EC9640 - ARTIFICIAL INTELLIGENCE

ASSIGNMENT 02

HULANGAMUWA S.C. – 2020/E/054

WIJESINGHE L.A.D.P.A. - 2020/E/182

SEMESTER 07

06 JAN 2024

1. Introduction

This project aims to develop a Sinhala spell and grammar checker that detects and corrects spelling and grammatical errors in Sinhala text. It includes a spell checker for automatic error correction and a grammar checker to identify and fix contextual grammatical issues. Three AI approaches such as Rule-based, LLM-based, and Retrieval-Augmented Generation with Fine-tuned LLM(XLM-RoBERTa) models were evaluated for grammar correction. The most effective approach was selected and tested on five paragraphs to measure its accuracy in spelling corrections and grammar suggestions.

2. Spell correction

a. Description of each considered spelling mistakes their correction approaches with example

Description:

The Sinhala Grammar Checker addresses spelling mistakes like typographical errors, phonetic similarities, and affix variations. Typographical errors are resolved using dictionary matching and edit distance techniques, while phonetic similarities are corrected through advanced phonetic mapping. Prefix and suffix errors are handled with affix rules and stemming to identify root words. The checker uses algorithms like Levenshtein distance, phonetic key generation, and dictionary validation to provide accurate corrections.

Example:

```
--- Spell Check for: සුද්ධාවත්
Spelling Errors: {'සී\u200dරද්ධාවත්': [('ශ්‍රද්ධාවෙන්', 76.16323529411764), ('ශ්‍රද්ධාවත්', 76.16323529411764), ('ස්පර්ධාවක්', 7
5.71323529411765), ('ස්පර්ධාවෙන්', 72.76944444444445), ('අශ්‍රද්ධාවත්', 70.41944444444444)]}
Auto-corrected Text: ශ්‍රද්ධාවෙන්

--- Spell Check for: ගෘහෂ්ථ
Spelling Errors: {'ගෘහෂ්ථ': [('ගෘහස්ථ', 82.38333333333333)]}
Auto-corrected Text: ගෘහස්ථ

--- Spell Check for: පනිවිඩය
Spelling Errors: {'පනිවිඩය': [('පණිවිඩය', 79.90833333333333), ('පිවිධ', 71.50681818181819), ('නිවිති', 71.075)]}
Auto-corrected Text: පණිවිඩය

--- Spell Check for: වානිජ්
Spelling Errors: {'වානිජ්': [('වාණිජ්', 78.025), ('වානේ', 71.25)]}
Auto-corrected Text: වාණිජ්

--- Spell Check for: මයිණා
Spelling Errors: {'මයිණා': [('මයිකා', 77.67857142857143), ('මයිනා', 77.67857142857143), ('මිණුම්', 70.42500000000001),
('මිණි', 70.42500000000001)]}
Auto-corrected Text: මයිකා
```

3. Grammar suggestions

a. Description of each considered grammar mistakes (2 different), their suggestions Description:

The Sinhala Grammar Checker code addresses two subject-verb agreement mistakes:

- For "මම" (I), the verb should end with "මි," as in correcting "මම බත් කමු" (I eat rice) to "මම බත් කමි."
- For "අපි" (We), the verb should end with "මු," as in changing "අපි පාඩම් කරමි" (We study) to "අපි පාඩම් කරමු."

The code uses a rule-based approach with predefined suffix mappings, tokenization, and POS tagging to identify and correct these errors, ensuring proper subject-verb agreement.

Example:

```
Processing Input Sentences...

Processing: මම ගදර යයී
Text after Spell Checking: මම ගෙදර යයී
Running Grammar Checker...

Processing: අපි කඩේට යයී
Text after Spell Checking: අපි කඩේ යයී
Running Grammar Checker...

Processing: අපි අදරය කරයී
Text after Spell Checking: අපි ආදරය කරයී
Running Grammar Checker...

Processing: මම කෑම ගනිමු
Text after Spell Checking: මම කෑම ගනිමු
Running Grammar Checker...

Processing: අපි රට යයී
Text after Spell Checking: අපි රට යයී
Running Grammar Checker...

Processing: අපි කලි යම
Text after Spell Checking: අපි කලින් යම
Running Grammar Checker...

Final Output:
Sentence 1: Grammar error: Verb 'යයී' (base: ය, affix: යී) should end with 'මී' when subject is 'මම'
Suggested correction: මම ගෙදර යමී
Sentence 2: Grammar error: Verb 'යයී' (base: ය, affix: යී) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි කඩේ යමු
Sentence 3: Grammar error: Verb 'කරයී' (base: කර, affix: යී) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි ආදරය කරමු
Sentence 4: Grammar error: Verb 'ගනිමු' (base: ගනි, affix: මු) should end with 'මී' when subject is 'මම'
Suggested correction: මම කෑම ගනිමී
Sentence 5: Grammar error: Verb 'යයී' (base: ය, affix: යී) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි රට යමු
Sentence 6: Grammar error: Verb 'යම' (base: ය, affix: ම) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි කලින් යමු
```

