Word-Level Indonesian Story Generator with Markov Chain and Bidirectional GRU

Cecilia Angieta Winata*[1], Handri Santoso[2], Ito Wasito[3], Haryono[4]
[1,2,3,4]Pradita University; Scientia Business Park, Jl. Gading Serpong Boulevard No.1, Curug Sangereng, Kec. Klp. Dua, Kabupaten Tangerang, Banten 15810, 02155689999
[1]Magister Teknologi Informasi, Pradita University, Banten
e-mail: *[1]**cecilia.angietawinata@gmail.com**, [2]handri.santoso@pradita.ac.id, [3]ito.wasito@pradita.ac.id, [4]haryono@pradita.ac.id

Abstrak

Story generator berperan penting bagi penulis cerita dalam membantu menghasilkan ide-ide maupun konsep awal cerita. Penggunaan Keras Tokenizer serta model word embedding membutuhkan kecepatan pelatihan model yang relatif lambat untuk menjalankan ratusan iterasi pelatihan. Pada penelitian ini, kami usulkan metode perancangan word-level story generator bahasa Indonesia menggunakan model Markov Chain dengan Bidirectional GRU yang mampu menghasilkan kualitas teks serupa dengan keluaran model word embedding, namun memiliki kecepatan pelatihan model yang lebih cepat. Performa model Markov Chain-BiGRU akan dibandingkan dengan performa model BiGRU (word-level) dan model GRU (character-level). Evaluasi model tahap pertama dilakukan dengan cara membandingkan nilai loss dan kecepatan pelatihan setiap model; evaluasi model tahap kedua dilakukan dengan cara survei kepada 33 orang assessors; sedangkan evaluasi model tahap ketiga dilakukan dengan cara membandingkan performa model dengan model penelitian serupa. Story generator bahasa Indonesia ini mampu meningkatkan kecepatan pelatihan model sebesar 66,38% dari model word embedding penelitian serupa, serta menghasilkan kualitas teks yang lebih baik dibandingkan dengan keluaran model konvensional neural-based maupun word embedding.

Kata kunci*— Story Generator, Keras Sequential Model, Markov Chain, Bidirectional GRU*

Abstract

Story generator plays an important role to help story writers generate story ideas, even initial concepts. Usage of Keras Tokenizer as well as word embedding model requires relatively slower model training speed in order to execute hundreds of training iterations. In this research, we propose design method to create a word-level Indonesian language story generator by implementing Markov Chain model and Bidirectional GRU, which is able to generate quality text as good as the outputs of word embedding models, while having faster model training speed. The performance of Markov Chain-BiGRU model was compared with the performance of word-level BiGRU model and character-level GRU model. The first stage of model evaluation was done by comparing each model's loss value and model training speed; the second stage was done by giving survey to 33 assessors; while the third stage was done by comparing model's performance with model from related work. The proposed Indonesian story generator succeeded on increasing the model training speed by 66.38% from related work's model, as well as producing better-quality text compared to outputs from conventional neural-based and word embedding models.

Keywords*— Story Generator, Keras Sequential Model, Markov Chain, Bidirectional GRU*

1. INTRODUCTION

The process of story writing requires a huge amount of time. Hence an alternative media is needed in order to speed up story writing process, as well as to maintain the originality of the story. Automated story generator is able to generate stories automatically by combining psychology and artificial intelligence (AI). Story generator chooses event or action sequences that fits certain criteria which can be considered as a story. Each story has setting, interactive characters, objects, even a moral message that the writers intend to convey [1]. Story writers can create a full-length story automatically by using story generator in only minutes or seconds.

Markov Chain model is stochastic, which is able to predict the next state based on relevant current state; therefore, the model is good for modeling short-term dependencies in the case of story or text generation [2]. Since the model has simplified processing, Markov Chain model is able to increase computational efficiency because it does not require backpropagation or gradient update. On the other side, neural language model which uses deep neural network differs from Markov Chain model. Neural language model is deterministic, has a loop in hidden layer to store information from the previous process to be used in value prediction in the current process. Recurrent Neural Network (RNN) architecture has a cyclic connection to update the current state based on the previous state and current inputs [3][4]. Several RNN variations, including bidirectional GRU (BiGRU), can capture long-term dependencies in the form of sequences and are proven to show well performances in sequence modeling [5].

Therefore, we propose a method to create word-level Indonesian story generator using Markov Chain model and BiGRU that can capture both short-term and long-term dependencies, resulting in more consistent and coherent text outputs, while having faster model training speed than word embedding models due to the stochastic Markov Chain model. We compare the performance of Markov Chain-BiGRU model with conventional word-level BiGRU model, conventional character-level GRU model, and word embedding model from related work. We develop evaluation metrics based on descriptive statistics in order to assess different aspects of stories for human evaluation.

