OCR correction for Indonesian historic newspapers using word repetition, stemmer and n-gram

D Purwantoro¹, H Akbar², A Hidayati³, Sfenrianto⁴
¹,²,⁴Bina Nusantara University, Indonesia
³Politeknik Negeri Jakarta, Indonesia

Email: danan.purwantoro@binus.ac.id

Abstract. Most digital archives of the old newspapers in Indonesia are only available as microfilm image file without their textual content. Manual transcription is certainly not effective and tiring for publishers that have large archives. Therefore, more automated transcription is required. As a part of that effort, this paper proposes OCR error correction of old spelling news articles utilizing new spelling databases. Spelling conversions which based on pattern are used to bridge spelling differences. The Error detection uses dictionary lookup and the phenomenon of word repetition and OCR errors that mostly are non-word errors. The Dictionary is built from KBBI and enriched with derivative words, English words, and entity names from validated news archives. Confix-stripping stemmer is used to validate derivative words while the English dictionary is used to validate English words in the news archive. The Error correction uses context-based methods by searching the phrase trigram/bigram word for each error in Google, then the Google Spelling suggestions are used as the correction. Experiments on 9 texts of OCR result of KOMPAS daily article between 1965 and 1966 is resulting the comparison of error rate before and after the correction (improvement ratio) of 193.01% compared to Hunspell spell checker of 61.47%.

Keywords—Optical Character Recognition, Error Detection, Error Correction, Google Spelling Suggestion, Dictionary, Stemmer, Repetition

1. Introduction
A more automatic transcription process usually consists of 3 phases, namely:
1. Pre-processing, image processing efforts to produce images that are more suitable for input OCR.
2. OCR, the phase of the text recognition process from pre-processing images.
3. Post-processing, the correction effort (post-correction) and further enrichment of the OCR result text.

Although there are a lot of researches and developments in the field of Pre-processing and OCR, such as [1,2], but the results of OCR transcriptions still have errors, so the error’s correction after OCR is still needed. This is the motivation in of this paper, which is developing OCR error correction methods. The old OCR image archive error is triggered by noise and poor print results, as shown in Figure 1. Especially for historic newspaper, the words that are old-spelling and sometimes in the form of entity names, foreign terms, and derivative words also complicate the correction process, because they’re not in standard dictionary (KBBI) which is commonly used as a basis for error correction system.
In general, word error OCR results can be divided into two [3]:
1. Non-word error: an error that is not in the word known in the dictionary.
2. Real-word error: an error in the form of a word known in the dictionary.

The error detection and correction method that are commonly used to overcome the word errors is divided into two [4]:
1. Isolated word-based.
2. Context-based error.

Isolated word-based correction is based on similar words. Some common measures of the words’ similarity of words are Levenshtein edit distance, n-gram character and LCS (Longest Common Subsequence). Different approaches use artificial neural networks and probability methods [5,6], but this method requires training data. The context-based method involves the context of the word. The context-based method when used for the detection and error correction process is able to correct real-word error [7-10]. According to [4], the error of OCR results is dominated by non-word error compared to real-word error. Therefore in this paper, we only use isolated word-based method for error detection because it’s enough to detect non-word error. Whereas for error correction we used context-based method so that the correction results are more appropriate since they are more contextual.

The isolated word-based error detection used in this paper uses dictionary lookup method. In order for the dictionary to cover more words that usually appear in news articles, the dictionary is built from a standard dictionary (KBBI) enriched with a collection of words from the news archives that have been validated. Validation is done so that only valid words will be taken, because there are also words resulting from typos in the news archive. Confix-stripping stemmer [11] is used to validate derivative words in the news archive. The Other research [12] uses morphological analysis to detect word errors, but this method risks recognizing the wrong words as derivative words because one of the results of the morphological analysis is lemma, for example: ‘memundukkan’ (should be ‘menundukkan’) when using morphological analysis [13] resulting lemma‘punduk’.

In addition to the word detection using dictionary lookup, this paper also utilizes the phenomenon of word repetition and OCR errors which are generally in the form of non-word errors. In one article, repeated words are generally the right words, while the wrong words are commonly unique because they’re are mostly non-word errors. The phenomenon of word repetition cannot be separated from the general principle that is applied in communication, namely 'principle of least effort'. This principle reveals the tendency to communicate efficiently with little effort [14]. This principle reveals a common phenomenon in which few words are often used and more words are rarely used. In a narrower context, such as in an article, the repeated words are not only words that are commonly used, but also words that are rarely used but become an important part of the article, e.g. a person's name will most likely be repeated in the article that tells about that person.