b. Existing 3 AI approaches details with example screenshots

1. Rule-Based AI Approach:

The Sinhala spell and grammar checker uses rule-based AI to correct spelling and grammar errors in Sinhala text. It leverages a preprocessed dictionary, phonetic mapping, stemming, and suffix matching to identify spelling mistakes and applies similarity metrics for corrections. The grammar checker ensures proper subject-verb agreement and word order using a POS tagger, offering accurate corrections for spelling and grammar.

```python
# Input Sentences
input_sentences = [
    "අපි කඩේ යයී.අපි උදේට කඩෙන කමු.",
    "මම අද ගදර යයී.අපි හෙට ගමට යම.",
    "අපි හෙට නුවර යන්නෙමී.මම දැන් වැවට යම.",
    "මම ඔයාට අදරය කරයී.මම ඔයාගේ ගෙදර යයී.",
    "අපි උත්සාහයෙන් වැඩ කරමී.අපි කොහොමහෙරි දිනයී."
]
```

```
Processing Input paragaphs...

Processing: අපි කඩේ යයී.අපි උදේට කඩෙන කමු.
Text after Spell Checking: අපි කඩේ යයී.අපි උදේට කඩෙන් කමු.
Running Grammar Checker...

Processing: මම අද ගදර යයී.අපි හෙට ගමට යම.
Text after Spell Checking: මම අද ගෙදර යයී.අපි හෙට ගමට යම.
Running Grammar Checker...

Processing: අපි හෙට නුවර යන්නෙමී.මම දැන් වැවට යම.
Text after Spell Checking: අපි හෙට නුවර යන්නෙමී.මම දැන් වැවට යම.
Running Grammar Checker...

Processing: මම ඔයාට අදරය කරයී.මම ඔයාගේ ගෙදර යයී.
Text after Spell Checking: මම ඔයාට ආදරය කරයී.මම ඔයාගේ ගෙදර යයී.
Running Grammar Checker...

Processing: අපි උත්සාහයෙන් වැඩ කරමී.අපි කොහොමහරි දිනයී.
Text after Spell Checking: අපි උත්සාහයෙන් වැඩ කරමී.අපි කොහොමහරි දිනය
Running Grammar Checker...
```

```
Final Output: Grammar Check Results

Results for paragaph 1:
Grammar Check Result: Grammar error: Verb 'යයි' (base: ය, affix: යි) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි කඩේ යමු
Grammar Check Result: The sentence is grammatically correct.

Results for paragaph 2:
Grammar Check Result: Grammar error: Verb 'යයි' (base: ය, affix: යි) should end with 'මි' when subject is 'මම'
Suggested correction: මම අද ගෙදර යමි
Grammar Check Result: Grammar error: Verb 'යම' (base: ය, affix: ම) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි හෙට ගමට යමු

Results for paragaph 3:
Grammar Check Result: Grammar error: Verb 'යන්නෙමි' (base: යන්නෙ, affix: මි) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි හෙට නුවර යන්නෙමු
Grammar Check Result: Grammar error: Verb 'යම' (base: ය, affix: ම) should end with 'මි' when subject is 'මම'
Suggested correction: මම දැන් වැවට යමි

Results for paragaph 4:
Grammar Check Result: Grammar error: Verb 'කරයි' (base: කර, affix: යි) should end with 'මි' when subject is 'මම'
Suggested correction: මම ඔයාට ආදරය කරමි
Grammar Check Result: Grammar error: Verb 'යයි' (base: ය, affix: යි) should end with 'මි' when subject is 'මම'
Suggested correction: මම ඔයාගේ ගෙදර යමි

Results for paragaph 5:
Grammar Check Result: Grammar error: Verb 'කරමි' (base: කර, affix: මි) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි උත්සාහයෙන් වැඩ කරමු
Grammar Check Result: Grammar error: Verb 'දිනය' (base: දින, affix: ය) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි සොහොමහරි දිනමු
```