2. METHODOLOGY

2.1 Literature Review

The story generator proposed in this research combines Markov Chain model and Bidirectional Gated Recurrent Unit (BiGRU) model. Model evaluation will be done using two main theories: story intrinsic elements and Likert scale. Each is explained as follows;

2.1.1 Markov Chain Model

According to Parsing, Markov Chain model structures on how random variable could change from one state to the next state [2]. Each variable has probability related to the next possible state, then Markov Chain model is in charge of choosing one of the next states based on its probability. This trait explains the stochastic side of Markov Chain because the choices are not always the same and the process has random element. Markov Chain predicts the next state based on the relevant current state, which is able to simplify the model but at the same time it loses beneficial information from the past [6].

Since Markov Chain predicts the next state based solely on the relevant current state, the model is considered good in modeling short-term dependencies in the case of text generation. By sampling from various next token candidates, Markov Chain model can produce text with more variety. The simplification of Markov Chain process also increases computational efficiency due to it not requiring backpropagation or gradient update.

2.1.2 Bidirectional Gated Recurrent Unit (BiGRU) Model

Gated recurrent unit (GRU) is a variation of RNN with gate structure to facilitate information flow in and between cells, while having less parameter and no output gate. GRU uses reset gate which is related to the hidden state, and update gate which is a combination of input and forget gate in Long Short-Term Memory (LSTM) architecture. Reset gate indicates how past information which has been stored could influence new inputs, whereas update gate indicates how past information flows to the future. These gates help to solve vanishing gradient problems found in RNN models [7][8]. GRU is designed for capturing long-term dependencies and able to show well performance in handling sequential data just like LSTM. But GRU train smaller dataset faster and easier than LSTM due to maintaining hidden state instead of cell state. Several experiments have proven that GRU converges faster and gives relatively better performance than LSTM [5].

States found in unidirectional GRU are transferred forward only so that it is easy to ignore the influence of beneficial next words. On the other hand, bidirectional GRU (BiGRU) differs from unidirectional GRU by transferring states forward and backward simultaneously. This results in the ability of GRU units to take beneficial next words influence into account and produce more accurate outputs [9].

2.1.3 Story Intrinsic Elements

Story intrinsic elements are fictional elements that build the literature itself as a whole essay [10]. Story intrinsic (and extrinsic) elements contains educational and entertaining values. Readers can grasp the moral message from story intrinsic elements and turning it into character building tool in everyday life. Each story intrinsic element is explained as follows [10]:

a. Theme is the main idea that connects structure, problem, and events in a short story.
b. Character is the actors that is related to the story and play an important role in conveying moral message. Characteristics are certain character traits found in character actions or explicitly written by story writers.
c. Plot is the order of events that creates the story whole.
d. Setting which consist of location, time, and mood/ambience.
e. Point of view (POV) is the writer's perspective position while telling the story.
f. Writing style is type of language or accent used in the story.
g. Moral message is a positive message that can be taken from the story.

2.1.4 Likert Scale

Likert scale consists of a rating scale where respondents can specify their approval to a certain condition by order or ranking. Likert scale results are equal to numerical values. Researches often use Likert scale in odd values (three, five, seven, or nine) to observe neutral responses and increase chances to produce slightly higher mean value in regards to the possible maximum value [11].

2.2 Related Works

Mustofa conducted an experiment on Indonesian story generator using Skip-Thoughts and GRU. The model consists of one encoder with one GRU layer (500 hidden units) and two decoders, each with one GRU layer (500 hidden units). The model was trained with several datasets: Indonesian folklore, slice-of-life, and romance short stories. Experiment shows that the model produced well-quality short stories, with the subject-predicate-object-description (*subjek-predikat-objek-keterangan*, S-P-O-K) and character criteria are evaluated as "good"; the cohesion, whole of the story, theme, plot, and settings criteria are evaluated as "fair" [12].

Pawade et al. conducted an experiment on story scrambler using word-level RNN-LSTM. The model is designed for generating new stories based on collection of short stories as input. Output of the model was evaluated subjectively by human based on the criteria: grammar, events linkage, interest level, and uniqueness. Model's accuracy is 63% [13].

In the other hand, Fu et al. conducted an experiment on repetition problem found in text generation. The experiment was conducted on a Markov Chain model with the implementation of Average Repetition Probability (ARP). Fu et al. proposed theory that explains why repetition problem could be found in text generation, and the answer is due to so many high inflow words from human language itself which causes the words to go back to themselves and increases repetition probability [6].

