Journal of Engineering Science and Technology
Vol. 14, No. 1 (2019) 498 - 508

DATA TO TEXT FOR GENERATING INFORMATION OF WEATHER AND AIR QUALITY IN THE *R* PROGRAMMING LANGUAGE

LALA SEPTEM RIZA*, BRAHMA PUTRA,
YAYA WIHARDI, BETA PARAMITA

Universitas Pendidikan Indonesia,
Jl. Setiabudhi 229, Bandung, Indonesia
*Corresponding Author: lala.s.riza@upi.edu

Abstract

Data obtained from weather monitoring stations and air quality acquired from several meteorological websites are numerical data. The research is aimed at developing a weather-generating Data to Text (D2T) that enables people to understand weather and air quality information. On developing D2T, there are four basic elements: signal analysis, data interpretation, document planning, as well as microplanning and realization. This research contributes to extending the signal analysis by adding new features: exponential smoothing for prediction time series data, a linear model with gradient descent for handling missing values and data summarization with statistical tools. In this study, several packages available in the *R* programming language were utilized. To validate the results, four measuring tools consisting of readability, computation time, relevance and truthfulness and comprehensibility and importance were used. After evaluating the results, it was stated that on the readability aspect the developed system can be understood by students at the age of 13-15 years; computational time is low (round 2.64 s) and sufficient scores for relevance and truthfulness and comprehensibility and importance were obtained. Therefore, the proposed system can be used to generate text for describing weather news and air quality.

Keywords: Data to text, Natural language generation, Natural language processing, *R* programming language, Time series data.

1. Introduction

1.1. Background

Today, many applications have been developed that can generate information in text form with non-linguistic input or numerical data. As introduced by Goldberg et al. [1], for example, which is the Forecast Generator (FOG). It can convert weather maps into predictions in sentence form by using natural language processors. Potret et al. [2] introduced another example in one of the health control applications is the BABYTALK family System. This application has been used to produce a summary of nursing changes only from electronic patient registration systems, at the Neonatal Intensive Care Unit (NICU) [3]. In general, the various examples that have been presented include the implementation of the D2T system.

The D2T system is part of the Natural Language Generation (NLG) system capable of generating sentences or text from numerical data automatically [4]. Because the system needs to convert from raw data, D2T must involve data analysis and linguistic processes. Reiter et al. [5] presented by developing the NLG architecture, the D2T system development stage consists of four steps: signal analysis, data interpretation, document planning and small planning and realization [4]. This is done to complement and improve the quality of information submitted to make it more easily understood by humans since the data obtained from the monitoring station is raw data.

Therefore, this research is focused on developing and implementing the D2T system in *R* programming language for describing and prediction weather condition and air quality. The proposed system is then implemented in the *R* programming language [6] since *R* provides many packages for dealing with time series datasets, machine learning and statistical tools.

1.2. Literature review

The D2T system is a Natural Language Generation (NLG) system capable of generating text from non-linguistic data inputs, such as sensor data and event logs [4]. Although D2T is part of the NLG, the biggest difference between D2T and NLG systems where inputs are knowledge-based, i.e., D2T systems must be able to analyse and translate inputs of data so that, there is knowledge that can be communicated and packed as possible in natural language.

Potret et al. [2] described one of the architecture of D2T that was applied in an application named BABYTALK. The D2T architecture of the BABYTALK application can be seen in Fig. 1.

From Fig. 1, it can be seen that the proposed architecture consists of four main stages, the stages consist of:

- Signal analysis: Potret et al. [2] implemented architecture in the D2T system. The main purpose of the signal analysis is to replace the numerical data into a discrete pattern.
- Data interpretation: The second step after getting the signals from the process of signal analysis. The main purpose of this Data Interpretation is to map the basic patterns and events into messages and relationships where people need them.
- Document planning: The third step in this architecture is to determine, which events will be mentioned in the text, as well as in the document structure.

Reiter and Dale [7] described that a series of processes Document Planning can be divided into content determination and document structuring.