2. Retrieval-Augmented Generation (RAG) approach:

The Sinhala grammar and spell checker uses a RAG approach with LangChain and Google's Gemini model to detect and correct errors in Sinhala sentences. It processes grammar rules from a JSON file, splits them into chunks for embedding, and stores them in a FAISS vector store for efficient retrieval. Upon receiving an input sentence, the system retrieves relevant grammar rules, constructs a prompt, and uses Gemini to generate corrections and explanations. The combination of LangChain, FAISS, and Gemini allows for accurate and fast error detection and correction.

```
Response:
Okay, let's break down the errors and correct the sentences.

**Sentence 1: අපි හෙට නුවර යන්නෙමි.**

*   **Error:** The verb ending "මි" is incorrect for the subject "අපි" (we). According to the provided context, "අපි" in the present tense should use endings "මු" or "මූ".
*   **Corrected Sentence:** අපි හෙට නුවර යමු. (We will go to Kandy tomorrow.)
*   **Explanation:** The correct verb ending for "අපි" in the present tense is "මු". The sentence is now grammatically correct.

**Sentence 2: මම දැන් වැවට යම.**

*   **Error:** The verb ending "ම" is incorrect for the subject "මම" (I). According to the provided context, "මම" in the present tense should use endings "මි" or "ෙමි".
*   **Corrected Sentence:** මම දැන් වැවට යමි. (I am going to the lake now.)
*   **Explanation:** The correct verb ending for "මම" in the present tense is "මි". The sentence is now grammatically correct.

**Summary of Corrections and Errors:**

*   **Original Sentence 1:** අපි හෙට නුවර යන්නෙමි.
    *   **Error:** Incorrect verb ending "මි" for subject "අපි".
    *   **Corrected Sentence 1:** අපි හෙට නුවර යමු.
*   **Original Sentence 2:** මම දැන් වැවට යම.
    *   **Error:** Incorrect verb ending "ම" for subject "මම".
    *   **Corrected Sentence 2:** මම දැන් වැවට යමි.

The provided context helped us identify the correct verb endings based on the subject and tense. The dative case marker "ට" was used correctly in the second sentence, so no changes we
```

Response:
Corrected Sentences:

*   මම අද ගෙදර යමි.
*   අපි හෙට ගමට යමු.

Error Explanations:

Sentence 1: "මම අද ගදර යයි."

*   **Error:** Subject-verb agreement. The subject "මම" (I) is first person singular. The verb ending "යි" is used for third person singular (he/she/it) in the present tense.
*   **Correction:** The correct verb ending for first person singular present tense is "මි". Therefore, "යයි" should be corrected to "යමි".

Sentence 2: "අපි හෙට ගමට යම."

*   **Error:** Subject-verb agreement. The subject "අපි" (we) is first person plural. The verb ending "ම" is not the correct ending for first person plural present tense.
*   **Correction:** The correct verb ending for first person plural present tense is "මු". Therefore, "යම" should be corrected to "යමු".

Summary of Errors:

Both sentences had errors related to subject-verb agreement. The verb endings did not match the person and number of the subject. The first sentence used a third-person singular ending

Response:
Corrected Sentences:

*   අපි කඩේ **යමු**.
*   අපි උදේට කඩේ**න්** කමු.

Explanation of Errors:

1. **Subject-Verb Agreement Error in "අපි කඩේ යයි":**
    *   The subject "අපි" (we) is first-person plural.
    *   The verb "යයි" (goes) uses the third-person singular/plural ending "යි".
    *   According to the provided context, first-person plural requires the "මු" ending.
    *   Therefore, the correct verb ending for "අපි" is "යමු".
    *   The error type is "subject_verb_agreement" and the pattern is "subject + incorrect_verb_ending".

2. **Case Marking Error in "කඩෙන කමු":**
    *   The sentence "අපි උදේට කඩෙන කමු" means "We eat at the shop in the morning".
    *   The word "කඩේ" (shop) is used as the location where the action of eating takes place.
    *   The provided context indicates that the dative case marker "ට" is used for indirect object marking. However, the location is not an indirect object.
    *   The correct case marker for location is the ablative case marker "න්".
    *   Therefore, the correct form is "කඩේන්" (from the shop).
    *   The error pattern is using "න" instead of "න්".