2.3 Research Methods

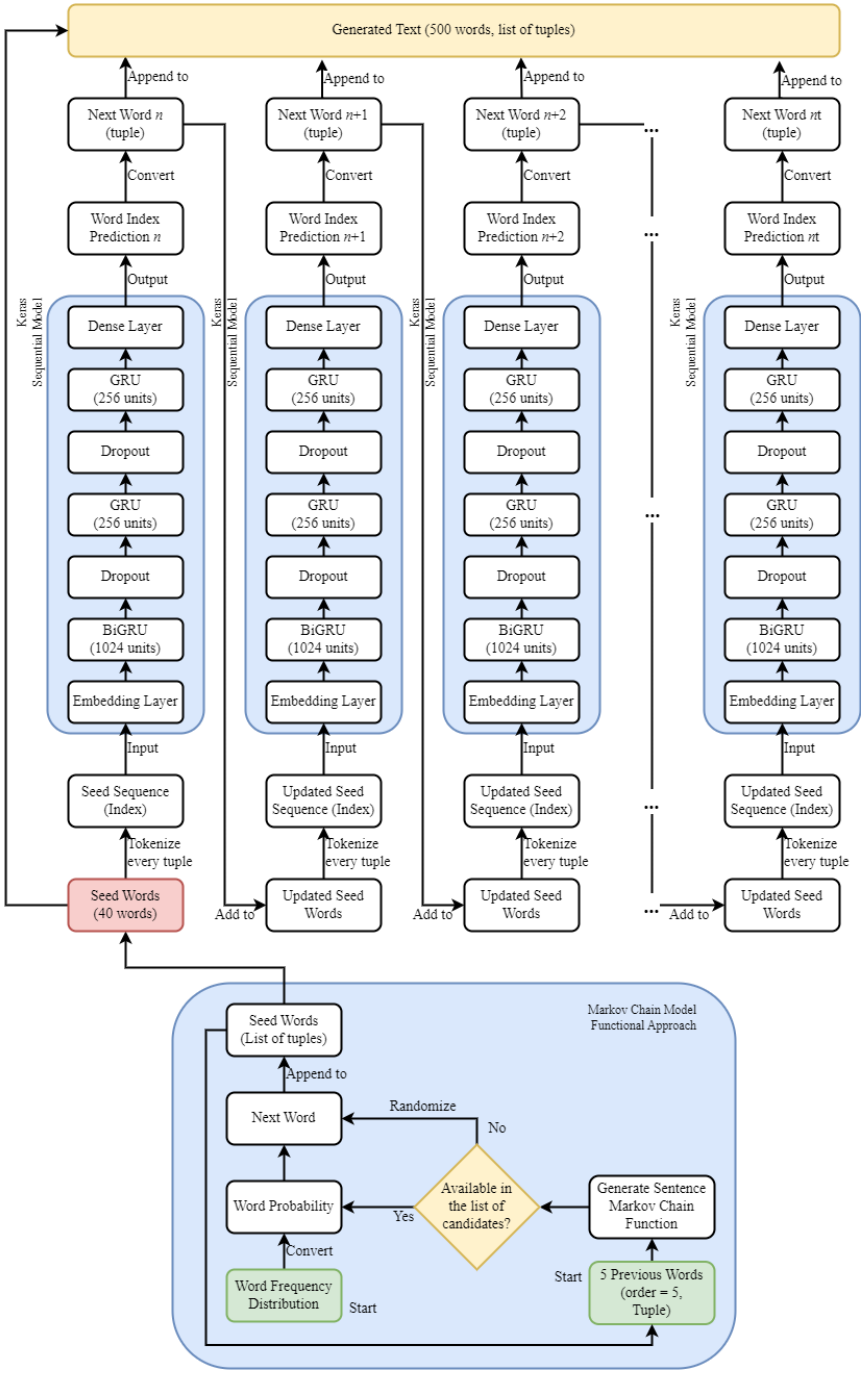

Figure 1. The architecture of Markov Chain model and sequential model with BiGRU

As illustrated in Figure 1, Markov Chain model uses functional approach and order of 5 instead of Markov Chain class due to the complexity of Markov Chain class which requires more computational power. The process begins with determining frequency distribution from each word followed by other word. Then, the frequency is converted into probability between words. The function generate_sentence_markov_chain is defined to take 5 previous words as input, observe the list of next word candidates, and take the next word based on probability between the words. If the previous words are not included in the list of next word candidates, then the model would take one word from the list in random. Those words are joined together to create the seed words. In the same time, 5 first words from the seed words are taken as input and the process is repeated until the seed words consist of 40 words. The final seed words are used as input to BiGRU model in order to predict the next words.

