Automatic Rule Extraction for Detecting and Correcting Burmese Spelling Errors

Ei Phyu Phyu Mon[†] **Ye Kyaw Thu**[‡] **Thida San**[†] **Zun Hlaing Moe**[†] **Hnin Aye Thant**[†]
[†]University of Technology (Yatanarpon Cyber City), PyinOoLwin, Myanmar
[‡]Language and Semantic Technology Research Team (LST),
National Electronics and Computer Technology Center (NECTEC), Pathum Thani, Thailand
{ephyumon, yktnlp, thidako22, zunhlaing, hninayethant}@gmail.com

Abstract

Spellcheckers have become important due to the increase in text-based communication at work and in society on social media. Detecting and correcting misspelled words in a written text is very important in many natural language processing tasks. Due to substantial ambiguity in the structure of human languages, especially Burmese, there is still no complete and practical spelling error detection and correction system that helps us avoid common mistakes. In this paper, we propose a rule-based spelling checker that can automatically extract the spelling correction rules from our manually constructed correction corpus and apply these extracted rules to correct the spelling errors. We analyzed and classified the nature of Burmese spelling errors into eleven categories, and also investigated the performance of each error type. The experimental results showed that the proposed approach is capable of extracting 38,124 spelling correction rules and applying these rules to correct the spelling errors of eleven different error types. By studying the evaluation results, although the phonetic "pho" and typographic "typo" error types are not well-handled, the other nine remaining error types are well-corrected with relevant outcomes by this automatic rule extraction approach.

Keywords: Spelling Checker, N-gram, Syllable Segmentation, Regular Expression (RE), Spelling error-correction corpus, Wdiff

1 Introduction

In this modern era, Natural Language Processing (NLP) is a collective term referring to the automatic computational processing of human languages. The Myanmar Language (Burmese) is similar to other Asian Languages including Indian, Chinese, Japanese, and Thailand. The number of social media users in Myanmar was equivalent to 53.1% of the total population in January 2021. In this technological age of computers and social media, the proper spelling is important for efficient communication between people. Spelling is a complex cognitive activity in which many interrelated skills are involved. A spell checker is a widely-used tool that aims to help in detecting and correcting various writing errors. It plays an important role in improving document quality by identifying misspelled words in the document. We analyzed and classified the Burmese spelling errors that are commonly found on social media into eleven categories of errors. The details of these eleven error types are discussed in Section 4.1. According to our study, the most frequent errors are the phonetic errors. The second most frequent errors are the typographic errors, which are caused by insertion, omission, substitution, and transposition. Consonant errors, the combination of phonetic and typographic errors, and sequence errors are the third, fourth, and fifth most frequent errors. Dialect, encoding, sensitive, stacked, short-form, and slang word usages all contribute to errors. In this proposed paper, we emphasize the syllable-level Burmese text with a rule-based spelling checker. No work has been found yet on the automatic rule extraction for Burmese spelling checking. Thus, most of the spellings for Burmese are going extinct. These spelling errors decrease the quality of the Myanmar language. The government recognized and officially announced Unicode as a standard encoding method in computer-based works. However, some people are still using de facto standard encoding named Zawgyi. Hence, both Unicode and Zawgyi encodings are mainly used in Myanmar. Burmese spelling errors can also occur because there is no standard keyboard layout yet. Currently, spell checkers for the English language are well established. Burmese does not have a publicly available spell checker.

Consequently, the spelling checker is a major issue and challenge not only for all computerized applications in Burmese but also for the essential parts of the field of NLP.

Our research contribution is to explore the high-performance spelling correction on the automatic rule extraction approach. One more contribution is we are developing a parallel corpus of spelling error and correction pairs, and we used the current version of this corpus for our experiment. The rest of this paper is organized as follows: related works are discussed in Section 2, the nature of the Myanmar Language is described in Section 3. Section 4 presents the parallel corpus preparation for this proposed spelling checker with examples. Section 5 explains the detail of experimental methodology. Section 6 discuss the evaluation results and Section 7 analyzes the errors of this proposed spelling checker approach. Finally, the paper is concluded in Section 8.

2 Related Work

In recent years, not much new work has been done on spelling checkers, especially for Burmese. Nwe Zin Oo et al. [1] proposed a Myanmar Language Spell Checker by using Levenshtein Distance Algorithm, Dynamic Threshold Algorithm, and Transformation Algorithm. This system used Zawgyi Myanmar font, and was implemented using Java Language and MySQL server to check the spelling of Myanmar words consulting with Animals and Plants, and added correct Myanmar words to the dictionary. Each Myanmar input word was compared against a dictionary of correctly spelled Myanmar words. It was just only for spelling checking with possible suggestions and correcting the erroneous Myanmar words using Levenshtein Distance algorithm.

