Building the Application to Identify Incorrect Capital Letters Writing in Bahasa Indonesia

D Gunawan¹, A Amalia², O N Maringga¹
¹Department of Information Technology, Universitas Sumatera Utara, Medan, Indonesia
²Department of Computer Science, Universitas Sumatera Utara, Medan, Indonesia
E-mail: danigunawan@usu.ac.id

Abstract. Ejaan Bahasa Indonesia (EBI) is a set of writing rules in Bahasa Indonesia which is implemented since 2015 according to the government rule Peraturan Kementerian Pendidikan dan Kebudayaan Republik Indonesia Nomor 5 Tahun 2015. This new set of rules replaces previously implemented rules which are known as Ejaan Yang Disempurnakan (EYD). The capital letter rule is one of the elements of EBI. According to the official dictionary of Bahasa Indonesia, the capital letter is a letter which has a bigger size and particular form than the common letter. This research yields an application to identify the incorrect capital letter writing in Bahasa Indonesia based on the user input. The application implements Part of Speech Tagging (POS), Named Entity Recognition (NER), N-Gram and set of rules based on EBI. As the conclusion, the application is able to check 81.57% words that should be written in capital form.

1. Introduction
Capital letters writing rule is one of the many rules in Ejaan Bahasa Indonesia (EBI) regulation. EBI is an official regulation for writing text in Bahasa Indonesia since 2015, replacing the previous regulation called Ejaan Yang Disempurnakan (EYD). Previous works found that many authors use more foreign computer terms instead of the official computer terms for Bahasa Indonesia on academic writing and government websites [1][2]. This issue might occur because of the lack of the application to check Bahasa Indonesia writing. In addition, there is also a limited number of applications to identify incorrect use of the capital letter in a text for Bahasa Indonesia. Therefore, this research is intended to yield an application to address the capital letter identification issue.

Previously conducted research in 2009 discussed restoring punctuation and capitalization in the transcribed speech by using finite state automata method [3]. This research found that more data training will improve performance. Meanwhile, the increasing of n-gram order does not have the significant contribution to the performance. This research yields the accuracy of 82.81% and 81.81% for trigram and quadrigram respectively. Another research observed automatic recovery of capitalization and punctuation of automatic speech transcript by using the Hidden Markov Model (HMM) algorithm and Maximum Entropy (ME) [4]. It is found that maximum entropy shows the better result for speech transcript. Meanwhile, hidden markov model performs better for the written-text corpus. In 2013, Beaufays utilized Finite State Transducer [5] and found that there is a common mistake in capitalization process, which is unable to recognize street name, business name and song title entities. This issue is addressed by adding the entities to the
corpus. The research conducted by Caranica [6] examined the punctuation and capital letter in Roman Language by using Statistical Language Model (SLM). This language model is used to predict the possibility of linguistic unit distribution and its order. The accuracy of this method reaches 80%.

This paper is organized as follow, section one describes the aim of this research and similar research in capital letter writing. The second section explains the research method and the third section discusses the result of the implemented research method. Then the last section summarizes the whole research.

2. Research Method

This capital letters checking is conducted by applying several steps. The first step is N-Gram calculation. This step is required to build the language model. This research limits the extraction to bigram (2-gram) and trigram (3-gram). For example, the sentence "Golkar menerima kekalahan ini dengan legowo", will be extracted to bigram and trigram as shown in table 1.

| bigram                   | trigram                                      |
|--------------------------|----------------------------------------------|
| Golkar menerima          | Golkar menerima kekalahan                   |
| Menerima kekalahan       | Menerima kekalahan ini                      |
| Kekalahan ini            | Kekalahan ini dengan                        |
| Ini dengan               | Ini dengan legowo                           |
| Dengan legowo            |                                              |

The second step is labelling the words by implementing Part-of-Speech (POS) tag algorithm. This research implements suggested POS tag by [7]. They suggested 35 POS tag to be used in Bahasa Indonesia. For example, the previous sentence example will be tagged into "Golkar/NN", "menerima/VBT", "kekalahan/NN", "ini/DT", "dengan/SC", and "legowo/NN". NN means common noun, VBT means transitive verb, DT means determiner, and SC means Subordination.

The third step is obtaining the entity by applying Name Entity Recognition (NER) algorithm. This step is required to determine the entity of a word, such as person, location, organization, quantity, time and other. The result of NER according to the previous sentence example is "Golkar/ORGANIZATION", "menerima/OTHER", "kekalahan/OTHER", "ini/OTHER", "dengan/OTHER", and "legowo/OTHER".

The last step is applying the rule to check whether the first letter of a word should be in capital letter or not. This rules are based on the official spelling rule of Bahasa Indonesia or known as Ejaan Bahasa Indonesia (EBI). The rules are:

(a) First letter in a sentence  
(b) First letter of a person’s name  
(c) First letter of a sentence in direct quotation  
(d) First letter of a religion, holy books, God (include the other names of God)  
(e) First letter of the honorary degree  
(f) First letter of the title
3. Result
This application is tested on a computer which has 2.3 GHz processor, 4 GB RAM and 500 GB hard disk. It requires Java version 1.8 and MySQL 5.5. This application has several modules such as N-Gram training module, POS tag training module, NER training module and Testing module.

3.1. N-Gram Training Module
This module is used to create a language model using text documents in Bahasa Indonesia. The input document should be written in correct Bahasa Indonesia (Follow the rule of Ejaan Bahasa Indonesia). This module will extract the text to bigram (2-gram) and trigram (3-gram), then calculate their frequency in the text. The result is the language model which will be used to compare the correct usage of the capital letters in a sentence.

