Surgical Masks

Neural Spell Checker

Our Team

Team

200199T - Akesh Samuditha

200073D - Tharusha Bandaranayake

200294F - Nadil Karunarathna

Supervisors

Dr. Nisansa de Silva

Dr. Surangika Ranathunga

Outline

Introduction

What is Spell Correction ?

- Process of identifying and correcting misspelled words in a text
- Computer based solutions taking the lead to automate the process with better accuracy.

Detection vs Correction

- Two main steps in spell correction
  a. **Detecting** a misspelled word
  b. Suggest **corrections** for the detected word

- A spell correction solution must perform both in detection and correction tasks

Research Problem

- Sinhala is a low-resource, morphologically rich Indo Aryan Language and it has a well defined rule set pertaining to spelling.
- Current spell corrector solutions which primarily use dictionary look-up or n-gram techniques are inherently not context aware.
- There have been publications of neural models for spell correction that are context aware but they lack accuracy [1].

1. Sudesh, P., Dashintha, D., Lakshan, R., & Dias, G. (2022, July). Erroff: A Tool to Identify and Correct Real-word Errors in Sinhala Documents. In 2022 Moratuwa Engineering Research Conference (MERCon) (pp. 1-6). IEEE.

The Gap

01

Comparison of different PLMs architectures

- A comparison between Encoder-only, Decoder-only, and Encoder-Decoder models has not been performed for the task of Spell Correction

02

Improvements to the architecture of PLMs

- Improvements to the architecture of PLMs beyond simple pre-training has not been explored yet for Sinhala spell correction

03

Sinhala Spell Correction Tool

- Existing open-source tools are limited in usability due to web-based only access.
- No functioning end user application based on PLMs for Sinhala spell correction

Objectives

- Perform a comparison between different PLMs
- Improve the best performing PLM further
- Implement an online tool

Literature Review

Rule Based Methods

- Rely on pre-defined linguistic rules to detect and correct spelling errors
- Typically use dictionaries and patterns (Ex: Common typos, Phonetic rules) to find misspelled words.
- Can correct straightforward errors but struggle with context-sensitive corrections or new words not in the dictionary.

Deep Neural Networks (DNN)

- Powerful models capable of learning complex patterns in text data
- Can identify and correct spelling errors by leveraging large amounts of training data.
- Automatically learn language patterns without relying on predefined rules, making them highly adaptable to different languages and contexts

Pre-trained Language Models (PLMs)

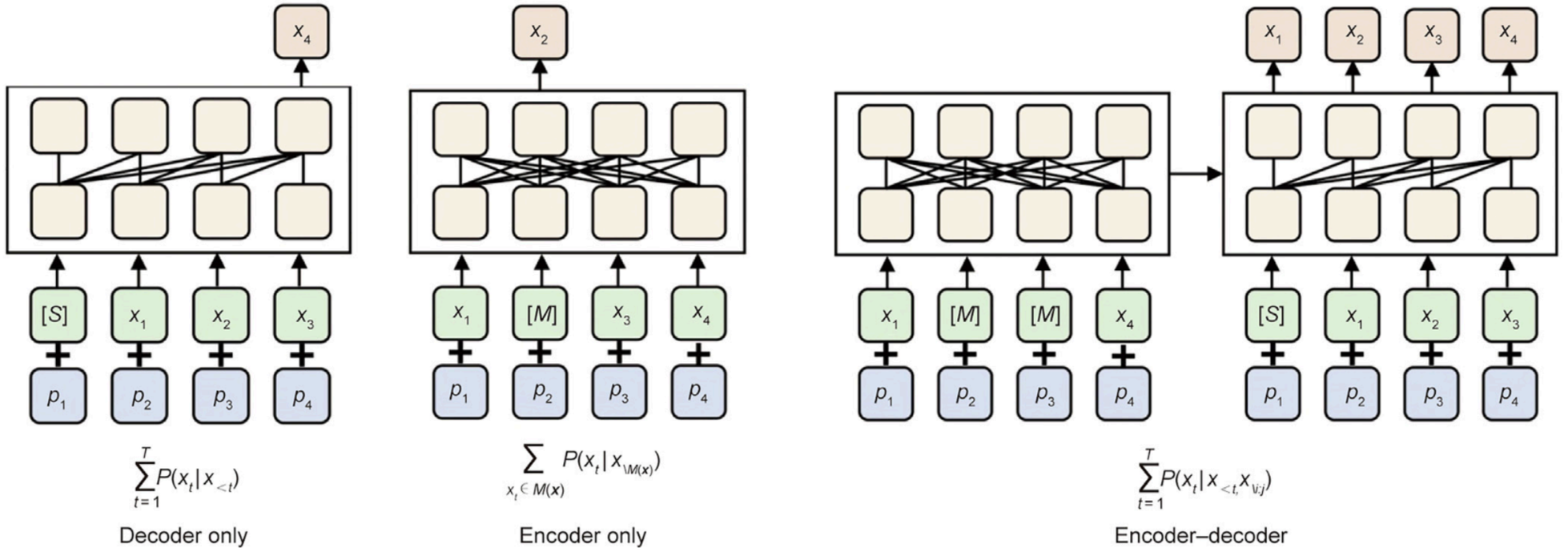
- Models that are trained on large text corpora to learn general representations of language
- Can then be fine-tuned to downstream tasks such as spell correction, summarization, sentiment analysis, translation, etc.
- Ex: mT5, mBART, BERT, RoBERTa, GPT-2