The authors [2] introduced KidSpell: a phonetic, rule-based spellchecker that corrects spelling errors generated by children. Experiments based on the essay and online search environments demonstrated that KidSpell outperforms well-known, general-purpose counterparts considered for analysis when applied to detect and correct child misspellings. Part of their contributions are new, freely available datasets of child spelling errors. To the best of their knowledge, they are the first of their kind and comprise of spelling errors in various environments from children in different contexts-hand-written essays (grade K-8) and web search environments (age 5-14 years). To assess the effectiveness of KidSpell, they compared the model's performance against several popular, mainstream spellcheckers in several offline experiments using existing and novel datasets. The results of these experiments show that KidSpell outperforms existing spellcheckers, as it accurately prioritizes relevant spelling corrections when handling misspellings generated by children in both the essay writing and online search tasks. Their experiments also showcased that the performance of the proposed model was comparable to existing counterparts when correcting adult spelling mistakes, enabling the use of KidSpell as a general spellchecker for a diverse audience.

Many researchers have been worked for spell checkers of Asian Languages but Burmese spell checker research has still in its vulnerability. Novan Zukarnain et al. [3] analyzed several studies conducted in various types of languages other than English, like Africa, Arabia, China, Indonesia, India, Japan, Malaysia, and Thailand. They investigated the problem, understanding, and solution of spelling checker technology in so many languages. They surveyed each language that has a different spelling check method, and observed that the Damerau-Levenshtein algorithm method is most often used for spelling checkers. Neha Gupta et al. [4] surveyed different types of error in text, techniques for detection and correction, and available spell checkers for Indian languages.

Aye Myat Mon et al. [5] studied the details on Myanmar language to identify the problem area of Myanmar spell errors. The author implemented a Myanmar spell checker system which can handle typographic errors, sequence errors, phonetic errors, and context errors by applying Myanmar text corpus. A compound misused word detection algorithm was proposed for phonetic errors checking, Bayesian classifier was applied for context errors checking, and Levenshtein distance algorithm was applied for generating suggestion list. This system emphasized Myanmar sentences that follow Myanmar grammar rules, and Pali words were not handled.

Zar Zar Hlaing et al. [6] proposed a Myanmar homonym disambiguation system. This system detected homonym errors or ambiguous homonyms, and then resolved these errors by using a corpus-based N-gram model [7]. Stacked consonant homonyms could not

be resolved in this system.

3 The Myanmar Language (Burmese)

Burmese is a resource-scarce language and the official language of the Republic of the Union of Myanmar. It is spoken by 36 million people as a first language and as a second language by another 10 million people, particularly ethnic minorities in Burma and those in neighboring countries. In addition, it is a syllabic writing system, and its script is written horizontally from left to right. Besides, it does not have a delimiter between syllables and words. However, the informal writing form often includes space after each word and a free word order. The Myanmar words are collated based on syllables. A Myanmar syllable has a base character, and may also have (or not) a pre-base (prefix) character, a post-base (suffix) character, an above-base character, and a below-base character. Each syllable boundary should begin with a base consonant. Its basic set of symbols consists of 35 consonants, 4 medials, 12 vowels, other symbols, and special characters. Symbols for vowels may be written before, above, below, or to the right of the letter representing an initial consonant. Some writers are writing with their pronunciation and carelessness about spelling errors. Burmese reflects the difference between spoken and written text, as spelling is often not an accurate reflection of pronunciation. Thus, there are serious discrepancies between orthography and pronunciation. In the Myanmar Language feature, the format of the character sequence [consonant (ဗျည်း), medial (ဗျည်းတွဲ), vowel (သရ)] is very important for writing Burmese words. Table 1 shows the groups of characters according to their sound letters, and they are arranged in the traditional order. Table 2 depicts 12 vowels, 4 medials, and independent vowels.

4 Corpus Preparation

The corpus is an essential component of NLP applications. Currently, there is no parallel corpus for the Burmese spelling checker. To outperform the spell checker, we created bigram syllable error and correction pairs data for eleven error types as our parallel corpus. The statistics of eleven error types for the closed-test and open-test data are shown in Table 3.

Table 1: Burmese Scripts

Consonants				
က /ka/	ခ /kha/	ဂ /ga/	ဃ /gha/	င /nga/
စ /ca/	ဆ /cha/	ဇ /ja/	ဈ /jha/	ဉ /nya/ ည /nnya/
ဋ /tta/	ဌ /ttha/	ဍ /dda/	ဎ /ddha/	ဏ /nna/
တ /ta/	ထ /tha/	ဒ /da/	ဓ /dha/	န /na/
ပ /pa/	ဖ /pha/	ဗ /ba/	ဘ /bha/	မ /ma/
ယ /ya/	ရ /ra/	လ /la/	ဝ /wa/	သ /sa/ ဿ /great sa/
	ဟ /ha/	ဠ /lla/	အ /a/	

Table 2: Vowels, Medials, and Independent vowels for Burmese

Vowels	အ /a/	အိ /i/	အီ /ii/	အု /u/
	အူ /uu/	အေ /e/	အဲ /ai/	အော့ /aa/
	အော် /au/	အံ /an/	အား /a:/	အက် /ae/
Medials	ျ /ya/	ြ /ra/	ွ /o/	ှ /ha/
Independent vowels	ဣ /i/	ဤ /ii/	ဥ /ou/	ဦ /uu/
	ဧ /a/	ဩ /o/	ဪ /au/	

4.1 Burmese Spelling Error Types

In order to propose an efficient spelling checker, we significantly need to study and categorize common error patterns. By studying the spelling errors involved in our Burmese error-correction parallel corpus, we defined eleven error types for our proposed spell checker. These eleven error categories are as follows:

1. *Consonant error (con)*

   These errors occur while misusing consonants, vowels, and independent vowels. e.g., ဥသြ → ဥသြ (“Asian Koel” in English) where the use of “ဥ” instead of

“ဥ”, and the use of writing with “သ” and “ြ” instead of “ြသ”.

