the OCR output texts or the text digitization process of an OCR system.

Recent OCR post-processing approaches explored different linguistic features and OCR error characteristics when employing *n*-gram language model [26] and error model [6]. Such important features have been successfully used in OCR error correction, including word *n*-gram frequency in [24, 25, 29, 35, 37, 39], string similarity in [24, 25, 35, 37, 39] and character confusion probability [18, 29, 37, 39].

Typically, Kissos and Dershowitz [29] suggested six linguistic features for a trained regression model, comprising probabilistic confusion matrix with a single edit, unigram frequency, backward and forward bigram frequency, term frequency in the OCR-ed text, and word confidence. In our opinion, the single edit in the confusion matrix has limitation when it is restricted to the maximum of two consecutive characters. This is due to that wrong character edits can happen at the different character positions of the erroneous words and character edit length can be more than two characters. Moreover, term frequency feature, computed by its frequency based on only the OCR text (instead of external large corpora), easily causes bias to consistent OCR errors. Besides, the important feature showing similarity between error and correction words used in real-word error correction approaches is not considered.

Mie *et al.* [35] employed a regression model to rank correction candidates. Six features of each candidate are extracted and used as inputs to the regression model. The features include Levenshtein edit distance, string similarity, unigram frequency and *n*-gram contexts. However, the confusion probability feature that captures OCR error characteristics is ignored in this approach. Furthermore, the edit distance feature shares some duplicate characteristic with the string similarity. It is obvious that the smaller the edit distance is, the higher the similarity is obtained.

Other OCR error correction schemes like [37, 39] have also considered such important features, for examples, employing *n*-gram context frequency and word similarity for the language model as well as constructing probabilistic confusion matrix for the OCR error model by the different methods. These schemes have shown successful and efficient performance in OCR error correction for English and Vietnamese languages, respectively.

Evolutionary algorithms are search methods that are used for solving optimization problems [9, 11, 42, 43, 45]. They mimic working principles from natural evolution by employing a population-based approach, labeling each individual of the population with a fitness value and including elements of randomness, although the randomness is directed through a selection process. Self-organizing migrating algorithm (SOMA) [51] is one of swarm-based algorithms which belong to the family of evolutionary algorithms. SOMA works on a population of candidate solutions in loops called migration loops. Each candidate will travel in the space of possible solutions. During the migration loops, new correction candidates will be found and scored with a fitness function.

In this paper, we introduce a novel automatic approach of Post-OCR error correction using *n*-gram language models and a candidate generation model based on multiple correction pattern edits under evolutionary loops. The *n*-gram data are constructed from a moderate-sized corpus (we use one billion word corpus, compared with one trillion word corpus used in other approaches such as [4, 24, 25, 35, 36]), and then further enriched with the training data. OCR output text is just tokenized on space with no restriction on punctuation in order to preserve the original OCR output words. With the help of *n*-grams, tokens are examined if they are erroneous in the error detection phase.

Unlike the approaches [29, 35, 36, 39] that have used the costly and complicated ensemble regression models to learn OCR error distribution and characteristics, we simply learn OCR errors by means of correction patterns directly obtained from the training data.

In the candidate generation model, we extensively explore candidates in candidate search space based on random correction pattern edits and evolved by evolutionary algorithm. It is shown that our model can achieve high-quality candidate generation through efficient algorithm parameter settings running on a normal PC with limited resources.

The candidates are scored and selected by a fitness function which is formulated based on a combination of modification versions of word-level linguistic features (using *n*-gram data) and substitution probability feature (using correction pattern data). Best candidates with highest scores are chosen as correction candidates of the error words. The evaluation results show that our approach has better performance than most of the top performers on the same evaluation dataset of English monographs in the ICDAR 2017 Post-OCR text correction competition [8].

Through employing the word-level linguistic features from the external corpus and obtaining correction patterns from the training data, our model is able to capture the diverse characteristics of both the language and the OCR errors. Our model can be used as a tool on OCR post-processing in various domains, wherein a sufficient training dataset in a domain is provided with OCR texts and corresponding Gold Standard (GS) / Ground Truth (GT) aligned at character level.

In summary, our model makes the following fivefold contributions:

- We introduce a novel automatic approach of OCR post-processing error correction using random correction pattern edits directed by evolutionary algorithm.