Large Language Models (LLMs)

- Subset of PLMs with billions or trillions of parameters
- Capable of understanding and generating human-like text
- Can perform a wide range of tasks without fine-tuning
- Ex: Llama 3.1, GPT-3, GPT-4, PaLM, Claude

Pre-Training Frameworks [12]

- Encoder only
- Decoder only
- Encoder-Decoder



12. Wang, H., Li, J., Wu, H., Hovy, E., & Sun, Y. (2023). Pre-trained language models and their applications. Engineering, 25, 51-65.

Spell Correction

Rule Based

- Dictionary Lookups [13]
- n-gram Techniques [14]
- Edit - Distance Techniques[15]

ML Based

Neural Network Based Models

DL Based

Combination of Architectures of CNN, LSTM, GRU, transformers

- Bi-LSTM model - Aytan et al. [16]
- LSTM model - Hertel et al. [17]
- RNN model - Salhab et al. [18]

PLM Based

- Encoder only
  - BERT - Liu et al. [19]
- Decoder only
  - GPT - Ramaneedi et al. [21]
- Encoder-Decoder
  - BART - Raju et al. [20]
  - T5 - Zhang et al. [9]

Existing Literature on SC on Other Languages using Neural Models

Paper	Language	Year	Models Used	Architecture
Ghosh et al. [6]	English	2017	CNN + GRU	Encoder-Decoder
Liu et al. [7]	Viatnamese	2019	mBERT	Encoder-Only
HINDIA[8]	Hindi	2020	Bi-RNN	Encoder-Decoder
Erroff [1]	Sinhala	2022	mT5, mBART	Encoder-Decoder
Zhang et al. [9]	Chinese	2023	GPT-3.5 turbo	Decoder-only
Li, Yinghui, et al. [22]	Chinese	2023	Baichuan 2	Decoder Only
Dutta et al. [10]	Persian	2024	mT5, mBART	Encoder-Decoder
ReLM [11]	Chinese	2024	BERT	Encoder Only

Sinhala Spell Correction

A Data-Driven Approach to Checking and Correcting Spelling Errors in Sinhala [23] 2010	**Data-Driven Spell Checking [2]** 2012	**Data-Driven Approach to Sinhala Spell Checking and Correction [3]** 2019	**SinSpell [4]** 2021	**Sinhala spell correction [5]** 2021	**ErrOff [1]** 2022
Permutation generation and n-gram	Permutation generation, n-gram and edit distance	Permutation generation and edit distance	A Rule Based Solution	Neural Model	Neural Model

1. Sudesh, P., Dashintha, D., Lakshan, R., & Dias, G. (2022, July). Erroff: A Tool to Identify and Correct Real-word Errors in Sinhala Documents. In 2022 Moratuwa Engineering Research Conference (MERCon) (pp. 1-6). IEEE.

2. Jayalatharachchi, E., Wasala, A., & Weerasinghe, R. (2012, December). Data-driven spell checking: the synergy of two algorithms for spelling error detection and correction. In International Conference on Advances in ICT for Emerging Regions (ICTer2012) (pp. 7-13). IEEE.

3. Subhagya, L. G. B., Ranathunga, L., Nimasha, W. H. A., Jayawickrama, B. R., & Mahaliyanaarchchi, K. L. (2018, September). Data driven approach to sinhala spellchecker and correction. In 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer) (pp. 01-06). IEEE.