- Microplanning and realization: The fourth step is to generate natural language in text form based on the content and structure selected at the planning stage of the document. In this process, the messages delivered will go through the following set of processes: lexicalization, aggregation, referring expression generation and structure realization.

Research related to the current D2T system has been of particular concern to researchers, as indicated by a large number of new studies related to this field (D2T and NLG) as explained in Table 1.

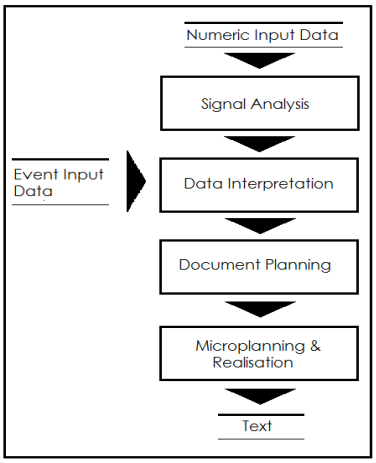

Fig. 1. Architecture of D2T [4].

Table 1. Related work on D2T and natural language generation.

References	Methods	Domains	Datasets
Boyd [8]	No content selection	Weather	Database
Sripada et al. [9]	Two stage model: (1) Domain reasoner; (2) communicative reasoner	Weather, Oil rigs	Sensor data, Numerical data
Sripada et al. [10]	Gricean maxims	Weather, Gas turbines, Health	Sensor data
Hallet et al. [11]	Rule-based	Heatlh	Database
Yu et al. [12]	Rules derived from corpus analysis and main knowledge	Gas turbines	Sensor
Sripada and Gao [13]	Decompression models	Dive	Sensor
Turner et al. [14]	Decision tree	Georeferenced data	Database
Gatt et al. [15]	Rule-based	Health	Sensor
Thomas et al. [16]	Document schema	Georeferenced data	Database
Demir et al. [17]	Rule-based	Domain independent	Graph-database
Redddington and Tintarey [18] and Tintarey et al. [19]	Threshold-based rules	Assistive technology	Sensor
Johnson and Lane [20]	Search algorithm	Autonomous underwater vehicle	Sensor
Banaee et al. [21]	Rule-based	Health	Grid of sensor
Schneider et al. [22]	Rule-based	Health	Sensor
Ramos-Soto et al. [23]	Fuzzy-sets	Weather	Database
Gkatzia et al. [24]	Rule-based	Weather	Numerical data with assigned probabilities

2. Methods

The D2T system model built in this research stands on the foundation of the architecture in the D2T system proposed in the literature [4] and some processes refer to what is done previously [7, 22, 25]. Therefore, the architecture developed in this research can be seen in Fig. 2. It is shown that our contributions are in signal analysis. Firstly, we utilize the *R* package "grad Descent", which is a package based on gradient descent dealing with regression tasks [26, 27]. In this case, gradient descent is used for handling missing values in data sets. Then, exponential smoothing was also used for analysing and predicting time series data sets. After that, the processes were adopted to accomplish the next steps in D2T (i.e., data interpretation, document planning and microplanning and realization) [4, 7, 22, 25].

Five test-case data were prepared for use, the dataset consists of 4 actual data, i.e., data obtained from http://www.MeteoGalicia.gal and one data modified by the author to identify system weaknesses. This division can be seen in Table 2.