BiGRU model uses sequential method from Keras that consists of seven layers. Input data flows through Embedding layer, BiGRU layer (1024 RNN units), dropout layer 1, GRU layer 1 (256 RNN units), dropout layer 2, GRU layer 2 (256 RNN units), and Dense layer. Seed words produced by Markov Chain model are tokenized into indexes and flows into Embedding layer as seed sequence. In this step, the sequential model processes the seed sequence. Output of the process is a predicted index n, which will be converted into next word n. The next word n is appended to seed words, creating generated text. Simultaneously, the next word n is added to the seed words and tokenized into updated seed sequence (in indexes). Updated seed sequence becomes the input for next process until the model produces predicted index n+1, which will be converted into next word n+1. The next word n+1 is appended to generated text. Simultaneously, the next word n+1 is added to the updated seed words and tokenized into updated seed sequence (in indexes). The loop goes on until generated text consists of 500 words. This Markov Chain-BiGRU model uses Adam optimizer with 0.0005 learning rate and categorical crossentropy loss function. The first model will be trained on folklore dataset with 100 epochs, while the second model will be trained on fairytales dataset with 210 epochs.

3. RESULTS AND DISCUSSIONS

3.1 Model Training

The result of every model training with two types of dataset (Indonesian folklore and fairytales) is shown in Table 1.

Table 1. Comparison of each model's training duration and lowest loss value

Model	**Dataset**	**Words**	**Epochs**	**Training Duration**	**Lowest Loss Value**
Markov Chain-BiGRU	Folklore	97,260	100	70 minutes	0.3374
Markov Chain-BiGRU	Fairytales	113,956	210	175 minutes	0.2694
Conventional Word-level BiGRU	Folklore	97,260	100	6 minutes 40 seconds	0.0097
Conventional Word-level BiGRU	Fairytales	170,471	100	11 minutes 40 seconds	0.0049
Conventional Character-level GRU	Folklore	97,260	200	13 minutes 20 seconds	0.4934
Conventional Character-level GRU	Fairytales	170,471	100	11 minutes 40 seconds	0.7681

3.2 Model Evaluation

Model evaluation was split into three stages. The first stage was done by comparing each model's loss value and model training speed. The second stage was done by giving survey to 33 respondents (later is called as assessors) to evaluate the quality of generated story. The third stage was done by comparing loss value, model training speed, and quality of the stories generated by Markov Chain-BiGRU model, to the Indonesian story generator with Skip-Thoughts and GRU model from related work [12].

3.2.1 The First Stage

The model training speed of Markov Chain-BiGRU model while processing folklore dataset decreased by 2,017.65% from the conventional word-level BiGRU model (about 20 times slower) and decreased by 414.29% from the conventional character-level GRU model (about 4 times slower). The model training speed of Markov Chain-BiGRU model while processing fairytales dataset decreased by 973.53% from both the conventional word-level BiGRU model and the conventional character-level GRU model (about 10 times slower). This result shows that Markov Chain-BiGRU model training speed highly decreased from the conventional neural-based models.

Comparing the mean of loss value and the mean of model training speed between models which were trained on folklore and fairytales dataset, as seen in Table 2, it is concluded that conventional word-level BiGRU model has the lowest loss value, followed by Markov Chain-BiGRU model and conventional character-level GRU model. The conventional word-level BiGRU model also has the fastest model training speed, followed by conventional character-level GRU model and Markov Chain-BiGRU model. Based on this result, the conventional word-level BiGRU model was determined as the most technically optimal model.

Table 2. Comparison of each model's mean of loss value and mean of model training speed

Model	Mean of Loss Value	Mean of Model Training Speed
Markov Chain-BiGRU	0.3034	0.00000725 minutes/word/epoch
Conventional Word-level BiGRU	0.0073	0.00000051 minutes/word/epoch
Conventional Character-level GRU	0.63075	0.00000104 minutes/word/epoch

3.2.2 The Second Stage

As for Markov Chain-BiGRU model with folklore dataset, as seen in Table 3, mean value of the theme, characteristics, location, mood, and point of view criteria are bigger than the median, so the evaluation of these 5 criteria tend to be "good" (between value "3" and "4"). This is supported by high occurring frequency of value "3" and "4". Mean value of the plot, grammar, cohesion and coherence, and moral message criteria are smaller than the median, so the evaluation of these 4 criteria tend to be "poor" (between value "2" and "3"). Mean value of the location criterion is smaller than the median, so the evaluation of this criterion tends to be "fair" (between value "3" and "4"). While mean value of the characters criterion is equal to the median, so the evaluation is neutral or "fair" (value "3"). High range value shows that the assessors have various opinions, but low interquartile range shows that data distribution in the middle half is dominated by value "3" and "4". Majority of the assessors are quite in agreement with the evaluation. In conclusion, the quality of folklore short stories generated by Markov Chain-BiGRU model is considered as "fair".

