A comparative review of extractive text summarization in Indonesian language

W Widodo*, M Nugraheni and I P Sari
Department of Informatics Education, Universitas Negeri Jakarta, Jl. Rawamangun Muka, Jakarta, Indonesia

*widodo@unj.ac.id

Abstract. Text summarization has important role in natural language processing. One of text summarization type is extractive summarization. Research on text summarization in Indonesian Language is still rare and not evaluated comprehensively. Each research is only conducted based on subjectivity of researcher. This paper reviewed and evaluated some works on Indonesian Language Text Summarization for obtaining the better method by analysing some aspects. This review also mapped Indonesian text summarization evaluation techniques and obtained its advantages and drawbacks. This research aims to provide a comprehensive review of text summarization in Indonesian Language. Result of this study is a comparative review of some works which showed detailed aspects in summarization method.

1. Introduction
Generally, a document contains explanation about a topic and for scientific document, it could contain many sentences that usually needs more times to read. Text summarization become a better solution for quick reading but its whole meaning is not missing. A shorter representation that simpler to read and understand is created. A summary should maintain important information from original text [1]. Hence, automatic text summarization becomes part of Natural Language Processing (NLP) field which its main purpose is to create summary based on relevant sentences in original document [2-3]. This purpose seems like an information retrieval process with a little different. This difference is, in information retrieval, relevance refers to document while in text summarization, relevance refers to sentence.

Automatic text summarization is categorized into several types, based on input, output, content, and language [4]. Summary based on input is summarization by considering input document type. It is classified by single document summarization and multi-document summarization [1,4]. Summary based on output is extractive and abstractive summarization. Extractive summarization is summarization which results text summary from original document without modification, while abstractive summarization results texts which needs advanced language generation techniques [5]. It needs such natural language processing techniques. Summarization also can be categorized by content, those are generic, domain-specific, and query-based summary. Generic and domain specific describe its scope while query-based is similar to information retrieval by selecting sentences as user needs [6]. Monolingual, multilingual, and cross-lingual summary are summarization based on lingual.

Many works for reviewing studies in summarizing has been done. A simple review for describing works in summarization [7] until comprehensive survey of summarization [8-9] is emerged. All review tried to explain text summarization, even summarization works in some language [10]. However, review
works described text summarization almost English summarization, not works in Indonesian language. Therefore, our main contribution is first review of summarization work in Indonesian language.

2. Methodology
We conducted this study and it is a preliminary work in reviewing summarization papers for Indonesian language. In this preliminary work we compared six main papers in some aspect. The methodology is shown in figure 1 below.

Figure 1 shows the methodology we used. We did not use systematic literature review (SLR) because our work is only still preliminary research in reviewing papers and study of summarization in Indonesian language is rare. Therefore, we review in traditional technique and download straight in relevant papers.

2.1. Download relevant papers
We collected data only from one search engine, Google Scholar. The phrase we used in downloading papers is “text summarization Indonesian Language”. It is direct and straight phrase to obtain relevant papers in earlier ranking. We did not limited time published in downloading. 24 relevant papers were obtained in this process.

2.2. Select papers
We selected papers for preliminary review. Only papers published after 2016 are included. Papers which are written in Indonesian language is excluded. Finally, six main papers were obtained.

2.3. Compare the papers
In this stage, six papers are analysed in many aspects. We compare the method, data, result/finding for the main review, and information about authors and their affiliation as secondary review.

2.4. Map of study
This is the final stage of the work. When all papers have been reviewed, the studies can be mapped for resulting the aspects we reviewed.