2. *Dialect error (dialect)*

   These errors occur due to a variety of languages (specifically, often a spoken variety) that are characteristic of a particular area, community, or group, often with relatively minor differences in vocabulary, style, spelling, and pronunciation.
   e.g., အကျည်းတန်လွန်း → အကျဉ်းတန်လွန်း (“Too shameful” in English) where the use of writing with “ကျဉ်း” instead of “ကျည်း”.

3. *Encoding error (encoding)*

   It is a mistake that happens during the process of encoding data.
   e.g., အိမ် → အိမ္ (“Home” in English) where the use of writing with “အိမ္” instead of “အိမ်”.

4. *Phonetic error (pho)*

   This error has been made by the lack of knowledge of the writer. The pronunciation of a misspelled word is the same or similar to the pronunciation of the intended correct word. We defined this error type as a phonological or pronunciation error.
   e.g., ဒါပေမဲ့ → ဒါပေမယ့် (“But” in English) where the use of writing with “မယ့်” instead of “မဲ့”.

5. *The combination of phonetic and typographic error (pho-typo)*

   These are errors that occur when the correct spelling of the word is not known, and the word is mistyped by mistake.
   e.g., သူငယ်ချင်း → သူငယ်ဂျင်း (“Friend” in English) where the use of writing with “ဂျင်း” instead of “ချင်း”.

6. *Sensitive word error (sensitive)*

   These errors occur when the correct spelling of the word reflects the sensitive word meaning. These Burmese sensitive words are not allowed by the Facebook team because they go against Facebook’s Community Standards on hate speech and inferiority. Thus, the writer writes English words together with Burmese text instead of these sensitive words.
   e.g., သတ်ဖြတ် → Tatဖြတ် (“Kill” in English) where the use of writing with an English letter “Tat” instead of Burmese “သတ်”.

7. *Sequence error (seq)*

   These errors can be caused by writing Burmese words with the wrong format sequence, which may be two or three consonant combinations, medials or vowels.
   e.g., မြို့ → ြမို့ (“City” in English) where the use of a writing sequence with “ြ+မ+ ိ+ ု+ ့” instead of “မ+ ြ+ ိ+ ု + ့”.

8. *Short-form error (short)*

   These errors occur when using a shortened form of a word or abbreviations for Burmese text. Shortcuts save time and effort. However, they are not always convenient and are inappropriate for certain settings.
   e.g., ပြီးပြီ → pီးp (“Finish” in English) where the use of short writing with “pီးp” instead of “ပြီးပြီ”.

9. *Slang word error (slang)*

   In this day and age, slang words play an important role on social media. It creates a barrier to communication for the uninitiated, and also decreases the quality of the Myanmar language.
   e.g., ဆရာမ → ဆျာမ (“Teacher” in English) where the use of writing with “ဆျာမ” instead of “ဆရာမ”.

10. *Stack word error (stack)*

    This error type typically occurs while misusing ***stacked words***:
    e.g., ကုမ္ပဏီ → ကုပ္မဏီ (“Company” in English) where the use of writing with “ပ္မ” instead of “မ္ပ”.
    double stacked word:
    e.g., မိတ္တူ → မိတ္ထူ (“Copy” in English) where the use of writing with “တ္ထ” instead of “တ္တ”.
    မင်္ဂလာ → မင်္လာ (“Mingalar” in English) where the use of writing with “ဂ်” instead of “ဂ််”.
    double-stacked Pali word:
    e.g., ပါဋ္ဌဆင့် → ပါ့ဆင့် where the use of writing with “ပါ့” instead of double-stacked Pali word “ပါဋ္ဌ”.

11. *Typographic error (typo)*

    These errors occur when the correct spelling of the word is known, but the word is mistyped by mistake. These errors are mostly related to the keyboard and, therefore, do not follow any linguistic criteria.
    e.g., နိုင် → နှိုင် (“win” in English) where the use of writing with “နှိုင်” instead of “နိုင်”.

Table 3: The statistics for eleven different types of errors

Error Types	Closed-test (sentences)	Open-test (sentences)
(1) con	3,812	282
(2) dialect	58	8
(3) encoding	384	32
(4) pho	19,722	2,096
(5) pho-typo	3,152	270
(6) sensitive	73	11
(7) seq	2,068	206
(8) short	149	16
(9) slang	1,170	127
(10) stack	795	109
(11) typo	18,733	2,205

