An idea based on sequential pattern mining and deep learning for text summarization

D S Maylawati¹²*, Y J Kumar², F B Kasmin¹ and M A Ramdhani¹

¹ Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia
² Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia

*diamsm@uinsgd.ac.id

Abstract. One of the Natural Language Processing (NLP) studies that has been widely researched is automatic text summarization. There are a lot of techniques and methods that are proposed for text summarization. However, not much attention has been given on the coherence and cohesion in text. The aim of this study is to present an idea to combine Sequential Pattern Mining (SPM) and Deep Learning (DL) for better text summarization process and result. In text summarization, it is important to produce understandable and readable summary, and SPM as text representation extracting algorithm is capable to maintain the meaning of text by giving attention of the order of words appearance. Whereas DL is a popular and powerful machine learning technique widely used recently in various data mining studies. This study uses descriptive research methodology that collects all of the facts and information which are related to SPM and DL for text summarization, where NLP as the body of knowledge, SPM and DL as the method, and text summarization as the domain problem that need to be solved. The findings of the study are presented as a logical design and mapping of current text representation that can be implemented to further improve automatic text summarization results, in particular, to improve its coherence and cohesion.

1. Introduction

Language is the learning ability [1], for humans to communicate and understand each other's information [2,3] both verbally and non-verbally [4]. The wealth of languages in the world that reaches 6000-7000 languages (not including local languages that are not identified) offers a very large resource for research in NLP studies [5,6]. This is because each language has its own characteristics, both from the form of letters, structure and grammar. NLP is a computational technique to process, learn, understand, extract important information or insight knowledge from human language content [7–9]. Among many NLP studies such as speech recognition, language translation, search engines, text similarity, sentiment analysis, opinion mining, intelligent personal assistants, chatbots, etc, text summarization is also widely studied.

Text summarization is a technique to summarize a documents computationally and automatically based on analytic result from main idea or key essence of documents [10,11]. To produce best summary result, it is important to prepare text representation well. This is because text representation is needed and play crucial roles to maintain the main idea or the meaning of text when performing the processes in text mining, text analytics or even NLP [12–14]. Generally, text representation has several types, among others Bag of Words (BoW) [15–19], binary representation [15], n-gram [15], [19], multiple of
words (MoW) or multi-words term [15,20], character and word embedding [21–23], semantic text representation [24–26], text representation for deep learning [21], symbolic and non-symbolic text representation [12], graph-based text representation [16], active descriptive text representation [23] and document specific representation [27]. Most text summarization research use BoW or n-gram as text representation and use TF-IDF to count the frequency of words [28–35]. However, not many research that have used Sequential Pattern (SP) or Frequent Pattern (FP) as MoW, called Sequence of Words (SoW), for text summarization research. The differences between SP and FP is that SP pays attention to the order of occurrence, while the FP does not pay attention to the order of occurrence. SP and FP as text representation make use of Sequential Pattern Mining (SPM) or Frequent Pattern Mining (FPM) technique, but based on previous research, from not less than 61 SPM and FPM algorithm [36-38], only around 22.69% that were used for text analytics. Most of the text summarization research that use SoW as text representation is conducted for English [30,39–44], several for Malay [45,46], and some for Spanish [47]. However, text representation for Indonesian language that use SoW has not been investigated. It will be interesting to see how it can be adapted for Indonesian language because the language has its own characteristic, structure, and grammar.