Table 3. Descriptive statistics calculation of folklore short stories generated by Markov Chain-BiGRU model

Markov Chain-BiGRU Model with Folklore Dataset						
Criteria	**Mean ($\bar{x}$)**	**Quartiles (Q_1, Q_2, Q_3)**	**Mode**	**Range**	**Standard Deviation**	**Interquartile Range**

Theme	3.09	3,3,4	3	4	0.91	1
Characters	3	2,3,4	3	4	1.03	2
Characteristics	3.03	2,3,4	3	4	1.02	2
Plot	2.94	2,3,4	3	4	1.09	2
Location	3.18	3,3,4	4	4	0.98	1
Time	3.33	3,4,4	4	4	0.96	1
Mood	3.24	3,3,4	4	4	1.06	1
Point of view	3.24	3,3,4	4	4	1.06	1
Grammar	2.94	3,3,4	3	3	0.97	1
Cohesion and coherence	2.88	2,3,4	3	3	1.05	2
Moral message	2.82	2,3,3	3	4	1.04	1

As for conventional word-level BiGRU model with folklore dataset, as seen in Table 4, mean value of all criteria are bigger than the median, so the evaluation of all criteria tends to be "poor" (between value "1" and "2"). High range value shows that the assessors have various opinions, but low interquartile range shows that data distribution in the middle half is dominated by value "1" and "2". Majority of the assessors are quite in agreement with the evaluation. Although, median value of "1" and high occurring frequency of value "1" shows that the quality of folklore short stories generated by conventional word-level BiGRU model can be considered as "very poor".

Table 4. Descriptive statistics calculation of folklore short stories generated by conventional word-level BiGRU model

Conventional Word-level BiGRU Model with Folklore Dataset						
Criteria	**Mean ($\bar{x}$)**	**Quartiles (Q_1,Q_2,Q_3)**	**Mode**	**Range**	**Standard Deviation**	**Interquartile Range**
Theme	1.52	1,1,2	1	3	0.91	1
Characters	1.64	1,1,2	1	4	0.99	1
Characteristics	1.60	1,1,2	1	4	1.09	1
Plot	1.48	1,1,1	1	3	1	0
Location	1.58	1,1,2	1	3	0.97	1
Time	1.67	1,1,2	1	3	1.02	1
Mood	1.64	1,1,2	1	4	1.14	1
Point of view	1.67	1,1,2	1	4	1.14	1
Grammar	1.48	1,1,1	1	4	1.09	0
Cohesion and coherence	1.55	1,1,1	1	4	1.20	0
Moral message	1.52	1,1,2	1	4	1	1

As for conventional character-level GRU model with folklore dataset, as seen in Table 5, mean value of the characters, characteristics, time, mood, point of view, and grammar criteria are bigger than the median, so the evaluation of these 8 criteria tend to be "fair" (between value "2" and "3"). Mean value of the plot and cohesion and coherence criteria are bigger than the median, so the evaluation of these 2 criteria tend to be "poor" (between value "1" and "2"). Mean value of the theme and location criteria are smaller than the median, so the evaluation of these 2 criteria tend to be "poor" (between value "2" and "3"). Mean value of the moral message criterion is smaller than the median, so the evaluation of this criterion tends to be "very poor" (between value "1" and "2"). High range value shows that the assessors have various opinions, and high interquartile range shows that data distribution in the middle half consist of various values. Majority of assessors are not in agreement with the evaluation. Median value of "2" and high occurring frequency of value "1" shows that the quality of folklore short stories generated by conventional character-level GRU model can be considered as "poor".

Table 5. Descriptive statistics calculation of folklore short stories generated by conventional character-level GRU model

Conventional Character-level GRU Model with Folklore Dataset						
Criteria	**Mean ($\bar{x}$)**	**Quartiles (Q_1,Q_2,Q_3)**	**Mode**	**Range**	**Standard Deviation**	**Interquartile Range**
Theme	2.67	1,3,3	3	4	1.34	2
Characters	2.42	1,2,3	1	4	1.17	2
Characteristics	2.21	1,2,3	1	4	1.24	2
Plot	2	1,1,3	1	4	1.27	2
Location	2.33	1,2,3	1	4	1.34	2
Time	2.64	1,3,4	1	4	1.32	3
Mood	2.36	1,2,3	1	4	1.22	2
Point of view	2.39	1,2,4	1	4	1.29	3
Grammar	2.24	1,2,3	1	4	1.29	2
Cohesion and coherence	1.79	1,1,2	1	4	1.11	1
Moral message	1.97	1,2,3	1	4	1.16	2

As for Markov Chain-BiGRU model with fairytales dataset, as seen in Table 6, mean value of the theme, characters, characteristics, plot, time, mood, point of view, grammar, and moral message criteria are bigger than the median, so the evaluation of these 9 criteria tend to be "good" (between value "3" and "4"). This is supported by high occurring frequency of the value "3" and "4". Mean value of the location criterion is smaller than the median, so the evaluation of this criterion tends to be "fair" (between value "3" and "4"). Mean value of the cohesion and coherence criterion is equal to the median, so the evaluation is neutral or "fair" (value "3"). High range value shows that the assessors have various opinions, but low interquartile range shows that data distribution in the middle half is dominated by the value "3" and "4". Majority of the assessors are quite in agreement with the evaluation. In conclusion, the quality of fairytale short stories generated by Markov Chain-BiGRU model is considered as "good".