Response:
Corrected Sentences:

*   අපි උන්සාහයෙන් වැඩ කරමු.
*   අපි කොහොමහරි දිනමු.

Explanation of Errors:

1. **"අපි උන්සාහයෙන් වැඩ කරමි"**
    *   **Error:** The verb ending "මි" is used with the subject "මම" (I) in the present tense. The subject here is "අපි" (we).
    *   **Correction:** The correct verb ending for "අපි" in the present tense is "මු" or "මූ". Therefore, "කරමි" should be "කරමු".

2. **"අපි කොහොමහරි දිනයි"**
    *   **Error:** The verb ending "යි" is not used with the subject "අපි" (we) in the present tense.
    *   **Correction:** The correct verb ending for "අපි" in the present tense is "මු" or "මූ". Therefore, "දිනයි" should be "දිනමු".

In summary:

The errors in the original sentences were due to incorrect verb endings being used with the subject "අපි". The correct verb endings for "අපි" in the present tense are "මු" or "මූ".

Response:
Corrected Sentences:

*   මම ඔයාට ආදරෙයි. (Mama oyāṭa ādarei.)
*   මම ඔයාගේ ගෙදර යමි. (Mama oyāgē gedara yami.)

Error Explanations:

1. **"මම ඔයාට අදරය කරයි" (Mama oyāṭa adaraya karayi) - Subject-Verb Agreement Error:**
    *   **Error:** The verb ending "කරයි" (karayi) is used for the third-person singular subject ("ඔහු/ඇය" - he/she). However, the subject of the sentence is "මම" (mama), which is
    *   **Correction:** The correct verb ending for the first-person singular in the present tense is "මි" (mi) or "යි" (yi). The verb "ආදරය කරනවා" (adaraya karanawa) is often short
    *   **Explanation:** The error falls under the "subject_verb_agreement" error type, specifically the pattern "subject + incorrect_verb_ending". The provided context also indicates

2. **"මම ඔයාගේ ගෙදර යයි" (Mama oyāgē gedara yayi) - Subject-Verb Agreement Error:**
    *   **Error:** Similar to the first sentence, the verb ending "යයි" (yayi) is used for the third-person singular subject. The subject is "මම" (mama), which requires a first-persor
    *   **Correction:** The correct verb ending for the first-person singular in the present tense is "මි" (mi). Therefore, the correct form is "මම ඔයාගේ ගෙදර යමි" (Mama oyāgē geda
    *   **Explanation:** This error also falls under the "subject_verb_agreement" error type, specifically the pattern "subject + incorrect_verb_ending". The provided context also indi

In summary:

Both sentences had subject-verb agreement errors. The verb endings were incorrectly conjugated for the third-person singular instead of the first-person singular. The corrected sentenc

3. Fine-tuned LLM approach with the XLM-RoBERTa model:

The Sinhala grammar and spell checker uses a fine-tuned XLM-RoBERTa model for classifying sentences as grammatically correct or incorrect. It processes sentences through tokenization, fine-tunes the model with the AdamW optimizer, and evaluates grammar issues like subject-verb agreement and sentence structure. The trained model is optimized and ready for deployment to detect and suggest corrections for errors in Sinhala sentences.

```
Sentence: අපි කඩේ යයි
Status: Incorrect
Suggestion: Grammar error: Verb 'යයි' (base: ය, affix: යි) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි කඩේ යමු

Sentence: අපි උදේට කඩේන කමු
Status: Incorrect
Suggestion: The sentence is grammatically correct.

Sentence: මම අද ගදර යයි
Status: Incorrect
Suggestion: Grammar error: Verb 'යයි' (base: ය, affix: යි) should end with 'මි' when subject is 'මම'
Suggested correction: මම අද ගදර යමි

Sentence: අපි හෙට ගමට යම
Status: Incorrect
Suggestion: Grammar error: Verb 'යම' (base: ය, affix: ම) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි හෙට ගමට යමු

Sentence: අපි හෙට තුවර යන්නෙමි
Status: Incorrect
Suggestion: Grammar error: Verb 'යන්නෙමි' (base: යන්නෙ, affix: මි) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි හෙට තුවර යන්නෙමු

Sentence: මම දැන් වැවට යම
Status: Incorrect
Suggestion: Grammar error: Verb 'යම' (base: ය, affix: ම) should end with 'මි' when subject is 'මම'
Suggested correction: මම දැන් වැවට යමි
```



```
Sentence: මම ඔයාට අදරය කරයි
Status: Incorrect
Suggestion: Grammar error: Verb 'කරයි' (base: කර, affix: යි) should end with 'මි' when subject is 'මම'
Suggested correction: මම ඔයාට අදරය කරමි

Sentence: මම ඔයාගේ ගෙදර යයි
Status: Incorrect
Suggestion: Grammar error: Verb 'යයි' (base: ය, affix: යි) should end with 'මි' when subject is 'මම'
Suggested correction: මම ඔයාගේ ගෙදර යමි

Sentence: අපි උත්සාහයෙන් වැඩ කරම්
Status: Incorrect
Suggestion: Grammar error: Verb 'කරම්' (base: කර, affix: ම්) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි උත්සාහයෙන් වැඩ කරමු

Sentence: අපි කොහොමහරි දිනයි
Status: Incorrect
Suggestion: Grammar error: Verb 'දිනයි' (base: දින, affix: යි) should end with 'මු' when subject is 'අපි'
Suggested correction: අපි කොහොමහරි දිනමු
```