3. Comparative analysis
This section explained comparison some aspects on text summarization in Indonesian language.

3.1. Methods
From six analysed papers, 9 methods have been obtained [11-16]. It is summarized in table 1 below.

| Study          | Methods                                      |
|----------------|----------------------------------------------|
| Slamet et al.  | Vector Space Model, TF-IDF                   |
| Yulita et al.  | Maximum Marginal Relevance (MMR)             |
| Gunawan et al. | TextTeaser and TextRank                      |
| Garmastewira   | Graph Convolution Networks (GCN)             |
| Sabuna         | Sentence Scoring, Decision Tree              |
| Christie       | Sentence Fusion                              |
3.2. Results
Work by Slamet using VSM and TF-IDF got accuracy level almost 40%. However, it cannot be really accurate since the amount of data only 5 articles. Slamet used journal article for this summarization. Their work then being cross-checked to manual abstracting, and three sentences or more are similar [11].

Work by Yulita using MMR with a little modification by utilizing WordNet and corpus to enhance result of summarization. This work is categorized by semantic-based MMR, that utilizes WordNet and corpus, and non-semantic based MMR by using TF-IDF. Data used in this work is 50 online news. The result shows with 30% compression ratio, semantic based MMR has $F1$-score 0.561 while non-semantic based MMR 0.598 [12].

Gunawan et al used 3075 online news for conducting their research. They employed TextTeaser which is previously only used for English and never being tested to others language. The result shows many things have to be improved because this method needs to be more adapted for Indonesian language. This resulted $F1$-score most in 0.7-0.79 with 25.3%. Average $F1$-score is not mentioned in this work, but it is around 0.5 [13].

Garmastewira used 87 news articles in their research. The utilization of Graph Convolution Networks (GCN) give a vary of this review, since this method is categorized as graph-based method while others are not. Garmastewira also used multi-document as input document, whereas others are single document. The result, by using ROUGE score 0.370 for 100 words summary and 0.378 for 200 words summary [14].

Sentence scoring and decision tree are used for summarizing by Sabuna. Sabuna used 8 features, those are tfidf, uppercase, proper noun, cue phrases, numerical data, sentence position, sentence length, and similarity to title. $F$-measure average result is 0.58 [15].

Christie used sentence fusion for integrating sentence summarization result. In its sentence selection, Integer Linear Programming (ILP) is employed. Some researchers categorize it as extractive method, however others categorize as semi abstractive, Data used came from news dataset. The result shows $f$-measure score is 3.91 out of 5 [16].

Six works on text summarization in Indonesian language shows that only 2 works using multi-document, other 4 are single document. The data used almost all use online news, unless work by Slamet [11] used journal articles. This can be a challenge for a future work. Detail summary of all six studies can be seen in next sub section.

3.3. Maps
This map shows some research position in text summarization in Indonesian language and potential future work to do. Table 2 shows categorisation of input document in our study.

| Input document | Numbers |
|----------------|---------|
| Single document| 4       |
| Multi document | 2       |
Table 3. Summary of study.

| Study id | Published in | Data | Methods | Result/Finding |
|----------|--------------|------|---------|----------------|
| [11]     | IOP Conf. Series: Materials Science and Engineering 288 (2017) 012037 | Journal articles – 5 | Vector Space Model (VSM) | Method weakness: manual abstracting has different words with original text (journal articles) |
| [12]     | IJCCS (Indonesian Journal of Computing and Cybernetics Systems) Vol.13, No.2, April 2019 | Online news – 50 | Marginal Maximum Relevance (MMR) | Higher compression rate – higher f-measure, non-semantic based produces better f-measure than semantic based |
| [13]     | IOP Conf. Series: Materials Science and Engineering 190 (2017) 012048 | Online news – 3075 | TextTeaser | TextTeaser still needs more adaptation in Indonesian language |
| [14]     | Journal of Information and Communication Technology, 18, No. 3 (July) 2019 | Indonesian news articles – 87 | Graph Convolution Network (GCN) | Three top sentences always talk same topic |
| [15]     | 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE) | Indonesian online news – 50 | Sentence Scoring – Decision Tree | Higher compression rate – higher f-measure |
| [16]     | 2016 International Conference on Advanced Informatics: Concepts, Theory and Application (ICAIICTA) | News dataset – [number is not mentioned] | Sentence Fusion | The process tends to follow semi abstractive |

Table 3 exhibits summary of this study, from table 3 there are still many studies in text summarization in Indonesian language can be conducted. Text summarization for journal article still rare in Indonesian language. It can be improved also by conducting text summarization for multi document with journal article as the data. What kind of evaluation should be used for this, due to journal articles more possibly being summarized by using abstractive method. While we utilize extractive method, is F-measure still relevant for evaluate the result? This can be good future work.