4.2 Manually Annotated Spelling Error Corpus

The Burmese spelling error-correction parallel corpus is developed by collecting manually spelled mistakes that are commonly found on actual data sources such as social media (Facebook), news (e.g., BBC Burmese, 7Day News, health-related news from Myanmar Times), and several blogs such as travel, food, and beauty (e.g., MYSTYLE Myanmar). This valuable data was gathered over a two-year period, and Myanmar3 Unicode was used to implement a Burmese spell checker. In this corpus, the pair of noisy and correct sentences are separated by three pipes, and organized as the parallel corpus for this spelling checker process. To save time, the annotator (part of speech tagging) worked on different types of spelling error tagging. The noisy word is manually annotated according to eleven error types, and these errors are provided with corrections to be prepared as the parallel correction corpus. These sentences are prepared as the testing data by using the corpus-based N-gram language model [7]. The N-gram analysis is described as a method to find incorrectly spelled words in a mass of text. Instead of comparing each entire word in a text to a dictionary, just the N-grams are controlled. We use an N-gram model (mainly bi-gram) trained on the largest available corpus to date based on the amount of noise present in the data. Then, this system uses the bi-gram syllable error and correction pair corpus calculation based on the annotated word by using a rule-based approach. The formula for the bi-gram model is as follows:

$$P(W_n|W_{n-1}) = \frac{C(W_{n-1}W_n)}{C(W_{n-1})} \qquad (1)$$

where, P is the conditional probability, C is the count of the N-gram in the corpus, and W_n is the n^{th} word in the sentence.

If N = 2 (known as bi-gram), the bi-gram syllable error and correction annotated parallel pair for the “pho-typo” error type would be:
(“The children are happy.” in English)

ကလေး<တေ|pho-typo>ပျော်တယ်|||ကလေး<တွေ>ပျော်တယ်

4.3 Syllable Segmentation

Generally, Myanmar words are composed of multiple syllables, and most of the syllables are composed of more than one character. Syllables are also the basic units for the pronunciation of Myanmar words. If we only focus on consonant based syllables, the structure of the syllable can be described with Backus Normal Form (BNF) as follows:

Syllable := CMV[CK][D]

Here, C is for consonants, M for medials, V for vowels, K for vowel killer characters, and D for diacritic characters [8]. In our experiments, we used a RE-based Myanmar syllable segmentation tool named “sylbreak” [9] or “sylbreak4all” [10]. The following is an example of syllable segmentation:

Unsegmented sentence: ကလေးတွေပျော်တယ်
(“The children are happy.” in English)

Syllable segmentation: က လေး တွေ ပျော် တယ်

5 Experimental Methodology

In this section, we describe the methodology used in our spelling checker experiment for this proposed paper. The various techniques that were designed on the basis of spelling errors and trends, also called error patterns, have been studied. For spell checkers, rule-based approaches are a popular choice. The pattern picks up incorrect parts of words, and replaces them with the relevant correct word using a special form of regular expression. They work by having a set of rules that capture common spelling errors and applying these rules to the misspelled word. Each correct word generated by this process is taken as a correction suggestion. This proposed spelling checker is experimented with by using an automatic rule extraction approach.

5.1 Automatic Rule Extraction Approach

In this approach, first we tokenize the input text of each error type into syllable (using sylbreak4all [10]). And then, we use Wdiff [11] for comparing two texts (i.e., error and correction parallel pair texts). Wdiff works by creating two temporary files, one word per line, and then executing the difference on these files. It collects the difference output and uses it to produce a nicer display of word differences between the original files. According to the nature of our data, some data sets are comprised of both the prefix and suffix syllable data on the error and correction, some include only the prefix or suffix syllable data, and some have neither prefix nor suffix. Hence, five different error-correction patterns are defined based on our data. These five different patterns are "prefix-error-correction-suffix" (PECS), "prefix-error-correction" (PEC), "error-correction-suffix" (ECS), "error-correction" (EC), and all patterns (ALL). And then, we convert this Wdiff pattern into five different patterns. By doing so, we get a total of 38,124 unique rules for the closed-test data of all eleven error types. These unique rules are changed to regular expression (RE) based "/search/replace/" pair rules for handling most common errors, as it is the only approach that enables users to receive specific error detection and correction. Based on extracted RE-based "/search/replace/" pair rules, we made the automatic spelling error correction for each error type. The overview of this proposed spelling checker is shown in Figure 1.

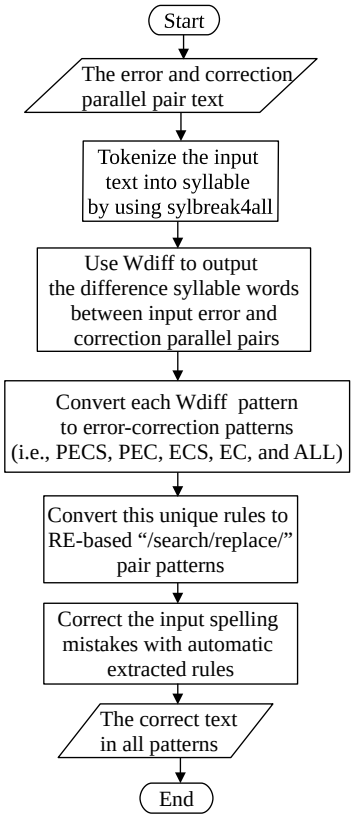

Figure 1: The overview of the automatic rule extraction and spelling correction

Some examples of five different error-correction patterns for the "pho-typo" error type "ကလေးတေပျော်တယ် <tab> ကလေးတွေပျော်တယ်" ("The children are happy." in English) are as follows:

1. **prefix-error-correction-suffix (PECS) pattern**

   လေး [-တေ-] {+တွေ+} ပျော် PECS

   Both prefix and suffix must be included in this pattern, where "လေး" is the prefix syllable, "[-တေ-]" is the error syllable, "{+တွေ+}" is the correction syllable, and "ပျော်" is the suffix syllable of the error correction pattern.