Basic text summarization research use Term Frequency-Inverse Document Frequency (TF-IDF) with sentence similarity or sentence weighting [32,34,48]. However, in many NLP research today, including text summarization, have employed Machine Learning (ML) approach such as Naive Bayes Classifier [35,47,49]; Vector Space Model [29]; Latent Dirichlet Allocation (LDA) and Genetic Algorithm (GA); K-Means [44]; and Adaptive K-Nearest Neighbor (AKNN) [33]. Moreover, Deep Learning (DL) was found to be effective and efficient for summarization process. Deep Learning is the enhancement of Artificial Neural Network (ANN) that contains multilayer between input layer and output layer [50,51]. There are several studies about text summarization that use DL method, among others: text summarization using unsupervised deep learning based on query-oriented extraction for document summarization with deep Auto Encoder (AE). The study proposed Ensemble Noisy Auto-Encoder (ENAE) as AE adds noise to the input text and selects the top sentence from an ensemble of noisy runs [52]. Other researchers use neural network model for learning phrase representation called RNN Encoder-Decoder [53]. Text summarization research has also implemented hybrid model based on neural network and fuzzy logic (neuro-fuzzy) [28]. Other related studies include removing redundant sentence for efficient text summarization using Restricted Boltzmann Machine (RBM) which is stochastic neural network [54]; using Recurrent Neural Networks (RNNs) to get meaningful summaries [55,56]; and using Deep Q-Network (DQN) for extractive summarization tasks [57]. Based on previous research studies discussed above, there is still wide opportunity to explore text representation for summarization research, especially for the Indonesian language, in particular, to use SP or FP to produce SoW as a structured representation of text and to use the currently trending DL that is used in NLP studies. Therefore, this article presents in-depth exploration about text representation using SPM and DL and the correlation between them for automated text summarization. In the next section, we present the methodology, result, discussion, conclusion, and further works of this research.

2. Review on text representation and sequential pattern mining

2.1. Current progress of text representation

Text is an unstructured data that must be structured before conducting the analysis or mining process. Text representation in this research means the structured form of text data which is prepared from text pre-processing so that it is ready to be mined [15,61]. Text representation has important role because it contains the meaning of text. Therefore, the impact of well-prepared text representation gives well text analytics result, even in text summarization. Text representation continued to develop from BoW as the basic of text representation until text representation that effectively used with DL method. From all of text representation that described in Figure 1, BoW is rarely used, because BoW is considered weak in maintaining the meaning of the text. N-Gram is better than BoW although it does not consider the order in which the word appears. Whereas the most frequently used are text representations that pay attention
to the order in which words appear or their semantics include MoW, SoW, and semantic text representation. Along with the development of NLP research using DL, then a text representation appears which is devoted to text analysis with DL.

Figure 1. Current progress of text representation.

2.2. Current progress of sequential pattern mining
Actually, Sequential Pattern Mining or (SPM) technique is a part of data mining that find sequential pattern (SP) from a recurring transaction sequentially [69–71]. Currently, SP is not only used for structured data, but also for unstructured data such as text. Over hundred of SPM and FPM algorithms are available and continue to be developed [36,38,72]. However, not many SPM or FPM is implemented for text analytics. PrefixSpan is one of SPM algorithm that is used for text data and FP-Growth (Frequent Pattern Growth) is the FPM algorithm that most used for text data [73,74]. Of course this result is influenced by the fact that PrefixSpan and FP-Growth are the forerunner to the development of the SPM and FPM algorithms after Association Rule algorithm.

3. Review on text summarization and deep learning for text summarization

3.1. Current progress of text summarization
Text summarization research is developed rapidly enough along with the development of text analytic research. The big challenge in text summarization research is how to produce readable summary [75–77], it means that the gap between summary result and reader understanding is not high. This article presents the elaboration of text summarization based on document type, summary unit, approach, impact, etc that is illustrated in Figure 2.
Figure 2. Current progress of text summarization.
3.2. Current progress of deep learning for text summarization

In trend of Deep Learning technique for NLP, text summarization research also uses DL for the efficiency and effectiveness of the summarization process. DL is basically used for data mining and was found to improve NLP, especially text summarization. Figure 3 describes DL methods that have been used for text summarization.