As shown in Figure 1, there are two text areas in the N-Gram module. The left text area is used to show the text that has been loaded. The right text area is used to show the bigram and trigram result and the frequency next to it. The example of the N-Gram calculation result is shown in table 2.

![Figure 1. N-Gram Training](image-url)
3.2. POS Tag Training Module

This module is used to obtain a label from a document text. As well as the input document in N-Gram training module, the input document for POS tag training module has to follow the rule of Ejaan Bahasa Indonesia. The result of this module is the text which consists of words followed by their labels. Figure 2 shows there are two text areas in the POS tag training module. The input text is loaded to the left text area, and the right text area is used to show the result.

![Figure 2. POS Tag Training](image)

The example of POS tag labelling is shown in figure 3. The label’s description is based on work of Wicaksono as mentioned in section 2. The words "Dia" and "Mereka" are labelled to PRP, which means Personal Pronouns. The words "bangkit", "duduk", "bersepeda", "datang" and "gugur" are labelled to VBI. VBI means Intransitive Verb. The words "dari" and "di" are labelled IN (Preposition). The words "keterpurukan", "bangku", and "tanaman" are labelled to NN, which means Common Noun.

3.3. NER Training Module

NER is used to label the entities in the document text. There is six type of entities that will be labelled as shown in table 3. As shown in figure 4, the input text document is loaded to the left text area. The input text should be written in correct Bahasa Indonesia. The result will be shown in the right text area, along with the tag label described in table 3.
Table 3. Entity Label

| Entity label | Color |
|--------------|-------|
| PERSON       | Blue  |
| LOCATION     | Pink  |
| ORGANIZATION | Red   |
| QUANTITY     | Cyan  |
| TIME         | Yellow|
| OTHER        | Blank |
3.5. Evaluation

To evaluate we utilize precision, recall and f-score. The first step is transforming all the text to the lowercase. This step is required to allow the system recognizing the incorrect form. This research evaluates the system by testing ten articles that are obtained from several online articles such as detik.com, kompas.com and liputan6.com.

Table 4. Evaluation

| Article No. | TP | TN | FP | FN | Precision (%) | Recall (%) | F-Score (%) |
|-------------|----|----|----|----|---------------|------------|-------------|
| Article 1   | 45 | 109| 5  | 21 | 90.0          | 68.18      | 77.58       |
| Article 2   | 30 | 113| 0  | 12 | 100.0         | 71.42      | 83.33       |
| Article 3   | 40 | 158| 1  | 21 | 97.56         | 65.57      | 78.43       |
| Article 4   | 67 | 375| 5  | 32 | 93.05         | 67.67      | 78.36       |
| Article 5   | 47 | 183| 2  | 13 | 95.92         | 78.33      | 86.24       |
| Article 6   | 66 | 182| 6  | 4  | 91.67         | 94.28      | 92.96       |
| Article 7   | 47 | 182| 0  | 24 | 100.0         | 66.19      | 79.66       |
| Article 8   | 34 | 68 | 1  | 13 | 97.14         | 72.34      | 82.93       |
| Article 9   | 58 | 194| 6  | 24 | 90.62         | 70.73      | 79.45       |
| Article 10  | 47 | 182| 0  | 24 | 100.0         | 66.19      | 79.66       |
| **Average** | 45 | 109| 5  | 21 | **95.59**     | **72.09**  | **81.86**   |

Figure 5. Capital letter checking
4. Conclusion
According to the official dictionary of Bahasa Indonesia, the capital letter has bigger size and particular form than a common letter. Indonesia has a set of rules to write the capital letter that is defined in Ejaan Bahasa Indonesia (EBI). Unfortunately, there is a limited number of applications to identify incorrect capital letter writing. Therefore, this research proposes the application to address incorrect capital letter writing in Bahasa Indonesia. This application is able to check 81.57% incorrect capital letter writing and suggest the correct writing.

Acknowledgments
This research is supported by The Ministry of Research and Technology and the Higher Education Republic of Indonesia. The support is under the research grant TALENTA USU of the Year 2018 Contract Number 2590/UN5.1.R/PPM/2018 on March 16, 2018.

References
[1] Gunawan D, Amalia A, Lydia M S and Muthaqin M I 2018 Journal of Physics: Conference Series 979 012034 URL http://stacks.iop.org/1742-6596/979/i=1/a=012034
[2] Amalia A, Gunawan D, Lydia M S, Bonofi D and Nasution T H 2017 Identification of bahasa indonesia official computer terms in indonesian government websites 2017 International Conference on Electrical Engineering and Informatics (ICELTICs) pp 302–306
[3] Gravano A, Jansche M and Bacchiani M 2009 Restoring punctuation and capitalization in transcribed speech 2009 IEEE International Conference on Acoustics, Speech and Signal Processing pp 4741–4744 ISSN 1520-6149
[4] Batista F, Moniz H, Trancoso I and Mamede N 2012 IEEE Transactions on Audio, Speech, and Language Processing 20 474–485 ISSN 1558-7916
[5] Beaufays F and Strope B 2013 Language model capitalization 2013 IEEE International Conference on Acoustics, Speech and Signal Processing pp 6749–6752 ISSN 1520-6149
[6] Caranica A, Cucu H, Buzo A and Burileanu C 2015 Polytechnical University of Bucharest. Scientific Bulletin. Series C: Electrical Engineering and Computer Science 77 95 – 106 ISSN 2286-3540
[7] Wicaksono A F and Purwarianti A 2010 Hmm based part-of-speech tagger for bahasa indonesia Fourth International MALINDO Workshop