8	**1.1588**

Ensemble

- Type I
  - Union
  - Intersection
- Type II
  - Single-checkpoint
  - Three-checkpoints

Ensemble

Table: Effectiveness of Ensemble I on Private Test Set

Type	**Levenstein Distance**
BanglaBERT-large only	1.2212
BanglaBERT-base+large Union	1.2524
BanglaBERT-base+large Intersection	**1.144**

Table: Effectiveness of Ensemble II on Public Test Set

Model	**Levenstein Distance**
Single-checkpoint	1.0648
Three-checkpoints	**1.0539**

Summary of Our Solution

- Task formalization: Four-class Token Classification
- Models: BanglaBERT-base and BanglaBERT-large
- Ensemble:
  - Union
  - Three-checkpoints
- Loss function: Label Smoothing Cross-Entropy
- Optimizer: AdamW with Linear LR Scheduling
- Preprocessing: Normalization
- Postprocessing:
  - Confidence Thresholding
  - Denormalization
  - Rule-based Fix

Future Works

- To employ self-training with in-domain unlabeled data.
- To combine self-training with feature-based learning to learn a more robust model.
- To use adversarial training strategies.