4. Proposed method

4.1. Methodology

This study uses descriptive research that does not involve measurements or numeric, but collecting and monitoring parameters, variables, or characteristics of research [58]. Descriptive research is one of the qualitative research methods that offer a deeper insight into the research problem that focuses on “what” than “why” [59,60]. As shown in Figure 1, this study proposes an idea on using text representation, text summarization, SPM and DL for text summarization. We will elaborate the correlation between SPM as a method that produces SoW as one of its text representations and how DL method can be used in text summarization process. Text representation that carries the meaning of text could affect the text summary result.
4.1.1. Relation between text representation, SPM, deep learning and text summarization. Text representation is a structured representation of text that contain the meaning of text. Thus, it must be prepared well. SPM is one of technique that can be used for producing SoW as one of text representation that have been proven capable to keep the meaning of text well. Not many researches that use SoW for text summarization, especially for Indonesian language and it can be a big opportunity for exploration. DL techniques which are currently in trend, are also a great opportunity to produce automatic summaries effectively and efficiently.

4.1.2. An idea to combine deep learning and sequential pattern mining for text summarization. Looking at the facts of the vastness of SPM studies that can be used as a technique to produce good text representation and developing DL technology that is proven to be effective and efficient in studying data mining, machine learning and NLP, Figure 5 presents a combination design between SPM and DL that can be used for automated text summarization in general.

Figure 5. General logical design to combine SPM and DL for text summarization.

5. Conclusion
In automated text summarization research, the summary result must be readable and understandable enough to the readers. The meaning of text can be maintained with well-prepare text representation that contain the meaning of text. This article elaborates and collects the facts and information related to text representation that are capable to be used in text summarization research to produce readable summary. From many text representations, Sequence of Words is a text representation with Sequential Pattern form and can be produced by Sequential Pattern Mining (SPM) algorithm. From many machine learning methods available today, Deep Learning (DL) is a popular method that has been used in Natural Language Processing including in text summarization. Based on the studies on text representation, SPM and DL methods have the potential to solve various type of cases in text summarization. This paper proposes an idea to combine SPM and DL to answer the challenges of text summarization for producing readable summary. In our ongoing work, combination of SPM and DL will be implemented and evaluated to investigate the process performance and the quality of summary result, in particular, to improve its coherence and cohesion.

Acknowledgments
Authors wishing to acknowledge Research and Publication Centre of UIN Sunan Gunung Djati Bandung that supports and funds this research publication.

References
[1] Fitch W T 2010 The Evolution of Language (New York: Cambridge University Press)
[2] Lieberman P 2015 The evolution of language Handbook of Intelligence: Evolutionary Theory, Historical Perspective, and Current Concepts
[3] Lieberman P 2016 The evolution of language and thought J. Anthropol. Sci.
[4] Anonymous 2019 Bahasa [Online] Retrieved from: https://id.wikipedia.org/wiki/Bahasa.
Accessed on: 09-Apr-2019

[5] Alyosha 2016 Jumlah Bahasa di Dunia WordPress Blog2 [Online] Retrieved from: https://alyoshaminded.wordpress.com/tag/jumlah-bahasa-di-dunia/ Accessed on: 09-Apr-2019

[6] Anonymous 2016 Terpetakan! Jumlah Bahasa di Seluruh Dunia, Di mana Posisi Indonesia? Republika Online [Online] Retrieved from: http://portalsatu.com/read/budaya/terpetakan-jumlah-bahasa-di-seluruh-dunia-di-mana-posisi-indonesia-5960 Accessed on: 09-Apr-2019

[7] Hirschberg J and Manning C D 2017 Advances in natural language processing Science

[8] Chaudhur B, Genthial G, Socher R, Fang M, Mundra R and Clark K 2017 Natural Language Processing with Deep Learning CS224n

[9] Kulkarni A and Shivananda A 2019 Deep Learning for NLP Natural Language Processing Recipes

[10] Thiagarajan V, Shah D B and Lalithamani N 2018 Text summarization J. Adv. Res. Dyn. Control Syst.

[11] Sarkar D 2016 Text Analytics with Python - A Practical Real-World Approach to Gaining Actionable Insights from Your Data

[12] Schuetze H, Adel H and Asgari E 2017 Non-symbolic Text Representation Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguist. Comput. Linguist. 1 785–796

[13] He Q, Chang K, Lim E-P and Zhang J 2013 Bursty Feature Representation for Clustering Text Streams Soc. Ind. Appl. Math. 491–496

[14] Maylawati D S, Aulawi H and Ramdhani M A 2019 Flexibility of Indonesian text pre-processing library Indones. J. Electr. Eng. Comput. Sci.