4. Liyanapathirana, U., Gunasinghe, K., & Dias, G. (2021). Sinspell: A comprehensive spelling checker for sinhala. arXiv preprint arXiv:2107.02983.

5. Sonnadara, C., Ranathunga, S., & Jayasena, S. (2021). Sinhala spell correction: A novel benchmark with neural spell correction.

23. Wasala, A., Weerasinghe, R., Pushpananda, R., Liyanage, C., & Jayalatharachchi, E. (2010). A data-driven approach to checking and correcting spelling errors in sinhala. Int. J. Adv. ICT Emerg. Reg, 3(01), 13.

ErrOff Sudesh et al. [1]

- **mBART50** and **mT5-base** models were fine-tuned using a custom created parallel dataset with correct sentences and sentences with synthetic errors

Dataset

- Improved from Sonnadara et el [5]
- Train set - 521283 sentences
- Test Set - 2037 Sentences
- Type of errors introduced
  - Insertion
  - Deletion
  - Substitution
  - Transposition
- Four separate training sets were used
  - **Dataset-1**: Non-word errors only
  - **Dataset-2**: Real-word errors only
  - **Dataset-3**: Real-word errors and non-word errors at 1:1 ratio
  - **Dataset-4**: Real-word errors and non-word errors at 3:1 ratio

Results

- mBART50 outperformed the mT5-base model at the initial comparison
- mBART50 was fine-tuned with all 4 datasets and the Dataset 4 displayed the best result.
- A beam search algorithm was used to improve context dependent suggestions

1. Sudesh, P., Dashintha, D., Lakshan, R., & Dias, G. (2022, July). Erroff: A Tool to Identify and Correct Real-word Errors in Sinhala Documents. In 2022 Moratuwa Engineering Research Conference (MERCon) (pp. 1-6). IEEE.
5. Sonnadara, C., Ranathunga, S., & Jayasena, S. (2021). Sinhala spell correction: A novel benchmark with neural spell correction.

Existing Solutions for SSC

Solution	Detection			Correction		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F0.5**
Erroff [1]	**93.7**	**80.0**	**90.6**	**86.7**	**35.0**	**50.0**
scRNN [5]	37.48	47.32	41.83	33.5	39.76	34.59
SinSpell [4]*	87.52	98.25	92.58	Not Given		

* These values are calculated only considering non-word errors. Because SinSpell cannot identify real-word errors.

1. Sudesh, P., Dashintha, D., Lakshan, R., & Dias, G. (2022, July). Erroff: A Tool to Identify and Correct Real-word Errors in Sinhala Documents. In 2022 Moratuwa Engineering Research Conference (MERCon) (pp. 1-6). IEEE.
4. Liyanapathirana, U., Gunasinghe, K., & Dias, G. (2021). Sinspell: A comprehensive spelling checker for sinhala. arXiv preprint arXiv:2107.02983.
5. Sonnadara, C., Ranathunga, S., & Jayasena, S. (2021). Sinhala spell correction: A novel benchmark with neural spell correction.

Methodology

Models

Evaluated following core architectures for Sinhala Spell Correction task:

- Encoder only
- Decoder only
- Encoder and Decoder

Model	#params	Architecture	# supported languages
mT5	580M	Encoder-Decoder	101
mBART50	680M	Encoder-Decoder	50
XLM-RoBERTa	550M	Encoder only	100
SinBERT	125.9M	Encoder only	1 (Sinhala only)
Llama 3.1	8B	Decoder only	8*

*Does not provide Official Support for Sinhala

Dataset

- Used the same dataset as in ErrOff [1] its self is a generated dataset from correct sentences.
- Removed the Overlapping Sentences from Train Set
- 95% Real word Errors
- Type of Errors
  - Insertion
  - Deletion
  - Substitution
    - na-Na-la (ළ -> ල)
    - retroflexand_palatal_sibilant (ෂ -> ශ)
    - prenasalized (ඳ -> ද)
    - aspirated_and_unaspirated_consonant (ධ -> ද)
    - diacritic ("ෟ" -> "□")
    - other (ඬු -> ඳු)
    - similar_shape (ඝ -> ඟ)

	Before Cleaning	After Cleaning	Validation Set	Test Set
Total Sentences	522885	510706	5282	2037
Unique Sentences	497030	495953	5279	2037
Total Words	7546241	7504701	77538	34420
Unique Words	306028	305933	20835	8698

1. Sudesh, P., Dashintha, D., Lakshan, R., & Dias, G. (2022, July). Erroff: A Tool to Identify and Correct Real-word Errors in Sinhala Documents. In 2022 Moratuwa Engineering Research Conference (MERCon) (pp. 1-6). IEEE.