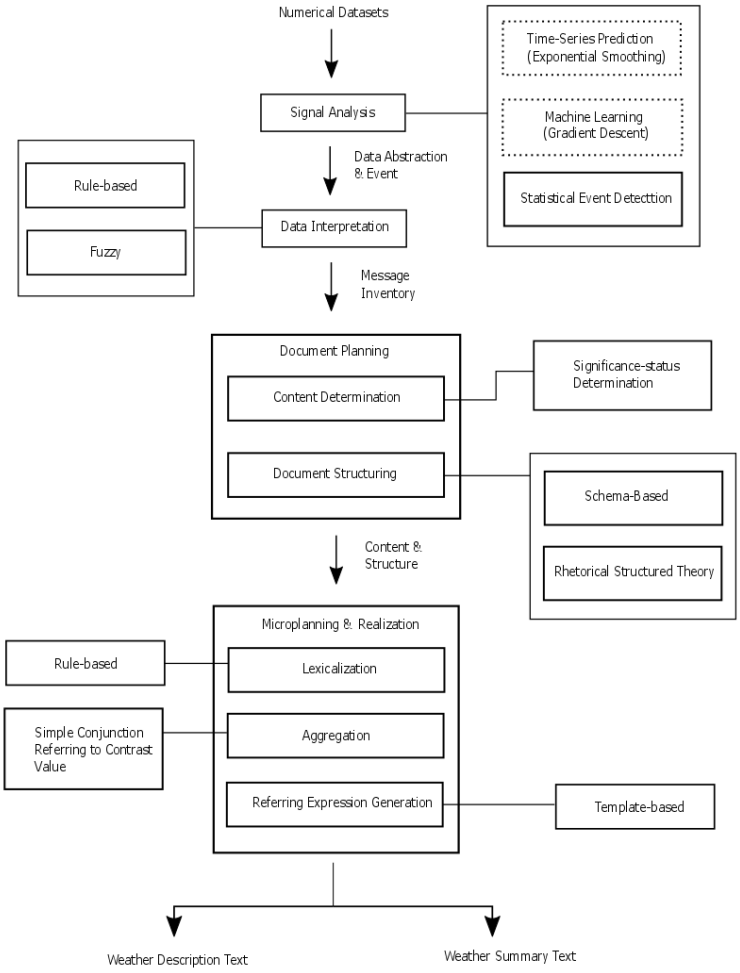

Fig. 2. Model of D2T with machine learning and time series for the weather information.

Table 2. Dataset used in experimentations.

Experimental	Dataset	Sources
1	2016-2017	Website www.MeteoGalicia.gal, during 2016-2017
2	2015-2016	Website www.MeteoGalicia.gal, during 2015-2016
3	2014-2015	Website www.MeteoGalicia.gal, during 2014-2015
4	2013-2014	Website www.MeteoGalicia.gal, during 2013-2014
5	Data TestCase1	Modified dataset to get all possible messages that will appear

After generating some texts from numerical input data, the results were evaluated and validated. Based on studies by Ramos-Soto et al. [25], testing was performed with 20 times NLG with different test-case inputs.

To evaluate and validate the experimental results the following four aspects were considered:

- Aspect of Readability, i.e., evaluating text quality by using Flesch Reading Ease assessment [28]. It is a method to evaluate levels of readability from texts.
- Aspect Computation Time, which is evaluating system-computing time by executing an *R* function.
- Relevance and truthfulness. Two questions were provided regarding the relevance and truthfulness aspects of the questionnaire for human experts.

  5 Likert scale was used for values of respondents' response containing strongly disagree (1), disagree (2), neither agree or disagree (3), agree (4) and strongly agree (5) [29].

Comprehensibility, i.e., evaluating the ease of information delivery, judged by the end-user. It can be seen that these aspects are measured by the questionnaire for end users.

3. Results and Discussion

3.1. Experimental results

For the first experiment, a dataset was used on the modified website. After the resurrection, the result can be seen in Fig. 3.

For the second experiment, a dataset is used on the modified website. After running the simulation, the results can be seen in Fig. 4.

In addition, Figs. 5 to 7 show results on the third, fourth and fifth experimentations respectively. It can be seen that the system was successful to generate texts from numerical data.