[15] Harish B S, Guru D S and Manjunath S 2010 Representation and classification of text documents: A brief review IJCA, Spec. Issue Recent Trends Image Process. Pattern Recognit. 2 110–119

[16] Jin W and Srihari R K 2007 Graph-based text representation and knowledge discovery SAC 807

[17] Pu W, Liu N, Yan S, Yan J, Xie K and Chen Z 2007 Local word bag model for text categorization Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguist. Comput. Linguist. 1 785–796

[18] Wallach H M 2006 Topic Modeling: Beyond Bag-of-Words ICML 1 977–984

[19] Sethy A and Ramabhadran B 2008 Bag-of-word normalized n-gram models Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 1594–1597

[20] Maylawati D S, Ramdhani M A, Rahman A and Darmalaksana W 2017 Incremental technique with set of frequent word item sets for mining large Indonesian text data 2017 5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017 1–6

[21] Young T, Hazarika D, Poria S and Cambria E 2018 Recent trends in deep learning based natural language processing [Review Article] IEEE Comput. Intell. Mag. 13 3 55–75

[22] Kageback M, Mogren O, Tahmasebi N and Dubhashi D 2014 Extractive Summarization using Continuous Vector Space Models Proc. 2nd Work. Contin. Vector Sp. Model. their Compos.

[23] Zhang Y, Lease M and Wallace B C 2016 Active Discriminative Text Representation Learning Proc. Thirty-First AAAI Conf. Artif. Intell. 3386–3392

[24] Zupanč K and Bosnič Z 2017 Automated essay evaluation with semantic analysis Knowledge-Based Syst.

[25] Dumais S T 2005 Latent semantic analysis Annu. Rev. Inf. Sci. Technol. 38 1 188–230

[26] Yeh J Y, Ke H R, W. P. Yang, and I. H. Meng, “Text summarization using a trainable summarizer and latent semantic analysis,” Inf. Process. Manag., vol. 41, no. 1, pp. 75–95, 2005.

[27] Singh K N and Devi H M 2017 Document representation techniques and their effect on the document Clustering and Classification : A Review Int. J. Adv. Res. Comput. Sci. 8 5 1780–1784

[28] Azhari M and Jaya Kumar Y 2017 Improving text summarization using neuro-fuzzy approach J. Inf. Telecommun. 1 4 367–379

[29] Slamet C, A. R. Atmadja, D. S. Maylawati, R. S. Lestari, W. Darmalaksana, and M. A. Ramdhani,
“Automated Text Summarization for Indonesian Article Using Vector Space Model,” in IOP Conference Series: Materials Science and Engineering, 2018.

[30] Ledeneva Y, Gelbukh A and García-Hernández R A 2008 Terms derived from frequent sequences for extractive text summarization Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 4919 LNCS 593–604

[31] Silvia, Rukmana P, Aprilia V R, Suhartono D, Wongso R and Meiliana 2014 Summarizing text for indonesian language by using latent dirichlet allocation and genetic algorithm Int. Conf. Electr. Eng. Comput. Sci. Informatics I 148–153

[32] Christian H, Agus M P and Suhartono D 2016 Single Document Automatic Text Summarization Using Term Frequency-Inverse Document Frequency ( Tf-Idf) ComTech Comput. Math. Eng. Appl. 7 4 285–294

[33] Putra J W G and Khodra M L 2017 Automatic Title Generation in Scientific Articles for Authorship Assistance: A Summarization Approach J. ICT Res. Appl. 11 3 253

[34] Hariardi W, Latief N, Febryanto D and Suhartono D 2016 Automatic Summarization from Indonesian Hashtag on Twitter Using TF-IDF and Phrase Reinforcement Algorithm Int. Work. Comput. Sci. Eng. (WCSE 2016) 17–19

[35] Najibullah A 2015 Indonesian Text Summarization based on Naïve Bayes Method Proceeding Int. Semin. Conf. 2015 Golden Triangle 67–78

[36] Fournier-Viger P, Gomariz A, Gueniche T, Soltani A, Wu C W and Tseng V S 2014 SPMF: A Java Open-Source Pattern Mining Library J. Mach. Learn. Res.