Established Benchmark

Model	Detection			Correction		
	Precision	Recall	F1	Precision	Recall	F0.5
mBart50	**64.22**	**96.45**	**76.12**	**59.54**	**79.72**	**62.09**
SinBERT	44.43	87.98	57.50	61.38	26.33	53.98
mT5	TBA	TBA	TBA	TBA	TBA	TBA
XLM-R	TBA	TBA	TBA	TBA	TBA	TBA
Erroff [1]	83.27	54.24	65.69	80.07	43.78	68.68
scRNN [5]	37.48	47.32	41.83	33.5	39.76	34.59

1. Sudesh, P., Dashintha, D., Lakshan, R., & Dias, G. (2022, July). Erroff: A Tool to Identify and Correct Real-word Errors in Sinhala Documents. In 2022 Moratuwa Engineering Research Conference (MERCon) (pp. 1-6). IEEE.
5. Sonnadara, C., Ranathunga, S., & Jayasena, S. (2021). Sinhala spell correction: A novel benchmark with neural spell correction.

Results: https://docs.google.com/spreadsheets/d/1-bjAYVEY2IXWEHTE9xXLOX4ZbjyLKPE1Y5h0yI4gbS0

Proposed Enhancements

01 Improving the performance of the Tokenizer

02 Develop a more Robust Loss Function for Training

03 Research on Pre and Post Processing Steps

04 Using self-consistency to explore different reasoning paths

01 Tokenizer

- SentencePiece[24] with Subword Tokenization is commonly used
- Improving the tokenizer to handle Sinhala specific special characters like the Zero Width Joiner
- Fine-tuning tokenizer for Sinhala laguage
- Experiment with Different Tokenization Granularities

```
ක්‍රීඩා විෂය හා අනුබද්ධ
ක්\u200dරීඩා විෂය හා අනුබද්ධ
Token: _ක්, ID: 16607
Token: _රී, ID: 11779
Token: ඩා, ID: 15747
Token: _වි, ID: 5216
Token: ෂය, ID: 69582
Token: _හා, ID: 4487
Token: _අනු, ID: 23812
Token: බ, ID: 4135
Token: ද්ධ, ID: 49576
```

Output from mT5 Tokenizer

```
You are an Expert
Sinhala Spell Corrector

Token: _You, ID: 1662
Token: _are, ID: 418
Token: _an, ID: 461
Token: _Expert, ID: 28235
Token: _S, ID: 320
Token: inhala, ID: 114453
Token: _Spell, ID: 156790
Token: _, ID: 259
Token: Correct, ID: 133759
Token: or, ID: 723
Token: </s>, ID: 1
```

24. Kudo, T. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226.

02 Loss Function

- Develop a Loss Function that penalizes the model for changing a correct word.

03 Pre and Post Processing

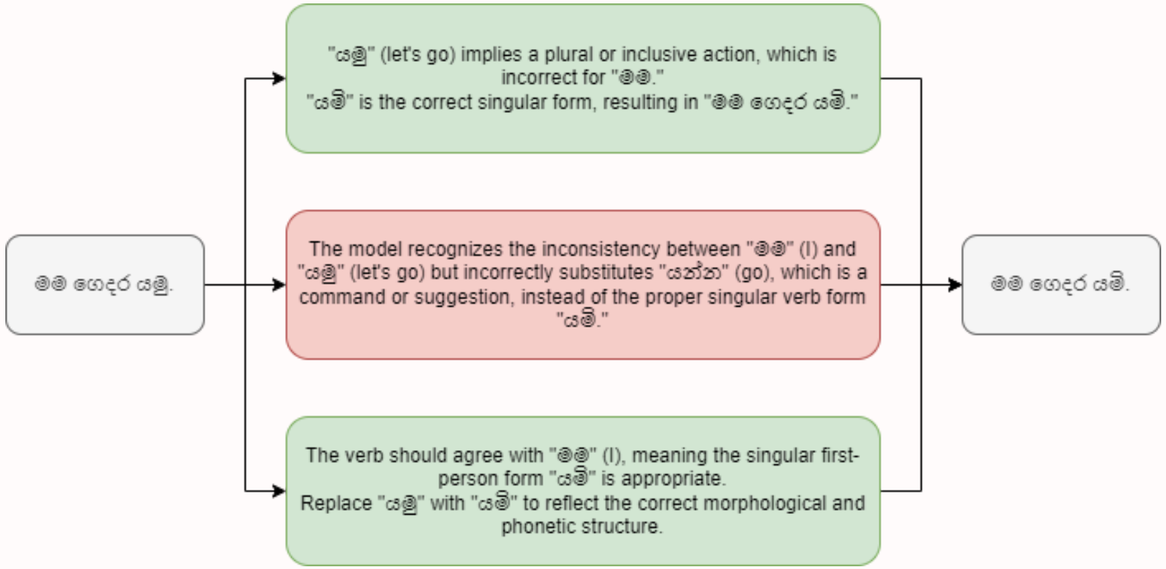
- Use Pre/Post Processing to handle special characters
  - Ex: Zero Width Joiner (\u200d)
- Using Rule-Based Methods for Improving detection of Non-Word Errors [5]