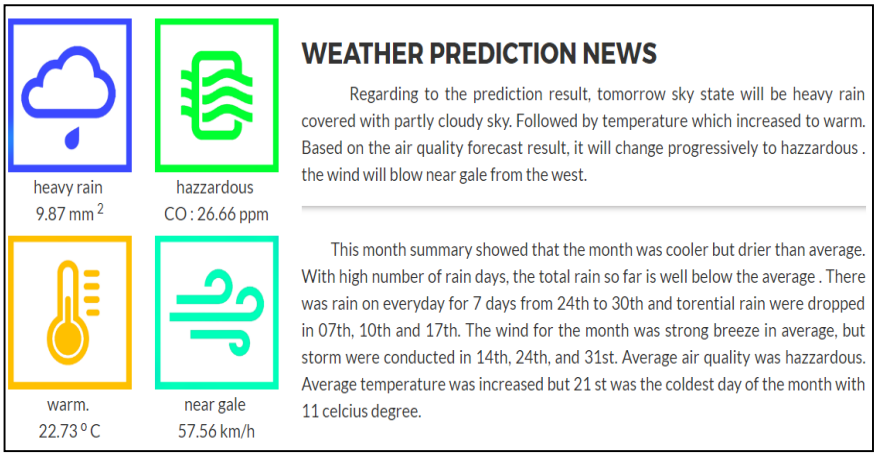

Fig. 3. Results on the first experiment.

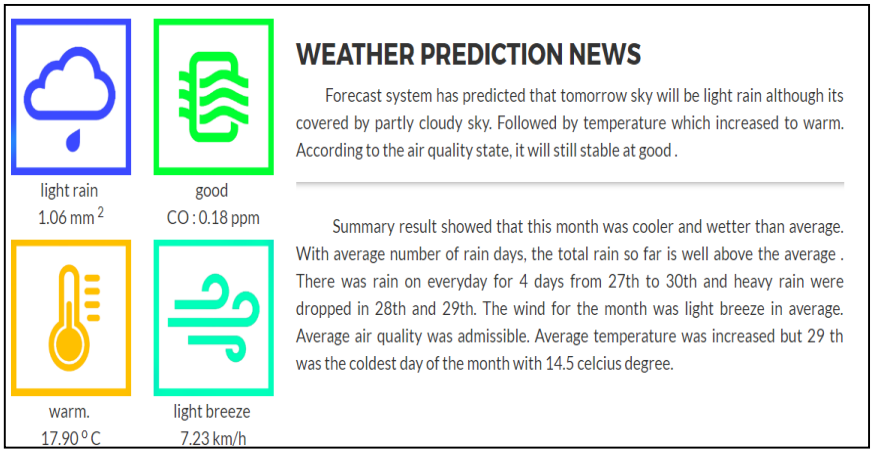

Fig. 4. Results on the second experiment.

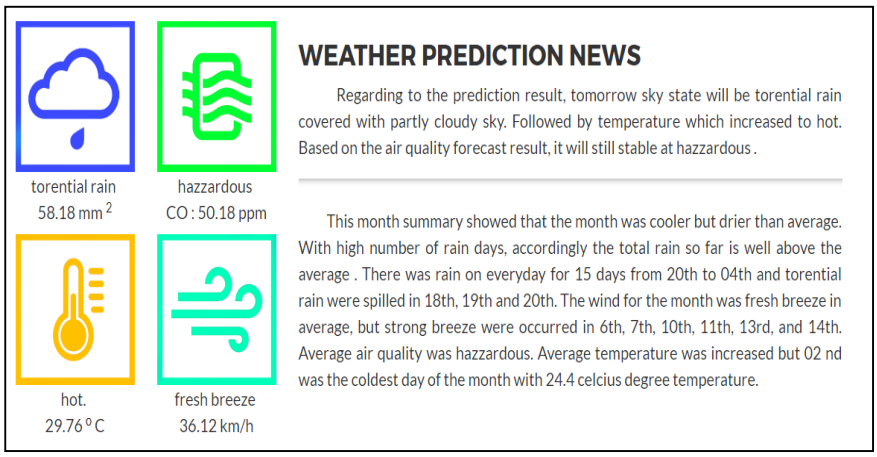

Fig. 5. Results on the third experiment.