The context-based error correction method that is used in this paper uses searching trigram/bigram word phrase on Google and then use Google spelling suggestion result as correction for the wrong word. Figure 2 shows an example of Google spelling suggestion from search results on Google.
Internally, Google spelling suggestion uses a probabilistic n-gram model with a database of term collections and n-gram words which derived from millions of pages of public sites [8]. Google spelling suggestion-based method can reduce errors from 21.4% to 3.1% or an improvement ratio of 690% for English-language OCR texts [8]. If [8] relies entirely on Google for error detection and correction, then in this paper, Google is used only for error correction. In [8] Google search is done for every 5 blocks of words, while in this paper the search is only done for words that are detected incorrectly. So this will reduce access and dependency to Google.

The dictionary for word detection in this paper uses modern spelling (EYD to PUEBI), because it originates from newspaper articles between 1991 to and 2017, and KBBI (KamusBesar Bahasa Indonesia) IV. Likewise, Google, which is the source of error correction, tends to only recognize modern spelling and consider old spelling word as error as seen in Figure 3. Meanwhile the correction target is the OCR result from the news archive between 1965 and 1966 which has Soewandi spelling. So that it requires spelling conversion. Spelling conversions are based on spelling differences pattern. The spelling differences pattern between the Soewandi spelling and EYD are shown in Table 1 [15]. There are no spelling difference between EYD (Ejaan Yang Disempurnakan) and PUEBI (PedomanUmumEjaan Bahasa Indonesia), but there are addition of a diphthong letter, changes in the use of bold letters, and changes in capital letters at PUEBI [16].

| Soewandi (1947) | EYD (1972) |
|----------------|-----------|
| j              | Y         |
| tj             | c         |
| dj             | j         |
| j              | ny        |
| sj             | sy        |
| ch             | kh        |

2. Methodologies
The stages in this paper are described in the following sections.

2.1. Dictionary Creation
Special dictionaries are used for word detection so that more words can be detected when compared to using only standard dictionary (KBBI). Since in news articles, there are not only lemmas there, but also derivative words, entity names, and foreign terms. For this reason, dictionaries are made from a combination of KBBI (KamusBesar Bahasa Indonesia) and news archive. Words in the news archive are validated using the English dictionary for English words, confix-stripping stemmer for derivative words, and manual validation for words in the form of entity names. The use of stemmer to validate
derivative words is done by looking at the results of stemming, if it results lemma then the word is considered as a derivative word. The whole process of dictionary creation can be seen in Figure 4

![Figure 4. Data Collection Process for Dictionary](image)

Here are the steps for data collection for dictionaries:
1. Take all lemmas from KBBI IV (resulting 33069 words).
2. Take all English words (validated using English dictionary) from KOMPAS daily between 1991 and 2017 whose frequency is greater or equal than 10 (resulting 11705 words).
3. Take all derivative words (validated using Confix-stripping stemmer) from KOMPAS daily between 1991 and 2017 whose frequency is greater or equal than 10 (resulting 41033 words).
4. Take all correct words (manually validated) from KOMPAS daily between 1991 and 2017 which are noted in KBBI and English dictionary and not derivative words (after validated using Confix-stripping stemmer), and whose frequency $\geq$ 200 (resulting 2081 words which generally are entity names).

All combined together to form word error detection dictionary.

2.2. Design and Construction Error Correction System
The OCR error correction system includes two parts, i.e. error detection and error correction part that work sequentially as shown in Figure 5. The Error detection part is responsible for detecting word error, and error correction part is responsible for correcting the detected error previously.

![Figure 5. The Process in OCR Correction System](image)

Error detection uses rule based on word repetition (if there is the same word in the article then it is considered as correct word, otherwise it could be correct or error so it needs to be checked further in the dictionary) and dictionary lookup (if the word is exist in the dictionary then it is considered as correct, otherwise it is considered as error). Spelling conversion from old spelling (Soewandi) to modern spelling (EYD/PUEBI) is done before word token is checked in the dictionary. Completed flowchart of error detection can be seen in Figure 6.

The error correction process produces the use of Google spelling suggestion results from bigram/trigram phrase search on Google. The trigram/bigram phrase is resulted from sliding window method around detected error as illustrated in Figure 8. This method is used to obtain all possible trigrams/bigrams around the detected error. Google search is done for each trigram/bigram phrase, where trigram phrase is executed first, because it is more contextual than bigram. Searching is stopped if it results the spelling suggestion, which will be used as a correction for the error after it is converted to the old spelling (Soewandi). The completed flowchart diagram of the error correction process can be seen in Figure 7.
2.3. Error Correction System Evaluation

The OCR error correction system is tested using 9 text KOMPAS newspaper articles between 1965 and 1966 OCR results consisting of 2640 words. The OCR software used is Tesseract 4.0 version. The metric which is used to measure the error correction system performance is an improvement ratio, because it can describe a number of errors reductions after error correction system applied. Improvement ratio is calculated from the number of error rate before correction and after correction as described in (1),(2) and (3).

\[
\text{error rate before correction} = \frac{\text{number of errors before correction}}{\text{total words}} \quad (1)
\]

\[
\text{error rate after correction} = \frac{\text{number of errors after correction}}{\text{total words}} \quad (2)
\]

\[
\text{improvement ratio} = \frac{\text{error rate before correction}}{\text{error rate after correction}} \quad (3)
\]

3. Result and Analysis

The following is the result of error correction system testing, in terms of the effect of spelling conversion, word error detection results, and error correction results.