Incorrect use of prenasalized consonants. E.g. writing දඩුවම(/daḍuvama) instead of දඬුවම (/daṅduvama-punishment). There can be instances where this type of errors produce real errors E.g. අද (/ada - today) and අඳ (/aṅda/ - blind)
Writing with incorrect letters that look similar to correct letters. E.g. writing සෘජු (/suju) instead of සෘජු (/ruju-straight)
Incorrect use of diacritics. E.g. writing භූමිය (/bhumiya) instead of භූමිය (/bhūmiya-land)
Merge or split errors. E.g. writing වැඩ සටහන් instead of වැඩසටහන් (/vædasatahan-programs). In this example, the split word pair වැඩ - work and සටහන් - notes carry individual meanings. But it is possible that some merge/split errors produce non-word errors.
Other insertion, deletion and substitution errors. E.g. writing තිපෙඛ instead of තිබේ (/thibē-has)
Encoding errors. E.g. writing ඔවුන්ේ instead of ඔවුන් (/ovun-them)
Combined errors. E.g. writing රඳවාගේන instead of රඳවාගෙන (/raṅdawāgena-detained)

5. Sonnadara, C., Ranathunga, S., & Jayasena, S. (2021). Sinhala spell correction: A novel benchmark with neural spell correction.

04 Self-Consistency

- Using self-consistency[12] to explore different reasoning paths and arriving at the final output can result in diverse answers.
- These candidate answers are then aggregated by marginalizing out the sampling reasoning paths and the most consistent answer among the generated answers is chosen as the final answer.
- This improves the chances of the model reaching the correct answer.



12. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Deliverables

01 Flexible and User Friendly Sinhala Spell Correction Tool

02 A Library Consisting of Fine Tuned models

03 Publications

Technology

Hugging Face

Accelerate

deepspeed

Dataset[1]