[37] Fournier-Viger P, Chun J, Wei Lin, Kiran R U, Koh Y S and Thomas R 2017 A Survey of Sequential Pattern Mining Ubiquitous Int. 1 1 54–77

[38] Fournier-Viger P, Lin J C W, Vo B, Chi T T, Zhang J and Le H B 2017 A survey of itemset mining Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery

[39] Baralis E, Caglierio L, Jabeen S and Fiori A 2012 Multi-document summarization exploiting frequent itemsets Proceedings of the 27th Annual ACM Symposium on Applied Computing - SAC ’12

[40] Baralis E, Caglierio L, Fiori A and Garza P 2015 MWI-Sum: A Multilingual Summarizer Based on Frequent Weighted Itemsets ACM Trans. Inf. Syst.

[41] Jayabhathary J, Kanmani S and Buvana 2012 Multi-document summarization based on sentence features and frequent itemsets Advances in Intelligent and Soft Computing

[42] Qiang J P, Chen P, Ding W, Xie F and Wu X 2016 Multi-document summarization using closed patterns Knowledge-Based Syst.

[43] Wu Y, Li Y and Xu Y 2019 Dual pattern-enhanced representations model for query-focused multi-document summarisation Knowledge-Based Syst. 163 736–748

[44] García-Hernández R A and Ledeneva Y 2009 Word sequence models for single text summarization Proc. 2nd Int. Conf. Adv. Comput. Interact. ACHI 2009 44–48

[45] Alias S, Mohammad S K, Hoon G K and Ping T T 2016 A Malay text summarizer using pattern-growth method with sentence compression rules 2016 3rd International Conference on Information Retrieval and Knowledge Management, CAMP 2016 - Conference Proceedings

[46] Alias S and Muhammad S K 2013 Sequential pattern based multi document summarization - An exploratory approach International Conference on Research and Innovation in Information Systems, ICRIIIS

[47] Villatoro-tello E, Pineda L V and Montes-y-gómez M 2006 Using Word Sequences for Text Summarization TSD, Text, Speech Dialogue 293–300

[48] Fachrurrozi M, Yusliani N and Yoanita R U 2013 Frequent Term based Text Summarization for Bahasa Indonesia Int. Conf. Innov. Eng. Technol. 30–32

[49] Ramanujam N and Kaliappan M 2016 An automatic multidocument text summarization approach based on naïve Bayesian classifier using timestamp strategy Sci. World J.

[50] Lecun Y, Bengio Y and Hinton G 2015 Deep learning Nature

[51] Ahmad J, Farman H and Jan Z 2019 Deep Learning Methods and Applications SpringerBriefs in
Yousefi-Azar M and Hamey L 2017 Text summarization using unsupervised deep learning Expert Syst. Appl.

Cho K 2014 Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) 1724–1734

Padmapriya G and Duraiswamy K 2014 An approach for text summarization using deep learning algorithm J. Comput. Sci.

Rossiello G 2016 Neural abstractive text summarization CEUR Workshop Proceedings

Shi T, Keneshloo Y, Ramakrishnan N and Reddy C K 2018 Neural Abstractive Text Summarization with Sequence-to-Sequence Models 1–28

Yao K, Zhang L, Luo T and Wu Y 2018 Deep reinforcement learning for extractive document summarization Neurocomputing

Ramdhani M A and Wulan E R 2012 The Analysis of Determinant Factors In Software Design For Computer Assisted Instruction Int. J. Sci. Technol. Res.