2. **prefix-error-correction (PEC) pattern**

   လေး [-တေ-] {+တွေ+} PEC

   The prefix must be included in this pattern, where "လေး" is the prefix syllable, "[-တေ-]" is the error syllable, and

"{+တွေ+}" is the correction syllable of the error correction pattern.

3. **error-correction-suffix (ECS) pattern**

   [-တော-] {+တွေ+} ပျော် ECS

   The suffix must be included in this pattern, where "[-တော-]" is the error syllable, "{+တွေ+}" is the correction syllable, and "ပျော်" is the suffix syllable of the error correction pattern.

4. **error-correction (EC) pattern**

   [-တော-] {+တွေ+} EC

   Neither prefix nor suffix must include in this pattern, where "[-တော-]" is the error syllable, and "{+တွေ+}" is the correction syllable of the error correction pattern.

5. **"ALL" pattern**

   လေး [-တော-] {+တွေ+} ပျော် ALL

   This "ALL" pattern is all of the pattern ranges that are extracted as the first syllable and the next one of the error-correction syllable pair. This pattern may include either PECS, or PEC, or ECS, or EC patterns.

The process flow of the automatic rule extraction approach for the Burmese Spelling Checker is shown in Figure 2. The input text is "ကလေးတော်ပျော်တယ် <tab> ကလေးတွေပျော်တယ်" ("The children are happy." in English) as error and correction parallel pair which is separated by Tab ("<tab>"). The left hand side of Tab is the error text (a red-colored syllable is an error syllable) and the right hand side is the correction text (a green-colored syllable is a correct syllable). Furthermore, these input text is tokenized into syllable words, and Wdiff is used to output the difference syllable words between input error and correction parallel pairs. The difference syllable words are [-တော-] {+တွေ+}, and the same syllable words are က လေး and ပျော် တယ်. Each Wdiff pattern is converted again to five error-correction patterns. And then, these unique rules are also changed to RE-based "/search/replace/" pairs "s/လေးတော်ပျော်/လေးတွေပျော်/g;" (as an example in PECS pattern).

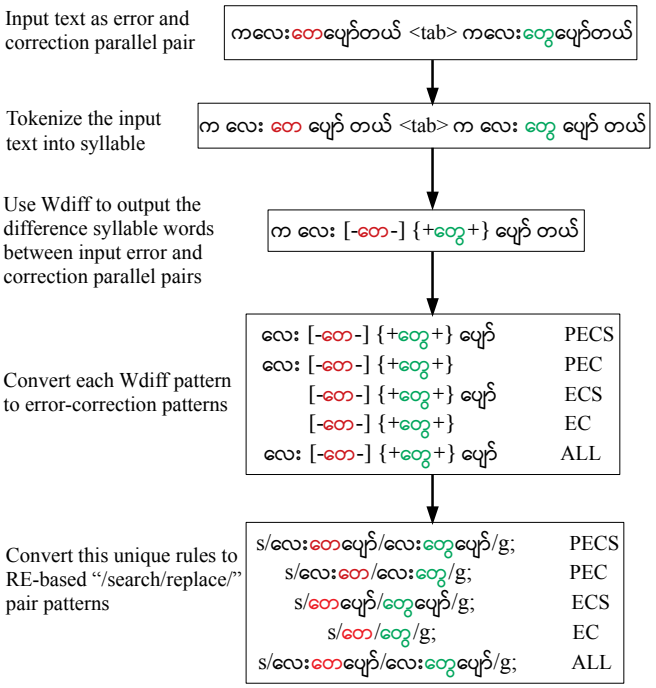

Figure 2: The process flow of automatic rule extraction approach for Burmese Spelling Checker

5.2 Evaluation

The performance of this proposed spelling checker is measured by calculating GLEU (the Generalized Language Evaluation Understanding) scores [12] and the F1-scores. The GLEU metric, a simple variant of BLEU, was proposed for evaluating grammatical error corrections using N-gram overlap with a set of reference sentences, as opposed to precision/recall of specific annotated errors (Napoles et al., 2015) [13]. GLEU more closely models human judgments than other metrics because it rewards correct edits while penalizing ungrammatical edits, while capturing fluency and grammatical constraints by virtue of using N-grams. This GLEU score's range is always between 0 (no matches) and 1 (all match). The modified precision calculation of GLEU^{+} equation is stated as in Equation 2 and Equation 3, in which, C is the correction candidate, S is the source or the input sentence with error, R is the reference sentence, and n is the maximum order of N-gram.