3.1. Spelling conversions

According to the experiment result, spelling conversion method that is developed in this paper is able to tackle spelling differences between correction target that uses old spelling (Soewandi) and
correction sources (dictionary and Google) that use modern spelling. There is only one correction error due to spelling conversion, i.e. the word 'Ljubljana' which becomes 'Ldjubldjana' (due to spelling conversion: j → dj). 'Ljubljana' is overseas place name that does not adhere to the old spelling (Soewandi), but not all of overseas place names like that, e.g. 'French' and 'Spanjol' which adhere to the old spelling (Soewandi).

3.2. Error Detection
The detection results compared with truth can be seen in Table 2. The number of correct words detected is 2263 out of the total 2364 correct words or the percentage is 95.73%. Meanwhile, the number of errors detected is 242 out of a total of 276 errors or the percentage is 87.68%.

Of the 2263 correct words detected, 1333 words or more than half are detected using rule based on word repetition phenomenon (repetitive word is considered correct). So this simple technique is able to detect the correct word very well. It can patch the lack of dictionary-based detection, as evidenced from those 1333 words, 64 of them are not in the dictionary, as seen in Table 3. These words include entity names such as ‘Radhakrishnan’, English terms such as ‘procede’, abbreviations such as ‘utk’ and old spelling words whose differences are not included in spelling differences pattern such as ‘anggauta’. There is only one detection error caused by this technique, that is the word ‘ada’ that should be ‘adalah’, because it appears more than once in the article so it is considered as correct word.

The usage of news archives to enrich dictionary other than KBBI (KamusBesar Bahasa Indonesia) has a significant effect on the results of word detection. Among all 850 words detected in the dictionary, 408 words of them come from news archives, mostly in the form of derivative words, entity names, and foreign terms, for example: ‘membantu’, ‘mesir’, and ‘clay’.

There are 34 errors but they are detected as correct words from the total of 276 errors, or 12.32% of errors. Among all 34 errors, almost all are real-word errors that cannot be detected by dictionary lookup because they are in the dictionary. Whereas 101 correct words that are detected as errors can be recovered in the error correction stage.

3.3. Error Correction
For error correction, our proposed system is compared with Hunspell spell checker, and the results are shown in Table 4. For all 9 articles, our system reduced number of errors from 276 to 143 or nearly half of them were reduced, or it has improvement ratio of 193.01%. Whereas Hunspell spell checker makes errors increased from 276 to 449, or has improvement ratio of 61.47%. So, our error correction system is better than Hunspell spell checker.

When it is examined per article, the results are varied. It was because each article has different errors, repeated words, and correct words that affects detection and correction process results. For example, article 2 contains many correct foreign words but they are not repeated and not in the dictionary so they are detected as errors. And when entering the correction stage, they are failed to be recovered because Google didn’t recognized trigram/bigram phrase related to them. It makes the improvement ratio is only 133.33%. Some of those words are: 'Michel', 'Kopchonge', 'Fujung', 'Hutjing', and 'SekoNoriko'.

Table 2. Detection result compared with truth

| Detection Result | Truth |
|------------------|-------|
| Correct          | 2263  |
|                  | 34    |
|                  | (seng, kana, den, caja, puter, mer, kedjuraan, uteri, pad, Setangan, Menter, kop, dima, paro, menjolok, Guruh, kapak, talah, diam, darat, dai, dati, Law, telan, Kelua, And, involusi, menanggapnja, Tetaplah, jai, lelah, meluru, Abdu, ada) |
| Error            | 101   |
| Total            | 2364  |
|                  | 242   |
Table 3. Correct word which is not in the dictionary but successfully detected because it repeated

| Article | Repeated word which is not in the dictionary |
|---------|---------------------------------------------|
| 1       | pint (2)                                    |
| 2       | Clarke(2), Jazy(2), jg(2), Keino(3), Li(3), Masako(2), Perantjis(2), Spasski(3), utk(2) |
| 3       | -                                           |
| 4       | proce(2)                                    |
| 5       | setya(2)                                    |
| 6       | Radhakrishnan(2), Shastri(4)                |
| 7       | -                                           |
| 8       | anggauta(2), Drs(2), Kombes(3), sd(2), Suhardjo(3), Surjobroto(2), tgl(2) |
| 9       | bersitegak(2), Harris(2), jg(3), Karno(2), kesetyaannja(2), konsekwen(2), setya(2) |
| Total   | 64 (28 distinct words)                      |