Table 6. Descriptive statistics calculation of fairytales short stories generated by Markov Chain-BiGRU model

Markov Chain-BiGRU Model with Fairytales Dataset						
Criteria	**Mean ($\bar{x}$)**	**Quartiles (Q_1,Q_2,Q_3)**	**Mode**	**Range**	**Standard Deviation**	**Interquartile Range**
Theme	3.24	3,3,4	3	4	1.03	1
Characters	3.21	3,3,4	4	4	1.05	1
Characteristics	3.06	3,3,4	3	4	0.99	1
Plot	3.06	2,3,4	4	4	1.14	2
Location	3.18	3,3,4	4	4	1.01	1
Time	3.55	3,4,4	4	4	1.06	1
Mood	3.27	3,3,4	4	4	1	1
Point of view	3.12	3,3,4	3	4	0.99	1
Grammar	3.06	2,3,4	4	4	1.17	2
Cohesion and coherence	3	3,3,4	3	4	1.09	1
Moral message	3.03	3,3,4	3	4	1.07	1

As for conventional word-level BiGRU model with fairytales dataset, as seen in Table 7, mean value of all criteria are bigger than the median, so the evaluation of all criteria tends to be "poor" (between value "1" and "2"). High range value shows that the assessors have various opinions, but low interquartile range shows that data distribution in the middle half is dominated by value "1" and "2". Majority of the assessors are in agreement with the evaluation. Although, median value of "1" and high occurring frequency of value "1" shows that the quality of folklore

short stories generated by conventional word-level BiGRU model can be considered as "very poor".

Table 7. Descriptive statistics calculation of fairytales short stories generated by conventional word-level BiGRU model

Conventional Word-level BiGRU Model with Fairytales Dataset						
Criteria	**Mean ($\bar{x}$)**	**Quartiles (Q_1,Q_2,Q_3)**	**Mode**	**Range**	**Standard Deviation**	**Interquartile Range**
Theme	1.52	1,1,2	1	4	0.94	1
Characters	1.42	1,1,2	1	3	0.79	1
Characteristics	1.39	1,1,1	1	4	0.89	0
Plot	1.24	1,1,1	1	3	0.71	0
Location	1.39	1,1,2	1	4	0.83	1
Time	1.36	1,1,1	1	3	0.74	0
Mood	1.39	1,1,1	1	3	0.79	0
Point of view	1.39	1,1,2	1	3	0.75	1
Grammar	1.39	1,1,2	1	4	0.83	1
Cohesion and coherence	1.27	1,1,1	1	4	0.80	0
Moral message	1.36	1,1,1	1	4	0.86	0

As for conventional character-level GRU model with fairytales dataset, as seen in Table 8, mean value of the characters, characteristics, plot, time, mood, point of view, grammar, cohesion and coherence, and moral message criteria are bigger than the median, so the evaluation of these 9 criteria tend to be "fair" (between value "2" and "3"). Mean value of the location criterion is bigger than the median, so the evaluation of this criterion tends to be "good" (between value "3" and "4"). Mean value of the theme criterion is smaller than the median, so the evaluation tends to be "poor" (between value "2" and "3"). High range value and various mode shows that the assessors have various opinions. High interquartile range shows that data distribution in the middle half is dominated by various values. Majority of the assessors are not in agreement with the evaluation. In conclusion, the quality of the fairytales short stories generated by conventional character-level GRU model is considered as "fair".

Table 8. Descriptive statistics calculation of fairytales short stories generated by conventional character-level GRU model

Conventional Character-level GRU Model with Fairytales Dataset						
Criteria	**Mean ($\bar{x}$)**	**Quartiles (Q_1,Q_2,Q_3)**	**Mode**	**Range**	**Standard Deviation**	**Interquartile Range**
Theme	2.88	2,3,4	2	4	1.36	2
Characters	2.69	2,2,4	2	4	1.21	2
Characteristics	2.45	2,2,3	2	4	1.12	1
Plot	2.52	2,2,3	2	4	1.23	1
Location	2.52	1,2,4	1	4	1.25	3
Time	3.03	2,3,4	4	4	1.24	2
Mood	2.67	2,2,4	4	4	1.31	2
Point of view	2.52	1,2,4	1	4	1.28	3
Grammar	2.42	2,2,3	2	4	1.19	1
Cohesion and coherence	2.15	1,2,3	1	4	1.18	2
Moral message	2.18	1,2,3	1	4	1.33	2

3.2.3 The Third Stage

In order to make the comparison equivalent, the performance of Markov Chain-BiGRU model and Skip-Thoughts with GRU model [12] are compared solely while processing Indonesian folklore datasets. We used Model I from the Skip-Thoughts with GRU model to compare with our Markov Chain-BiGRU model. The comparison data are shown in Table 9.