This spell checking performance is also evaluated using three measures of assessment: Precision (P), Recall (R), and F-measure (F1). Precision (P) means the

$$p_n^* = \frac{(\sum_{ngram \in \{C \cap R\}} count_{C,R}(ngram) - \sum_{ngram \in \{C \cap S\}} max[0, count_{C,S}(ngram) - count_{C,R}(ngram)])}{\sum_{ngram \in \{C\}} count(ngram)} \quad (2)$$

$$count_{A,B}(ngram) = \min(\#\, occurrences\, of\, ngram\, in\, A,\ \#\, occurrences\, of\, ngram\, in\, B) \quad (3)$$

percentage of the correct word suggested by the system, which is divided by the total number of errors detected by the system. Recall (R) means the percentage of correct words suggested by the system, which is divided by the total number of sentences. The F1-score (F1) is the mean of recall and precision as follows:

$$F_1 = \frac{2PR}{P+R} \quad (4)$$

6 Results and Discussion

By observing the GLEU score results on both the closed-test and open-test (as shown in Figure 3 and Figure 4), the correction results of "pho", "pho-typo", and "typo" error types do not match in the error-correction (i.e., EC) pattern experiment. Besides, the corrected results for the error types such as "con", "dialect", "encoding", "pho-typo", "sensitive", "seq", "slang", and "stack" errors closely match the reference outputs in all of the patterns (i.e., PECS, PEC, ECS, and ALL) other than the "EC" pattern.

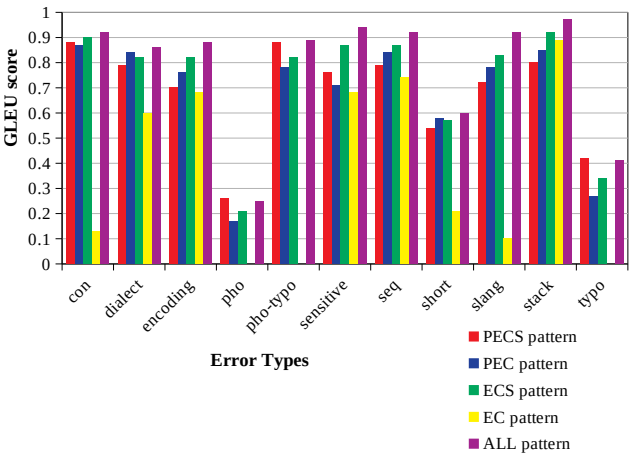

Figure 3: GLEU Score of the automatic rule extraction approach for the closed-test data

According to the F1-score results in the closed-test data experiment (as shown in Figure 5), the "ALL" pattern is more correct

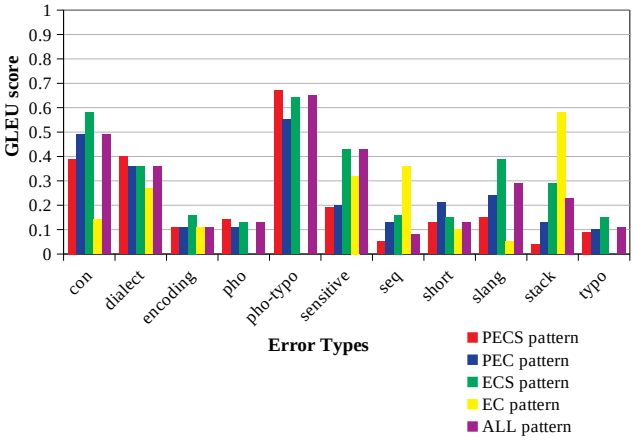

Figure 4: GLEU Score of the automatic rule extraction approach for the open-test data

than other patterns (i.e., EC, PECS, PEC, and ECS). In the open-test data experiment (as shown in Figure 6), the error types such as "pho", "pho-typo", and "typo" errors are corrected a little on the prefix-error-correction-suffix (i.e., PECS) pattern. We studied that these error types absolutely depend on the prefix and suffix. The "stack" word, "sensitive" word, and "sequence" error types are corrected efficiently on the error-correction (i.e., EC) pattern. Hence, these types of errors can be well-corrected without having a prefix or suffix. The short-form errors are totally dependent on only the prefix (i.e., PEC) pattern, and the dialect errors are also dependent on either the prefix or suffix (i.e., PECS, PEC, and ECS) patterns. The other error types, such as "con", "encoding", and "slang" word errors are closely corrected on the only suffix (i.e., ECS) pattern. On the basis of the original spelling error, the rate of F1-score increased from 72.62% to 96.42% (for "con"), from 70.97% to 91.31% (for "dialect"), from 59.92% to 90.05% (for "encoding"), from 73.94% to 44.36% (for "pho"), from 73.58% to 94.38% (for pho-typo), from 48.70% to 96.88% (for "sensitive"), from 68.75% to 94.03% (for "seq"), from 69.70% to 72.77% (for "short"), from

62.44% to 93.14% (for "slang"), from 69.00% to 97.92% (for "stack"), and from 72.33% to 71.14% (for "typo"). Therefore, we observed that all of the error types except for "pho" and "typo" are expertly corrected in their corresponding patterns according to the nature of the error types.