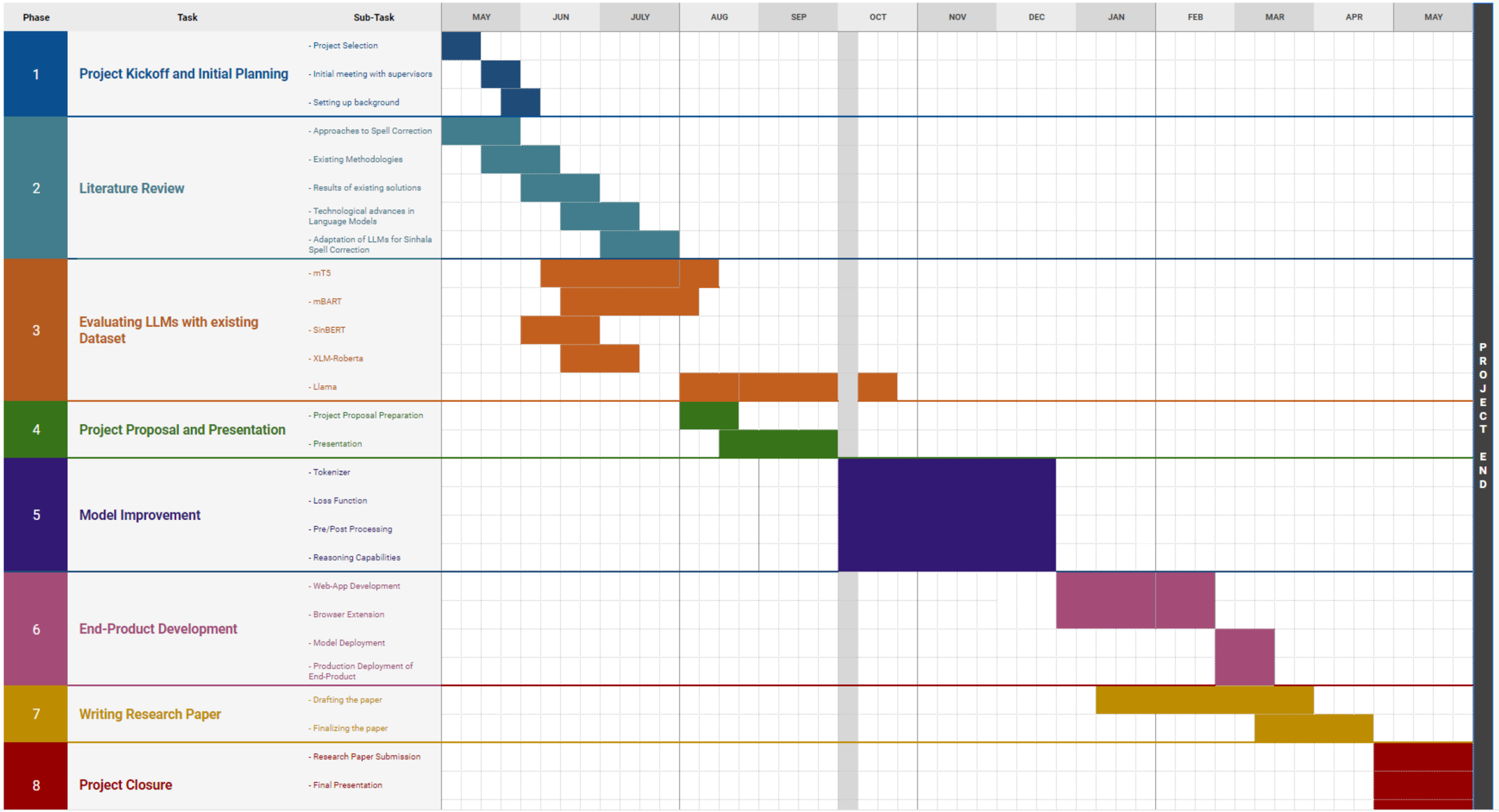
php

kaggle

Google Cloud

1. Sudesh, P., Dashintha, D., Lakshan, R., & Dias, G. (2022, July). Erroff: A Tool to Identify and Correct Real-word Errors in Sinhala Documents. In 2022 Moratuwa Engineering Research Conference (MERCon) (pp. 1-6). IEEE.

Timeline


Phase
Task
Sub-Task
MAY
JUN
JULY
AUG
SEP
OCT
NOV
DEC
JAN
FEB
MAR
APR
MAY
1
Project Kickoff and Initial Planning
- Project Selection
- Initial meeting with supervisors
- Setting up background
2
Literature Review
- Approaches to Spell Correction
- Existing Methodologies
- Results of existing solutions
- Technological advances in Language Models
- Adaptation of LLMs for Sinhala Spell Correction
3
Evaluating LLMs with existing Dataset
- mT5
- mBART
- SinBERT
- XLM-Roberta
- Llama
4
Project Proposal and Presentation
- Project Proposal Preparation
- Presentation
5
Model Improvement
- Tokenizer
- Loss Function
- Pre/Post Processing
- Reasoning Capabilities
6
End-Product Development
- Web-App Development
- Browser Extension
- Model Deployment
- Production Deployment of End-Product
7
Writing Research Paper
- Drafting the paper
- Finalizing the paper
8
Project Closure
- Research Paper Submission
- Final Presentation
PROJECT END

Publication Plans

- Comparison between Encoder only, Decoder only and Encoder-Decoder model performance for Spell Correction

- New improvements introduced by this research

Questions ?