Table 9. Comparison between Markov Chain-BiGRU model and Skip-Thoughts with GRU model, while processing similar folklore dataset [10]

Model	Sentences	Epochs	Training Duration	Mean of Loss Value	Mean of Model Training Speed
Markov Chain-BiGRU	8,938	100	70 minutes	0.3374	0.000078 minutes /sentence/epoch
Skip-Thoughts with GRU	3,872	79	71 minutes	*Pre-loss:* 0.0037 *Post-loss:* 0.0033	0,000232 minutes /sentence/epoch

Based on the comparison data, Markov Chain-BiGRU model has higher loss value than both pre-loss and post-loss value from Skip-Thoughts with GRU model. But in the term of model training speed, Markov Chain-BiGRU model has faster model training speed than Skip-Thoughts with GRU model. There is an increase in model training speed of 66.38% while using Markov Chain-BiGRU model. This result shows that the Markov Chain-BiGRU model succeeded on increasing the model training speed compared to word embedding model, such as Skip-Thoughts with GRU model.

3.2.4 Analysis

Two comparisons as seen in Table 10 and Table 11 show that the quality of short stories generated by Markov Chain-BiGRU model is better than the conventional word-level BiGRU model and conventional character-level GRU model. Although the conventional word-level BiGRU model has the lowest loss value and the fastest model training speed, in the end its generated stories are very poor in quality. The usage of larger dataset, such as fairytales dataset which has larger size compared to folklore dataset, does not cause degradation to the performance of Markov Chain-BiGRU model. Hence, it can be concluded that Markov Chain-BiGRU model is the most optimal story generator to generate Indonesian short stories, with loss values lower than the conventional character-level GRU model, although having the slowest model training speed compared to both conventional neural-based models.

Table 10. Comparison of folklore short stories quality interpretation of the three models

Criteria	Markov Chain-BiGRU	Conventional Word-level BiGRU	Conventional Character-level GRU
Theme	Good	Very poor	Poor
Characters	Fair	Very poor	Fair
Characteristics	Good	Very poor	Fair
Plot	Poor	Very poor	Poor
Location	Good	Very poor	Fair
Time	Fair	Very poor	Poor
Mood	Good	Very poor	Fair
Point of view	Good	Very poor	Fair
Grammar	Poor	Very poor	Fair
Cohesion and coherence	Poor	Very poor	Poor
Moral message	Poor	Very poor	Very poor

Table 11. Comparison of fairytales short stories quality interpretation of the three models

Criteria	Markov Chain-BiGRU	Conventional Word-level BiGRU	Conventional Character-level GRU
Theme	Good	Very poor	Poor
Characters	Good	Very poor	Fair
Characteristics	Good	Very poor	Fair
Plot	Good	Very poor	Fair
Location	Good	Very poor	Fair
Time	Fair	Very poor	Good
Mood	Good	Very poor	Fair
Point of view	Good	Very poor	Fair
Grammar	Good	Very poor	Fair
Cohesion and coherence	Fair	Very poor	Fair
Moral message	Good	Very poor	Fair

Based on the comparison as shown in Table 12, the quality of short stories generated by Markov Chain-BiGRU does not differ much from the quality of short stories generated by Skip-Thoughts with GRU model. But Markov Chain-BiGRU model excels in 2 criteria: theme and settings (overall from location, time, and mood). Other than the criteria used to evaluate Skip-Thoughts with GRU model, the Markov Chain-BiGRU model is also evaluated as "good" in the characteristics and point of view criteria.

Table 12. Comparison of folklore short stories quality interpretation between Markov Chain-BiGRU model and Skip-Thoughts with GRU model [10]

Criteria	Markov Chain-BiGRU	Skip-Thoughts with GRU
Theme	Good	Fair
Characters	Fair	Fair
Characteristics	Good	- (No information)
Plot	Poor	Poor
Location	Good	Fair
Time	Fair	
Mood	Good	
Point of view	Good	- (No information)
Grammar	Poor	Good
Cohesion and coherence	Poor	Poor
Moral message	Poor	- (No information)

By increasing the model training speed to 66.38% faster than word embedding model, as well as generating better quality stories, it is concluded that the Markov Chain-BiGRU model can produce equivalent, even better, quality stories than conventional neural-based and word embedding models, while having faster model training speed.

