computing@computingonline.net
www.computingonline.net

Print ISSN 1727-6209
On-line ISSN 2312-5381
International Journal of Computing

SEQUENCE-TO-SEQUENCE LEARNING FOR MOTION-AWARE CLAIM GENERATION

Derwin Suhartono [1)], Aryo Pradipta Gema [1)], Suhendro Winton [1)], Theodorus David [1)], Mohamad Ivan Fanany [2)], Aniati Murni Arymurthy [2)]

[1)] Computer Science Department, School of Computer Science, Bina Nusantara University,
Jl. K. H. Syahdan No. 9 Kemanggisan, Jakarta 11480, Indonesia
dsuhartono@binus.edu; {aryo.gema, suhendro.winton, theodorus.david}@binus.ac.id
[2)] Machine Learning and Computer Vision Laboratory,
Faculty of Computer Science,
Universitas Indonesia,
Depok 16424, Indonesia
{ivan, aniati}@cs.ui.ac.id

Paper history:
Received 12 May 2020
Received in revised form 26 August 2020
Accepted 17 October 2020
Available online 30 December 2020

Keywords:
seq2seq;
argumentation mining;
deep learning;
negative log likelihood;
BLEU.

Abstract: The goal of this research is to generate a motion-aware claim using a deep neural network approach: sequence-to-sequence learning method. A motion-aware claim is a sentence that is logically correlated to the motion while preserving its grammatical structure. Our proposed model generates a motion-aware claim in a form of one sentence and takes motion as the input also in a form of one sentence. We use a publicly available argumentation mining dataset that contains annotated motion and claim data. In this research, we propose a novel approach for argument generation by employing a scheduled sampling strategy to make the model converge faster. The BLEU scores and questionnaire are used to quantitatively assess the model. Our best model achieves 0.175 ± 0.088 BLEU-4 score. Based on the questionnaire results, we can also derive a conclusion that it is hard for the respondents to differentiate between the human-made and the model-generated arguments.

1. INTRODUCTION

Argumentation is a quintessential element in our interdependent life. It is almost impossible for individuals to avoid argument or debate with other individuals. Conveying a multitude of thoughts in a logically ordered manner is the essence of argumentation. Properly formed arguments can be a potent instrument to influence the listeners [1]. However, research in the computational argument has just recently dawned, despite the commonality of argumentation in our daily life. Fortunately, computational linguistic researchers already exhibited their interests in this research domain [2], for instance, research on claim detection [3], evidence detection [4], and stance classification [5]. Examples of research that attempt to analyze the qualitative behavior of arguments are recognizing insufficiently supported arguments using Hierarchical Attention Networks (HANs) and XGBoosts [6] and recognizing argumentative relations using Siamese Networks [7].

Despite much proliferation in the argumentation recognition research, the argument generation is still not well developed. To the best of our knowledge, the only study that focused on topic-dependent argument generation is done in [8]. The paper attempted to generate argument using a sentence retrieval method. Sentences that are retrieved will then be ordered using two machine learning tasks, claim sentence selection and supporting sentences ordering respectively. This system consists of a complex processing pipeline, and as a result, this system needs to address three different problems: (1) identification errors, (2) polarity errors, (3) motion format limitation. The method may also incorrectly identify the relevant keywords of the topic/motion, stance, and variety of debate topic. We believe that a

different approach needs to be taken to produce an argument close to a human-made sentence.

Deep learning is an appealing and natural answer to tackle these problems. Deep learning algorithms have been proven successful in many natural language modelling tasks [9, 10, 11]. However, generating a meaningful and coherent sentence is still a challenging task for deep learning. The fundamental objective of producing a sentence is to determine a distribution over sentences in a training corpus and use it to sample naturally formed sentences [12]. This will potentially allow a generation of distinct sentences which still preserve semantic and syntactic properties of natural sentences.

All these hidden potentials that can be achieved and the lack of argumentation dataset, in general, motivate us to pursue this argument generation research. We believe that the advances in argument generation field would greatly benefit researchers by providing more available datasets for argument recognition, analysis, model development, and other general argument mining tasks. This would in return, makes it easier to understand the creative process of human thoughts mathematically.

General sentences generation study has been vastly conducted in the computational language community. A prior approach attempted to use Recurrent Neural Network (RNN)-based encoder-decoder (autoencoder) framework [13], for sentence generation [14]. This framework will learn to represent sentences into their latent representations in encoder block; then the decoder block will attempt to generate synthetic sentences from these latent representations.

One immense challenge of generating realistic sentences is due to the nature of RNNs. RNNs generate words from previously generated words sequentially and do not take the ground truth words into account. As a result, error accumulates proportional to the length of the sentence. This is a problem known as exposure bias [15]. When deciding the next token in the training stage, [16] a training strategy is proposed which is called scheduled strategy (SS), where the synthetic model data is partially fed to the model as a prefix rather than the true data. This training strategy helps the model to converge faster.