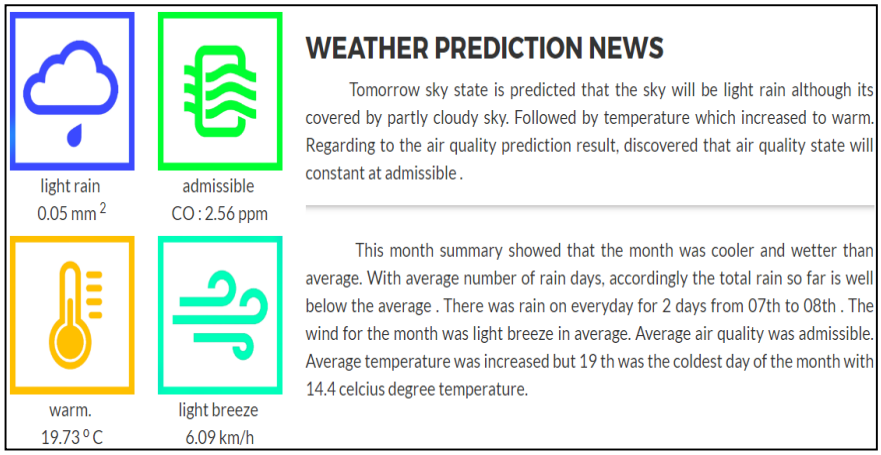

Fig. 6. Results on the fourth experiment.

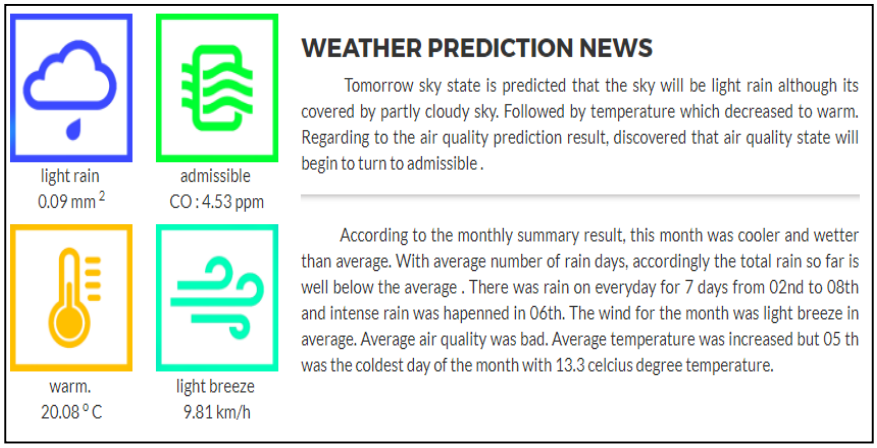

Fig. 7. Results on the fifth experiment.

3.2. Discussion

The following explanation is the analysis of the results obtained from the experiments:

- Result of analysis of readability aspects: The first analysis conducted is to determine the level of readability in the resulting text. Then, each generated text is evaluated using the Flesch Reading Ease method. The results of this test can be seen in Table 3.

  From Table 3, it can be seen that the highest Flesch Score is in the first experiment, with 67.24, while the lowest score is in the fifth experiment that is equal to 64.30. Then, since the average score obtained is at intervals of 60-70, the readability quality of the text generated from this system is: "Plain English, Easily understood by 13- to 15-year-old students".
- Computation Time Aspect: To evaluate computing time, each experiment is computed by computing time using the *R* function (i.e., system.time()). The

result of the computation time can be seen in Table 4. It can be seen that the average of the computation time is 2.64 s. It means that it is relatively fast.

- Relevance and Truthfulness Aspect: To evaluate the relevance and truthfulness aspects, A questionnaire was assessed by the expert. The results of this assessment can be seen in Table 5. It can be seen that the texts have good remarks on relevance and trustfulness.
- Evaluation of comprehensibility: To evaluate aspects of comprehensibility and importance, in this study used user-based evaluation (user-based). This assessment obtained data that can be seen in Table 6.

Therefore, according to the experimental results, the texts generated by the system with five input data can be easily understood by 13 to 15-year-old students, are taking 2.64 s. This results showed that making a simple protocol can make system to be user-friendly [30]. The results also have sufficient relevance, truthfulness, comprehensibility and importance. This is in a good agreement with the literature regarding the higher logical thinking of students [31].