4. Conclusion
In section 2, it is mentioned that this work is a preliminary for reviewing text summarization in Indonesian language. For this preliminary study, six studies have been reviewed for papers which is published after 2016. All papers claimed using extractive method to summarize and all using different methods. 5 papers used online news data, while only 1 use journal article data. In average, from six papers, the best method is sentence scoring and decision tree by Sabana with f-score 0.58.

We can improve research in text summarization in Indonesian language by considering journal article as the data. It is also better for conducting it for multi document as input document.
References

[1] Yogan J K, Goh O S, Halizah B, Ngo H C and Puspalata C 2016 A review on automatic text summarization approaches Journal of Computer Science 12(4) 178-190

[2] Nazari N and Mahdavi M A 2019 A survey on automatic text summarization Journal of AI and Data Mining 7(1) 121-135

[3] Ježek K and Steinberger J 2008 Automatic text summarization (the state of the art 2007 and new challenges) Proceedings of Znalosti pp 1-12

[4] Nenkova A and McKeown K 2011 Automatic summarization Foundations and Trends in Information Retrieval 5(2-3) 103-233

[5] Das D and Martins A F 2007 A survey on automatic text summarization. Literature survey for the course Language and Statistics II (Technical report, Carnegie Mellon University)

[6] Gong Y and Liu X 2001 Generic text summarization using relevance measure and latent semantic analysis Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval pp 19-25

[7] Gaikwad D K and Mahender C N 2016 A review paper on text summarization International Journal of Advanced Research in Computer and Communication Engineering 5(3) 154-160

[8] Rudrapal D, Das A and Bhattacharyya B 2018 A Survey on Automatic Twitter Event Summarization Journal of Information Processing Systems 14(1) 79-100

[9] Moratanch N and Chitrakala S 2017 A survey on extractive text summarization 2017 international conference on computer, communication and signal processing (ICCCSP) (IEEE) pp 1-6

[10] Abualigah L, Bashabsheh M Q, Alabool H and Shehab M 2020 Text Summarization: A Brief Review Recent Advances in NLP: The Case of Arabic Language (Cham.: Springer) pp 1-15

[11] Slamet C, Atmadja A R, Maylawati D S, Lestari R S, Darmalaksana W and Ramdhani M A 2018 Automated text summarization for Indonesian article using vector space model IOP Conf. Series: Materials Science and Engineering 288 012037

[12] Yulita W, Priyanta S and Azhari S N Automatic Text Summarization Based on Semantic Networks and Corpus Statistics IJCCS (Indonesian Journal of Computing and Cybernetics Systems) 13(2) 137-148

[13] Gunawan D, Pasaribu A, Rahmat R F and Budiarto R 2017 Automatic text summarization for Indonesian language using textteaser MS&E 190(1) 012048

[14] Garmastewira G and Khodra M L 2019 Summarizing Indonesian News Articles Using Graph Convolutional Network Journal of Information and Communication Technology 18(3) 345-365

[15] Sabuna P M and Setyohadi D B 2017 Summarizing Indonesian text automatically by using sentence scoring and decision tree 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISSEE) (IEEE) pp 1-6

[16] Christie F and Khodra M L 2016 Multi-document summarization using sentence fusion for Indonesian news articles 2016 International Conference on Advanced Informatics: Concepts, Theory and Application (ICAICTA) (IEEE) pp 1-6