In this paper, we propose the use of Sequence-to-Sequence framework (one of deep neural network approaches) with a scheduled strategy to generate realistic-looking claims based on given debate topics via GAN. Specifically, the Gated Recurrent Unit (GRU)-based [17], Sequence-to-Sequence framework [18] is used. We implement a Negative Log Likelihood (NLL) loss function to train the model. This strategy forces the model to extract features and generates new sentences that are grammatically correct and related to the topic.

2. RELATED WORKS

Our research is tightly related with argumentation mining, Natural Language Generation (NLG) and adversarial training. In argumentation mining, there is a plethora of research that focuses on understanding elements of debating, for instance, research on claim detection, evidence detection, and stance classification. However, research in argumentation mining does not only revolve around classification or detection, but some also focus on qualitative assessment problem, for example, predicting convincingness of an argument using Bidirectional LSTM [19] and assessing the sufficiency of an argument using CNN [20] and HANs and XGBoosts [6].

At the same time, NLG gains a lot of research interest. Most of the works in natural language generation focus on particular task domains, such as poetry generation [21], and jokes generation [22]. Three requirements that must be fulfilled were proposed to define a text as a poem [21]. First, a poem should be grammatically correct. Second, the poem must be semantically understandable. Third, the poem should exhibit poeticness. Derivation tree of natural language into its syntax is just an example of many NLP techniques used by them. The research on jokes generation uses keywords, templates, and rules [22].

Defining the set of rules for keywords extraction and text generation is a painstaking process. Due to this impracticality of manual rules setting, deep learning was introduced as a more favorable approach in NLG. Deep learning algorithms statistically learn to create and adjust their own set of rules to holistically extract the input feature and generate new linguistic data. One of the first approaches to NLG using deep learning introduced Recurrent Neural Network (RNN) based on encoder-decoder (autoencoder) framework [13, 14]. This framework learns to represent sentences into their latent representations in encoder block; then the decoder block attempts to generate synthetic sentences from these latent representations. Since this approach relies heavily on encoded latent representations of sentences from a corpus, this approach often fails to generate realistic sentences from random latent representations [23, 12].

Deep generative model is making a significant leap of progress lately, where [24] tries to provide a view on some set of broad generative methods. There is Adversarial Domain Adaptation (ADA) [25] which aims to transfer prediction knowledge by learning domain-invariant features from a source

domain with labeled data to a target domain without labels. The ADA trains its discriminator to adversarially distinguish between the two domains to achieve the domain invariance of features. Another method of generative model is the Variational Autoencoders (VAEs). VAEs consist of encoder and generator networks which encode data to a latent representation and generate samples from latent space. The model is trained by maximizing a variational lower bound on the log-likelihood of the data [26].

Even though the task of a broad generative model is gaining significant progress, argumentation generation task with a deep generative model is still under-developed. Besides our current work, the only work we found on this task has been done in [8]. They used a large text data from Gigaword corpus [27] and annotated it on the pre-processing step. The work will first accept input of sentence and the stance of the generated argument. The keywords on the sentences will be detected, and then the system will retrieve the sentences from the corpus with the same keywords. However, we could not find other work on argumentation generation which uses generative model. Hence, our work can further the advancement of argumentation generation with a generative model.

3. PROPOSED METHODS

3.1 DATASET

In this research, we used a publicly available dataset provided by IBM Haifa, Israel [4]. This dataset contains 2294 labelled claims and 4690 labelled evidence for 58 different motions. Therefore, on average one motion is connected to 39 claims as can be seen in Fig. 1.

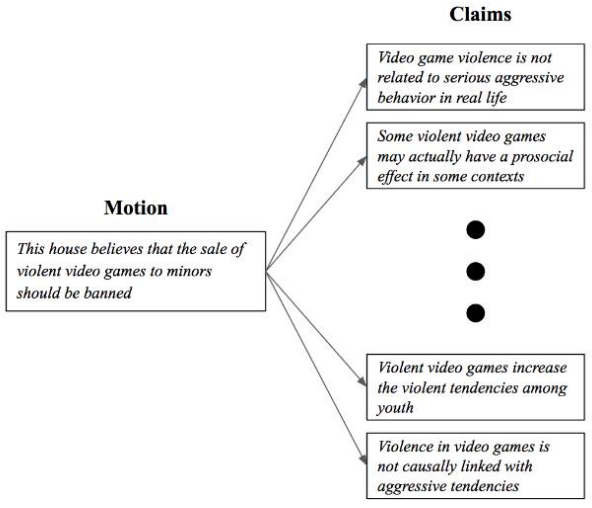

Figure 1 – One-to-many relations between claim and motions

3.2 DATA PREPROCESSING

The raw data is transformed into data that is ready to be trained on. We firstly map claim and motion in a one-to-one form. We use a simple brute force algorithm to do this. We read through the file containing the claims and motions, and temporarily save the combination as a key-value pair. The mapping is created under 1 minute for 2294 data.