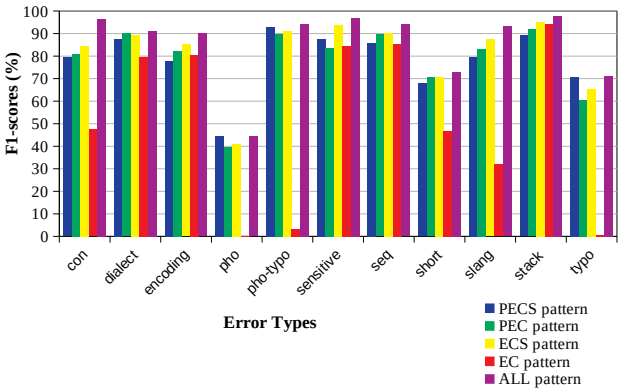

Figure 5: F1-score of the automatic rule extraction approach for the closed-test data

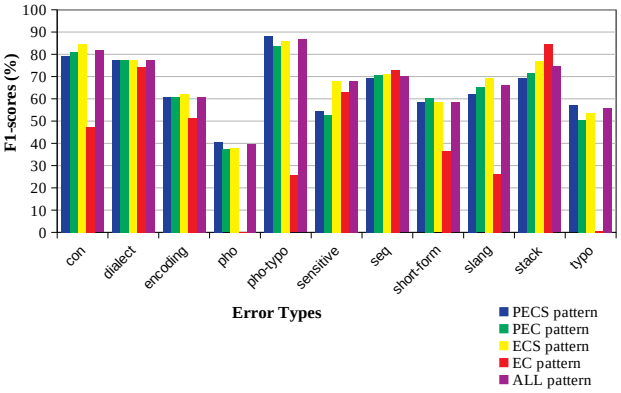

Figure 6: F1-score of the automatic rule extraction approach for the open-test data

7 Error Analysis

We analyzed the correct word suggestion of the proposed approach using the Word Error Rate [14] and the SCLITE toolkit [15]. "SOURCE" is the test sentence with error words, "Scores" are operation scores of the Edit Distance, "C" is the number of correct words, "S" is the number of substitutions, "D" is the number of deletions, "I" is the number of insertions, and N is the number of words in the reference ($N = S + D + C$). "REF" for reference (human written out), "HYP" for hypothesis (output from the proposed model), and "Eval" is the ordered sequence of edit operations. Note that if the number of insertions is very high, the WER can be greater than 100%. The lower the WER, the better the correction result is. The lower WER percentage is highlighted in green, and the greater WER percentage is also highlighted in red. The WER percentages for the closed-test and open-test are shown in Figure 7 and Figure 8, respectively. The formula for WER can be expressed as Equation 5:

$$WER = \frac{(I + D + S) * 100}{N} \quad (5)$$

Error Types / Five Patterns	WER (%)				
	PECS	PEC	ECS	EC	ALL
(1) con	7.2	6.8	5.4	51.1	**3.6**
(2) dialect	11.9	9.3	10.2	19.5	**8.0**
(3) encoding	24.2	18.9	16.1	20.3	**10.6**
(4) pho	**54.3**	60.4	58.3	**187.4**	54.7
(5) pho-typo	6.7	10.0	7.8	**100.2**	**5.0**
(6) sensitive	18.4	18.4	8.9	11.3	**4.5**
(7) sequence	15.2	11.3	10.2	16.0	**6.3**
(8) short	25.3	22.4	22.5	70.0	**19.9**
(9) slang	19.0	14.9	11.4	67.9	**6.4**
(10) stack	13.0	9.7	5.7	6.3	**2.3**
(11) typo	30.8	41.2	35.6	**101.2**	**30.0**

Figure 7: WER percentage of the automatic rule extraction approach for closed-test data

Error Types / Five Patterns	WER (%)				
	PECS	PEC	ECS	EC	ALL
(1) con	21.6	19.4	**15.8**	52.4	18.9
(2) dialect	22.6	22.6	22.6	25.8	22.6
(3) encoding	48.3	48.3	**46.7**	55.0	48.3
(4) pho	**58.5**	63.5	61.7	**186.0**	59.7
(5) pho-typo	**11.9**	16.2	12.8	**100.6**	12.4
(6) sensitive	77.8	77.8	46.7	**42.2**	46.7
(7) sequence	31.9	30.5	29.7	**27.4**	31.2
(8) short	38.2	**36.8**	38.2	86.8	38.2
(9) slang	37.2	34.9	**30.3**	75.8	33.6
(10) stack	39.7	36.7	29.7	**17.9**	33.2
(11) typo	**45.3**	52.2	48.0	**101.4**	46.1

Figure 8: WER percentage of the automatic rule extraction approach for open-test data

We also conducted manual error analysis on the corrected outputs of the automatic rule extraction approach, and we found

that several “confused” words and “falsely recognized” words are found in five different error correction patterns.

Analysis 1: An example of the “confused” words of the phonetic error type (see underlined word)
SOURCE: ကွန် ယက် ကုမ္ပ ဏီ (“Network Company” in English)
Scores: (#C #S #D #I) 3 1 0 0
REF: ကွန် ရက် ကုမ္ပ ဏီ
HYP: ကွန် ယက် ကုမ္ပ ဏီ
Eval: S

In Analysis 1, the pronunciations of the syllables “ရက်” and “ယက်” are the same, but the meanings are different. This proposed method does not well-handle or detect confusion errors.