Table 3. Results on evaluation using Flesch Reading Ease.

No.	Data-test-case	Flesch score
1	DT1	67.24
2	DT2	65.69
3	DT3	59.57
4	DT4	64.81
5	MT1	64.30
	Average	64.12

Table 4. Results on calculating computation cost by using system.time()

No.	Data-test-case	Running time(s)
1	DT1	2.60
2	DT2	2.64
3	DT3	2.94
4	DT4	2.51
5	MT1	2.51
	Average	2.64

Table 5. Results of questionnaire assessed by expert.

No.	Data-test-case	Relevance	Truthfulness
1	DT1	4	4
2	DT2	4	4
3	DT3	4	4.25
4	DT4	3.5	4
5	MT1	4	4

Table 6. Results on the questionnaires filled by end-users.

No.	Data-test-case	Comprehensability	Importance
1	DT1	4	4.5
2	DT2	4	4.5
3	DT3	3.5	3.5
4	DT4	4.5	5
5	MT1	3.5	3.5

4. Conclusion

Based on a series of research processes that have been done, it can be concluded as follows: D2T has been developed by adding functions implementing exponential smoothing for prediction time series data, a linear model with gradient descent for handling missing values and data summarization with statistical tools. Moreover, based on the experimental results, the quality of readability obtained by the Flesch Reading Ease method states that the resulting text is plain text that can be understood even by students at the age of 13-15 years. While the quality of relevance and truthfulness assessed by human experts obtained good results. In the aspect of acceptability and importance from the end user perspective, the system obtained sufficient value.

References

1. Goldberg, E.; Driedger, N.; and Kitterdge, R.I. (1994). Using natural-language processing to produce weather forecast. *IEEE Expert*, 9(2), 45-53.
2. Potret, F.; Reiter, E.; Gatt, A.; Hunter, J.; Sripada, S.; Freer, Y.; and Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8), 789-816.
3. Hunter, J.; Freer, Y.; Gatt, A.; Reiter, E.; Sripada, S.; and Sykes, C. (2012). Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial Intelligence in Medicine*, 56(3), 157-172.
4. Reiter, E. (2007). An architecture for data-to-text systems. *Proceedings of the Eleventh European Workshop on Natural Language Generation*. Schloss Dagstuhl, Germany, 97-104.
5. Reiter, E.; Sripada, S.; Hunter, J.; Yu, J.; and Davy, I. (2005). Choosing words in computer-generated weather forecast. *Artificial Intelligence*, 167(1-2), 137-169.
6. Ihaka, R.; and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3), 299-314.
7. Reiter, E.; and Dale, R. (2000). *Building natural language generation systems*. New York, United States of America: Cambridge University Press.
8. Boyd, S. (1998). TREND: A system for generating intelligent descriptions of time series data. *Proceedings of the IEEE International Conference on Intelligent Processing Systems (ICIPS1998)*. 5 pages.
9. Sripada, S.G.; Reiter, E.; Hunter, J.; and Yu, J. (2001). A two-stage model for content determination. *Proceedings of the 8th European Workshop on Natural Language Generation*-Volume 8. Toulouse, France, 1-8.
10. Sripada, S.G.; Reiter, E.; and Davy, I. (2003). SumTime-Mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3), 4-10.