Thank You

References

1. Sudesh, P., Dashintha, D., Lakshan, R., & Dias, G. (2022, July). Erroff: A Tool to Identify and Correct Real-word Errors in Sinhala Documents. In 2022 Moratuwa Engineering Research Conference (MERCon) (pp. 1-6). IEEE.
2. Jayalatharachchi, E., Wasala, A., & Weerasinghe, R. (2012, December). Data-driven spell checking: the synergy of two algorithms for spelling error detection and correction. In International Conference on Advances in ICT for Emerging Regions (ICTer2012) (pp. 7-13). IEEE.
3. Subhagya, L. G. B., Ranathunga, L., Nimasha, W. H. A., Jayawickrama, B. R., & Mahaliyanaarchchi, K. L. (2018, September). Data driven approach to sinhala spellchecker and correction. In 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer) (pp. 01-06). IEEE.
4. Liyanapathirana, U., Gunasinghe, K., & Dias, G. (2021). Sinspell: A comprehensive spelling checker for sinhala. arXiv preprint arXiv:2107.02983.
5. Sonnadara, C., Ranathunga, S., & Jayasena, S. (2021). Sinhala spell correction: A novel benchmark with neural spell correction.
6. Ghosh, S. and Kristensson, P.O., 2017. Neural networks for text correction and completion in keyboard decoding. arXiv preprint arXiv:1709.06429.
7. Liu, L., Wu, H., & Zhao, H. (2024). Chinese Spelling Correction as Rephrasing Language Model. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17), 18662-18670.
8. Singh, S., Singh, S. HINDIA: a deep-learning-based model for spell-checking of Hindi language. Neural Comput & Applic 33, 3825–3840 (2021).
9. Zhang, X., Zhang, X., Yang, C., Yan, H. and Qiu, X., 2023. Does correction remain an problem for large language models?. arXiv preprint arXiv:2308.01776.
10. Dutta, A., Polushin, G., Zhang, X., & Stein, D. (2024, August). Enhancing E-commerce Spelling Correction with Fine-Tuned Transformer Models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 4928-4938).
11. Trung, H.N., Ham, D.T., Huynh, T. and Hoang, K., 2024. A Combination of BERT and Transformer for Vietnamese Spelling Correction. arXiv preprint arXiv:2405.02573.
12. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

References

13. Singh, S., & Singh, S. (2020). Systematic review of spell-checkers for highly inflectional languages. Artificial Intelligence Review, 53(6), 4051-4092.
14. Zamora, E. M., Pollock, J. J., & Zamora, A. (1981). The use of trigram analysis for spelling error detection. Information Processing & Management, 17(6), 305-316.
15. Islam, M. I. K., Meem, R. I., Kasem, F. B. A., Rakshit, A., & Habib, M. T. (2019, May). Bangla spell checking and correction using edit distance. In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) (pp. 1-4). IEEE.
16. Aytan, B., & ŞAKAR, C. O. (2023). Deep learning-based Turkish spelling error detection with a multi-class false positive reduction model. Turkish Journal of Electrical Engineering and Computer Sciences, 31(3), 581-595.
17. Hertel, M. (2019). Neural language models for spelling correction. Methods, 1(2).
18. Salhab, M., & Abu-Khzam, F. (2024). Araspell: A deep learning approach for arabic spelling correction. arXiv preprint arXiv:2405.06981.
19. Liu, L., Wu, H., & Zhao, H. (2024). Chinese Spelling Correction as Rephrasing Language Model. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17), 18662-18670. https://doi.org/10.1609/aaai.v38i17.29829
20. Raju, R., Pati, P. B., Gandheesh, S. A., Sannala, G. S., & Suriya, K. S. (2024). Grammatical versus Spelling Error Correction: An Investigation into the Responsiveness of Transformer-Based Language Models Using BART and MarianMT. Journal of Information & Knowledge Management, 2450037.
21. Ramaneedi, S., & Pati, P. B. (2023, April). Kannada Textual Error Correction Using T5 Model. In 2023 IEEE 8th International Conference for Convergence in Technology (I2CT) (pp. 1-5). IEEE.
22. Li, Y., Huang, H., Ma, S., Jiang, Y., Li, Y., Zhou, F., ... & Zhou, Q. (2023). On the (in) effectiveness of large language models for chinese text correction. arXiv preprint arXiv:2307.09007.
23. Wasala, A., Weerasinghe, R., Pushpananda, R., Liyanage, C., & Jayalatharachchi, E. (2010). A data-driven approach to checking and correcting spelling errors in sinhala. Int. J. Adv. ICT Emerg. Reg, 3(01), 13.
24. Kudo, T. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226.