Table 4. Error detection result compared with hunspell spell checker

| Article | Total words | Error before correction | Error after correction | Improvement Ratio |
|---------|-------------|-------------------------|------------------------|-------------------|
|         |             | Proposed | Hunspell | Proposed | Hunspell |
| 1       | 214         | 12       | 3        | 23       | 400.00%  | 52.17%   |
| 2       | 676         | 56       | 42       | 149      | 133.33%  | 37.58%   |
| 3       | 80          | 11       | 6        | 15       | 183.33%  | 73.33%   |
| 4       | 189         | 33       | 17       | 43       | 194.12%  | 76.74%   |
| 5       | 409         | 40       | 18       | 39       | 222.22%  | 102.56%  |
| 6       | 228         | 4        | 1        | 27       | 400.00%  | 14.81%   |
| 7       | 162         | 26       | 11       | 29       | 236.36%  | 89.66%   |
| 8       | 256         | 34       | 19       | 60       | 178.95%  | 56.67%   |
| 9       | 426         | 60       | 26       | 64       | 230.77%  | 93.75%   |
| Total   | 2640        | 276      | 143      | 449      | 193.01%  | 61.47%   |

4. Conclusion

Based on the results of the research in this paper, error detection that utilizes the word repetition phenomenon (repetitive words in the article are mostly correct words) and dictionary lookup with dictionary that is rich with lemmas, derivative words, entity names, and English words, can detect 95.73% correct words and 87.68% errors. And only makes mistakes: as much as 12.32% errors detected as correct words. Confix-stripping stemmer can be used to find and validate derivative words in the news archive, proven in the experiment, 41033 derivative words have been obtained from KOMPAS archives between 1991 and 2017.

When it is viewed from error correction results, our proposed system that utilizes context-based error correction using Google spelling suggestion from trigram/bigram phrase search on Google can reduce almost a half of OCR errors (improvement ratio of 191.67%), instead of the Hunspell spell checker which makes errors increase for almost twice (improvement ratio of 61.47%).
Since the improvement error detection can be improved using context-based methods, so that it can detect real-word errors. Meanwhile, for the error correction, it can use its own local resources instead of internet resources, such as Google.

Acknowledgement
Thank you to the KOMPAS Daily for making this research happens.

References
[1] Koistinen, Mika, Kimmo K and Tuula P 2017. Proceedings of the 21st Nordic Conference on Computational Linguistics Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing.
[2] Bukhari S S, Kadi A, Jouneh M A, Mir F M and Dengel A 2017. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) anyOCR: An Open-Source OCR System for Historical Archives.
[3] Tong, Xiang and David A E 1996. Fourth Workshop on Very Large Corpora A statistical approach to automatic OCR error correction in context.
[4] Kukich K 1993. Proceedings of the 1993 ACM conference on Computer science - CSC 93 Techniques for automatically correcting words in text.
[5] Brill, Eric and Robert C M 2000. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics An improved error model for noisy channel spelling correction.
[6] Jurafsky and Martin J H 2014. Speech and language processing (London: Pearson).
[7] Islam A and Inkpen D 2009. Proceeding of the 18th ACM conference on Information and knowledge management - CIKM 09 Real-word spelling correction using Google web ITn-gram data set.
[8] Bassil Y and Alwani M 2012. Ocr post-processing error correction algorithm using google online spelling suggestion. arXiv preprint arXiv:1204.0191
[9] Bassil Y 2012. Parallel spell-checking algorithm based on Yahoo! N-grams dataset. arXiv preprint arXiv:1204.0184
[10] Bassil Y and Alwani M 2012. Computer and Information Science Context-sensitive Spelling Correction Using Google Web 1T 5-Gram Information. 5(3)
[11] Adriani M, Asian J, Nazief B, Tahaghoghi S M and Williams H E 2007. ACM Transactions on Asian Language Information Processing Stemming Indonesian. 6(4) pp. 1–33
[12] Soleh M Y and Purwarinti A 2011. Proceedings of the 2011 International Conference on Electrical Engineering and Informatics A non word error spell checker for Indonesian using morphologically analyzer and HMM.
[13] (2012) Morphological Analyzer - Language Grid Playground – Kyoto University. [Online], http://langgrid.org/playground/morphological-analyzer.html, tanggalakses: 28 Mei 2018.
[14] Zipf G 1972. Human behavior and the principle of least effort: an introduction to human ecology. (New York: Hafner)
[15] Kushartanti, Untung Y, dan Multamia RMT L 2005. Pesonabahasa: langkahawalmemahamilinguistik. (Jakarta: Gramedia Pustaka Utama)
[16] Badan Pengembangan dan Pembinaan Bahasa Kemdikbud 2016. Pedoman Umum Ejaan Bahasa Indonesia (edisi 4)