c. Selected approach and why

Selected approach: Rule based approach

The Rule-Based AI Approach was chosen due to its high accuracy of 90%, outperforming the LLM model at 70% and the RAG model at 60%. By using predefined rules, it ensures precise handling of Sinhala grammar and sentence structure, making it ideal for tasks requiring accuracy. The limited availability of large Sinhala datasets negatively impacted the performance of the LLM and RAG models, restricting their ability to generalize and learn diverse language patterns.

Although the LLM and RAG models offer greater flexibility, they struggled with issues like subject-verb agreement and sentence structure errors. For instance, the LLM model failed to produce correct sentences such as "අපි කඩේට යමු" and "මම අද ගෙදර යමි." The lack of sufficient training data exacerbated these problems, leading to lower reliability compared to the Rule-Based model. As a result, the Rule-Based approach emerged as the more dependable choice for accurate Sinhala language processing.

4. Conclusion:

In conclusion, the Rule-Based Model achieved the highest accuracy of 90%, outperforming both the LLM and RAG models, which scored 70% and 60%, respectively. The Rule-Based Model's precision is due to its reliance on predefined rules, ensuring consistent grammar and sentence structure. In contrast, while the LLM and RAG models offer flexibility, they struggled with issues such as incorrect subject-verb agreements and inconsistent sentence structures, leading to errors in output. These results highlight the trade-off between flexibility and precision, suggesting that rule-based systems are ideal for tasks requiring strict language rules, while LLM and RAG models can be useful for more dynamic language processing with further refinement.

```
correct_outputs = [
    "අපි කඩේට යමු",
    "අපි උදේට කඩෙන් කමු",
    "මම අද ගෙදර යමි",
    "අපි හෙට ගමට යමු",
    "අපි හෙට නුවර යන්නෙමු",
    "මම දැන් වැවට යමි",
    "මම ඔයාට ආදරය කරමි",
    "මම ඔයාගේ ගෙදර යමි",
    "අපි උත්සාහයෙන් වැඩ කරමු",
    "අපි කොහොමහරි දිනමු"
]
```

```
Accuracy Scores:
LLM Model: 70.00%
RAG Model: 60.00%
Rule-based Model: 90.00%

Errors in LLM Model:
Line 1:
Expected: අපි කඩේට යමු
Got     : අපි කඩේ යමු

Line 2:
Expected: අපි උදේට කඩෙන් කමු
Got     : අපි උදේට කඩෙන කමු

Line 3:
Expected: මම අද ගෙදර යමි
Got     : මම අද ගදර යමි
```

```
Errors in RAG Model:
Line 1:
Expected: අපි කඩේට යමු
Got     : අපි කඩේ යමු

Line 2:
Expected: අපි උදේට කඩෙන් කමු
Got     : අපි උදේට කඩෙන් කමු

Line 5:
Expected: අපි හෙට නුවර යන්නෙමු
Got     : අපි හෙට නුවර යමු

Line 7:
Expected: මම ඔයාට ආදරය කරමි
Got     : මම ඔයාට අදරය කරනවා


Errors in Rule-based Model:
Line 1:
Expected: අපි කඩේට යමු
Got     : අපි කඩේ යමු
```

GitHub link: https://github.com/SalindaHulangamuwa/Spelling-corrector-and-grammar-checker-for-Sinhala