11. Hallett, C.; Power, R.; and Scott, D. (2006). Summarisation and visualisation of e-health data repositories. *Proceedings of the UK E-Science All-Hands Meeting*. Nottingham, United Kingdom, 9 pages.
12. Yu, J.; Reiter, E.; Hunter, J.; and Mellish, C. (2007). Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(1), 25-49.
13. Sripada, S.G.; and Gao, F. (2007). Summarizing dive computer data: A case study in integrating textual and graphical presentations of numerical data. *Proceedings of the Workshop on Multimodal Output Generation*. Aberdeen, Scotland, United Kingdom. 149-157.
14. Turner, R.; Sripada, S.G.; Reiter, E.; and Davy, I.P. (2008). Using spatial reference frames to generate grounded textual summaries of georeferenced data. *Proceedings of the Fifth International Natural Language Generation Conference*. Salt Fork, Ohio. 16-24.
15. Gatt, A.; Portet, F.; Reiter, E.; Hunter, J.; Mahamood, S.; Moncur, W.; and Sripada, S. (2009). From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *AI Communications*, 22(3), 153-186.
16. Thomas, K.E.; Sripada, S.; and Noordzij, M.L. (2012). Atlas. txt: Exploring linguistic grounding techniques for communicating spatial information to blind users. *Universal Access in the Information Society*, 11(1), 85-98.
17. Demir, S.; Carberry, S.; and McCoy, K.F. (2012). Summarizing information graphics textually. *Computational Linguistics*, 38(3), 527-574.
18. Reddington, J.; and Tintarev, N. (2011). Automatically generating stories from sensor data. *Proceedings of the 16th International Conference On Intelligent User Interfaces*. Palo Alto, California, United States of America, 407-410.
19. Tintarev, N.; Reiter, E.; Black, R.; Waller, A.; and Reddington, J. (2016). Personal storytelling: Using natural language generation for children with complex communication needs, in the wild. *International Journal of Human-Computer Studies*, 92-93, 1-16.
20. Johnson, N.A.R.; and Lane, D.M. (2011). Narrative monologue as a first step towards advanced mission debrief for AUV operator situational awareness. *Proceedings of the 15th International Conference on Advanced Robotics (ICAR)*. Tallinn, Estonia, 241-246.
21. Banaee, H.; Ahmed, M.U.; and Loutfi, A. (2013). Towards NLG for physiological data monitoringwith body area networks. *Proceedings of the 14th European Workshop on Natural Language Generation*. Sofia, Bulgaria, 193-197.
22. Schneider, A.; Mort, A.; Mellish, C.; Reiter, E.; Wilson, P.; and Vaudry, P.L. (2013). MIME-NLG in pre-hospital care. *Proceedings of the 14th European Workshop on Natural Language Generation*. Sofia, Bulgaria, 152-156.
23. Ramos-Soto, A.; Bugarin, A.; and Barro, S. (2016). On the role of linguistic descriptions of data in the building of natural language generation systems. *Fuzzy Sets and Systems*, 285, 31-51.
24. Gkatzia, D.; Lemon, O.; and Rieser, V. (2016). Natural language generation enhances human decision-making with uncertain information. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistic*s. Berlin, Germany, 264-268.

25. Ramos-Soto, A.; Bugarin, A.; Barro, S.; Gallego, N.; Rodriguez, C.; Fraga, I.; and Saunders, A. (2015). Automatic generation of air quality index textual forecasts using a data-to-text approach. *Proceedings of the Conference of the Spanish Association for Artificial Intelligence*. Albacete, Spain, 164-174.
26. Riza, L.S.; Handian, D.; Megasari, R.; Abdullah, A.G.; Nandiyanto, A.B.D.; and Nazir, S. (2018). Development of R package and experimental analysis on prediction of the CO_2 compressibility factor using gradient descent. *Journal of Engineering, Science, and Technology (JESTEC)*, 13(8), 2342-2351.
27. Riza, L.S.; Nasrulloh, I.F.; Junaeti, E.; Zain, R.; and Nandiyanto, A.B.D. (2016). gradDescentR: An R package implementing gradient descent and its variants for regression tasks. *Proceedings of the 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*. Yogyakarta, Indonesia, 125-129.
28. Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221-233.
29. Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 55 pages.
30. Firdaus, C.; Wahyudin, W.; and Nugroho, E.P. (2017). Monitoring system with two central facilities protocol. *Indonesian Journal of Science and Technology*, 2(1), 8-25.
31. Haristiani, N.; Aryanti, T.; Nandiyanto, A.B.D.; and Sofiani, D. (2017). Myths, islamic view, and science concepts: The constructed education and knowledge of solar eclipse in Indonesia. *Journal of Turkish Science Education* (*TUSED*), 14(4), 35-47.