Association for Educational Communications and Technology 2001 What Is Descriptive Research? The Handbook of Research fo Educational Communications and Technology

Bhat A 2019 Descriptive Research: Definition, Characteristics, Methods, Examples And Advantages QuestionPro [Online] Retrieved from: https://www.questionpro.com/blog/descriptive-research/ Accessed on: 26-Feb-2019

Maylawati D S and Saptawati G A P 2017 Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang Journal of Physics: Conference Series

Doucet A 2006 Advanced document description, a sequential approach (University of Helsinki, Finland)

Ahonen-Myka H 2002 Discovery of Frequent Word Sequences in Text Proc. ESF Explor. Work. Pattern Detect. Discov. {LNCS} 24 Teollisuuskatu 23 180–189

Doucet A and Ahonen-Myka H 2004 Non-contiguous word sequences for information retrieval MWE ’04 Proc. Work. Multitword Expressions 26 88–95

Hoonlor A 2011 Sequential Patterns and Temporal Patterns for Text Mining (New York)

Rahmawati D, Saptawati G A P and Widyani Y 2015 Document Clustering using Sequential Pattern(SP) 2015 Int. Conf. Data Softw. Eng. (ICoDSE 2015) 98–102

Nalistia R As 2008 Feature-Based Clustering untuk Pengelompokan Dokumen dengan Representasi Text Terstruktur (Bandung)

Maylawati D S 2015 Pembangunan Library Pre-processing untuk Text Mining dengan Representasi Himpunan Frequent Word Itemset: Studi Kasus Bahasa Gaul Indonesia (Bandung Institute of Technology)

Agrawal R, Mannila H, Srikant R, Toivonen H and Verkamo A I 1996 Fast discovery of association rules Advances in knowledge discovery and data mining 12 307–328

Agrawal R and Srikant R 1994 Fast Algorithms for Mining Association Rules in Large Databases J. Comput. Sci. Technol. 15 6 487–499

Srikant R and Agrawal E 1996 Mining Sequential Patterns: Generalization and Performance Improvements 5th Int. Conf. Extending Database Technol. (EDBT ’96) 3–17

Fournier-Viger P, Lin J C W, Vo B, Chi T T, Zhang J and Le H B 2017 A survey of itemset mining Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 7 4

Maylawati D S 2018 The Concept of Frequent Itemset Mining for Text AASEC

Maylawati D S, Aulaw I H and Ramdhani M A 2018 The Concept of Sequential Pattern Mining for Text AASEC

Kumar Y J, Goh O S, Basiron H, Choon N H and Suppiah P C 2016 A review on automatic text summarization approaches Journal of Computer Science

Yao J, Wan X and Xiao J 2017 Recent advances in document summarization Knowl. Inf. Syst.

Nandhini K and Balasundaram S R 2013 Improving readability through extractive summarization
for learners with reading difficulties *Egypt. Informatics J.* **14** 3 195–204

[78] Allahyari M 2017 *Text Summarization Techniques : A Brief Survey*

[79] Patel M, Chokshi A, Vyas S and Maurya K 2018 Machine Learning Approach for Automatic Text Summarization Using Neural Networks *Int. J. Adv. Res. Comput. Commun. Eng.* **7** 1 194–202

[80] Khan A 2018 Abstractive Text Summarization based on Improved Semantic Graph Approach *Int. J. Parallel Program*

[81] Gambhir M and Gupta V 2017 Recent automatic text summarization techniques: a survey *Artif. Intell. Rev.*

[82] Gupta M 2018 *Automatic Text Summarization Manish Gupta*

[83] Johnson M E 2018 *Automatic Summarization of Natural Language Literature Review and Synthesis*

[84] Day M and Chen C 2018 Artificial Intelligence for Automatic Text Summarization *IEEE Int. Conf. Inf. Reuse Integr.* 478–484

[85] Gupta V, Bansal N and Sharma A 2019 Text Summarization for Big Data: A Comprehensive Survey *International Conference on Innovative Computing and Communications*

[86] Zopf M, Mene E L and Johannes F 2018 Which Scores to Predict in Sentence Regression for Text Summarization? *Proceedings of NAACL-HLT* 1782–1791

[87] Zhang Y, Liu Q and Song L 2018 Sentence-State LSTM for Text Representation *ACL2018*