Analysis 2: An example of the “Falsely Recognized” words of the phonetic error type (see underlined word)
SOURCE: များ သ လည်း (“How much?” in English)
Scores: (#C #S #D #I) 2 1 0 1
REF: များ သ ********* လဲ
HYP: များ သ လည် မ်း
Eval: I S

In Analysis 2, this automatic rule extraction approach falsely recognizes the error syllable “လည်း” as being inserted and substituted by “လည်” and “မ်း” instead of the correct syllable “လဲ”. The “falsely recognized” words “လည်” and “မ်း” are those hypothesis words which the recognizer incorrectly inserts and substitutes for the reference word “လဲ”.

8 Summary

This proposed paper contributes to the first investigation of the automatic rule extraction approach on the Burmese Spelling Checker. This approach can automatically extract correction rules from our parallel corpus and apply these rules to fix the spelling errors. We applied our developing Burmese spelling error-correction parallel corpus (23,860 sentences) and also containing 50,116 bi-gram syllable errors and correction parallel pairs. We analyzed the nature of Burmese spelling error types, and investigated the performance of eleven error types (i.e., con, dialect, encoding, pho, pho-typo, sensitive, seq, short, slang, stack, and typo) on five patterns (i.e., PECS, PEC, ECS, EC, and ALL) of automatically extracted error-correction pairs. Consequently, we tested the proposed spelling checker approach on both the closed-test and open-test data (statistical details are provided in Table 3). According to the experimental results, the proposed approach was found to be capable of extracting 38,124 spelling correction rules and applying them to correct the spelling errors of eleven different error types in the test corpus. We observed that while the phonetic “pho” and typographic “typo” errors are not well-corrected, the remaining error types are well-handled with meaningful outcomes by our automatic rule extraction approach. At present, our patterns are extracted based on 1-gram syllable before and after of the current error pattern. In the future, we plan to extract and study on more than 1-gram syllable. Furthermore, there are out-of-vocabulary (OOV) issues with this proposed approach because the current patterns are simple RE-based patterns such as the “/search/replace/” patterns. Therefore, we plan to study the performance of spell checkers by using sophisticated RE-based patterns such as fuzzy matching, which is a type of search that will find matches even when users misspell words or enter only partial words for the search. Moreover, we also have a plan to compare the experimental results between the manual and automatic rule extraction approaches.

References

[1] Nwe Zin Oo, and Tin Myat Htwe, “Myanmar Words Spelling Checking Using Levenshtein Distance Algorithm”, The proceeding of the 5th Conference on Parallel and Software Computing (PSC 2010), December 16, 2010, Yangon, Myanmar.

[2] Brody Downs, Oghenemaro Anuyah, Aprajita Shukla, Jerry Alan Fails, Maria Soledad Pera, Katherine Wright, and Casey Kennington, “KidSpell: A Child-Oriented, Rule-Based, Phonetic Spellchecker”, Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 11–16 May, 2020, Marseille, France, pp. 6937–6946.

[3] Novan Zukarnain, Bahtiar Saleh Abbas, Suparta Wayan, Agung Trisetyarso, Chul Ho Kang, “Spelling Checker Algorithm Methods for Many Languages”, International Conference on Information Management and Technology (ICIMTech), Jakarta/Bali, Indonesia, August 2019.

[4] Neha Gupta, and Pratistha Mathur, “Spell Checking Techniques in NLP: A Survey”, International Journal of Advanced Research in Computer Science and Software Engineering, India, Volume 2, Issue 12, December 2012, pp. 217-221.

[5] Aye Myat Mon, and Thandar Thein, “Myanmar Spell Checker”, International Journal of Science and Research (IJSR), India, [Online]. ISSN:

2319-7064, Volume 2 Issue 1, January 2013.

[6] Zar Zar Hlaing, Aye Thida, "Myanmar Homonym Disambiguation System", Seventeenth International Conference On Computer Applications (ICCA 2019), Yangon, 2019.

[7] N-gram: https://en.wikipedia.org/wiki/N-gram

[8] Myanmar Unicode Table, Range:1000–109F, [Online]. Available: (http://www.unicode.org/charts/PDF/U1000.pdf)

[9] Syllable segmentation tool for Burmese: [Online]. Available: (https://github.com/ye-kyaw-thu/sylbreak)

[10] Syllable Breaking Tool for Nine Ethnic Languages of Myanmar: [Online]. Available: (https://github.com/ye-kyaw-thu/sylbreak4all)

[11] GNU Wdiff, [Online]. Available: (https://www.gnu.org/software/wdiff/)

[12] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, "GLEU Without Tuning", 2016, arXiv:1605.02592. [Online]. Available: (http://arxiv.org/abs/1605.02592)

[13] Napoles, Courtney and Sakaguchi, Keisuke and Post, Matt and Tetreault, Joel, "Ground Truth for Grammatical Error Correction Metrics", Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), July 2015, Beijing, China, pp. 588-593. [Online]. Available: (http://www.aclweb.org/anthology/P15-2097)

[14] Word error rate (WER): [Online]. Available: (https://en.wikipedia.org/wiki/Word_error_rate)

[15] The National Institute of Standards and Technology (NIST), "SCTK", the NIST Scoring Toolkit, version: 2.4.11, November 11, 2018.