4. CONCLUSION

In this paper, we propose a method to create word-level Indonesian story generator using Markov Chain model and BiGRU that can capture both short-term and long-term dependencies, resulting in more consistent and coherent text outputs, while having faster model training speed than word embedding models due to the stochastic Markov Chain model. Experiment shows that our Markov Chain-BiGRU model has the second lowest loss value, with the conventional word-level BiGRU model in the first place and the conventional character-level GRU model in the third place. Markov Chain-BiGRU model has the slowest model training speed compared to both conventional neural-based models, but it increases the model training speed to 66.38% faster than

word embedding model. Overall, the Markov Chain-BiGRU model produces better quality stories than conventional neural-based and word embedding models.

5. SUGGESTION

Due to the main purpose of this paper, which is to increase the model training speed and producing better quality texts, several processes are simplified and limited to some extent. In the future, the proposed Markov Chain-BiGRU model can be explored and developed further into a more complex model which is able to produce even better-quality texts. Various methods can be implemented, such as stop words, word lemmatization, Keras Tokenizer, hyper tuning the model, implementing Average Repetition Probability (ARP) to solve the repetition problem, using other algorithms like GAN, or using word embedding and pretrained models like Transformers from Hugging Face.

REFERENCES

[1] Alhussain, A. I., & Azmi, A. M., 2021, Automatic Story Generation: A Survey of Approaches, *ACM Computing Surveys*, *54*(5), https://doi.org/10.1145/3453156.

[2] Meek, C., 2019, *Grammar Introduction into Markov Chain Text Generation*.

[3] Schmidt, R. M., 2019, *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*, *1*, 1–16, http://arxiv.org/abs/1912.05911

[4] Yu, Y., Si, X., Hu, C., & Zhang, J, 2019, A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures, *Neural Computation*, *31*, 1235–1270, https://doi.org/10.1162/NECO

[5] He, Y., Chen, R., Li, X., Hao, C., Liu, S., Zhang, G., & Jiang, B, 2020, Online at-risk student identification using RNN-GRU joint neural networks, *Information (Switzerland)*, *11*(10), 1–11, https://doi.org/10.3390/info11100474.

[6] Fu, Z., Lam, W., So, A. M. C., & Shi, B., 2021, A Theoretical Analysis of the Repetition Problem in Text Generation, *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, *14B*, 12848–12856, https://doi.org/10.1609/aaai.v35i14.17520.

[7] Raza, M. R., Hussain, W., & Merigó, J. M., 2021, Cloud Sentiment Accuracy Comparison using RNN, LSTM and GRU, *Innovations in Intelligent Systems and Applications Conference (ASYU)*, 1–5, https://doi.org/10.1109/ASYU52992.2021.9599044.Cloud.

[8] Shewalkar, A. N., 2018, *Comparison of RNN, LSTM and GRU on Speech Recognition Data*, https://doi.org/10.4324/9781315721606-101.

[9] Cheng, Y., Yao, L., Xiang, G., Zhang, G., Tang, T., & Zhong, L., 2020, Text Sentiment Orientation Analysis Based on Multi-Channel CNN and Bidirectional GRU with Attention Mechanism, *IEEE Access*, *8*, 134964–134975, https://doi.org/10.1109/ACCESS.2020.3005823.

[10] Pramidana, I. D. G. A. I., 2020, Unsur Intrinsik dan Ekstrinsik Dalam Cerpen "Buut" Karya I Gusti Ayu Putu Mahindu Dewi Purbarini, *Jurnal Pendidikan Bahasa Bali Undiksha*, *7*(2), 61, https://doi.org/10.23887/jpbb.v7i2.28067.

[11] Pimentel, J. L., 2019, Some Biases in Likert Scaling Usage and its Correction, *International Journal of Sciences: Basic and Applied Research*, *45*(1), 183–191, http://gssrr.org/index.php?journal=JournalOfBasicAndApplied.

[12] Mustofa, 2022, *Story Generator Bahasa Indonesia Menggunakan Skip-Thoughts*, https://dspace.uii.ac.id/handle/123456789/39883%0Ahttps://dspace.uii.ac.id/bitstream/handle/123456789/39883/18917123.pdf?sequence=1.

[13] Pawade, D., Sakhapara, A., Jain, M., Jain, N., & Gada, K., 2018, Story Scrambler - Automatic Text Generation Using Word Level RNN-LSTM, *International Journal of Information Technology and Computer Science*, *10*(6), 44–53, https://doi.org/10.5815/ijitcs.2018.06.05.