Claims and motions come in a form of array of words. However, computation cannot be done in string. Therefore, we need to mask each word using computable variable, in this case, float. In order to achieve that, we firstly tokenize all claims and motions word-by-word and store them in an array. In total, there are 4156 words and 4 additional special characters: "?", ".", "-", and "<SOS>". "<SOS>" symbolize the start of a sentence. 4160 tokens will then be saved in a python dictionary variable along with their indices as the dictionary key. "<SOS>" character will be put in front of every claim sentence. We transform every claim and motion words into their corresponding indices. Consequently, every claim and motion are now stored in a form of an array of indices (Fig. 2). The raw form of claims and motions (the string format of claims and motions) are also saved in the same dictionary variable for further monitoring purposes.

Figure 2 – Example of Representation Masking

3.3 AUTOENCODERS

Autoencoder is trained to encode the input in some representation so that the input can be reconstructed from that representation as the output [28]. The network is a combination of two blocks: an encoder block and a decoder block.

Autoencoder has a hidden layer h that describes the feature representation of the input. Mathematically, autoencoder can be seen as:

$$\textit{Encoder function: } h = f(x) \quad (1)$$

$$\textit{Decoder function: } r = g(h) \quad (2)$$

Autoencoder must learn to copy only approximately because it often learns useful properties of the data when it is forced to prioritize which aspects of the input should be copied [28]. In short, autoencoder should not be too specific and general for it to produce useful features out of input data. Autoencoder has become the first milestone of many generative models such as Generative Adversarial Network (GAN). In this research, both

the encoder and decoder used Vanilla GRU (Gated Recurrent Unit) with 1024 units. We also implemented scheduled sampling as a strategy to help the Sequence-to-Sequence architecture training, especially the decoder part. The calculation is as follows:

$$\log P(Y|X) = \log P(y_1^T|X) = \sum_{t=1}^{T} \log P(y_t|y_1^{t-1}, X)$$

$$\log P(y_t|y_1^{t-1}, X; \theta) = \log P(y_t|h_t; \theta)$$

$$h_t = \begin{matrix} f(X;\theta) \\ f(h_{t-1}, y_{t-1}; \theta)_{otherwise} \end{matrix} \quad if\ t=1,$$

Scheduled sampling itself is a training strategy for RNN to handle the expected output from a prior time step randomly as an input. Scheduled sampling randomly replaces the generated word with the expected output as the input for the next timestep to prevent the exposure bias problem. An example for this is, given the following input sequence: "<SOS> This is a nice day."

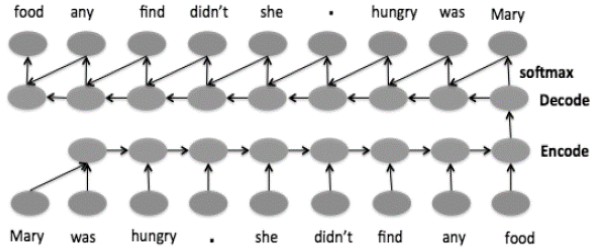

Figure 3 – Visualization of Autoencoder Network to Reconstruct a Sentence [29]

Autoencoder must learn to copy only approximately because it often learns useful properties of the data when it is forced to prioritize which aspects of the input should be copied [28]. In short, autoencoder should not be too specific and general for it to produce useful features out of input data. Visualization of autoencoder network to reconstruct a sentence is presented in Figure 3.

Autoencoder has become the first milestone of many generative models such as Generative Adversarial Network (GAN). In this research, both the encoder and decoder used Vanilla GRU (Gated Recurrent Unit) with 1024 units. We also implemented scheduled sampling as a strategy to help the Sequence-to-Sequence architecture training, especially the decoder part. Scheduled sampling itself is a training strategy for RNN to handle the expected output from a prior time step randomly as an input. Scheduled sampling randomly replaces the generated word with the expected output as the input for the next timestep to prevent the exposure bias problem. An example for this is, given the following input sequence: "<SOS> This is a nice day."

If then the model is fed <SOS>, the model generates a word. In a conventional RNN architecture, the output will always be the input for the next time step. Assuming the generated word is wrong, the next generated word will more likely to be wrong as well. It is what is called as exposure bias. Scheduled sampling will replace the generated output with the expected output randomly as the input for the next timestep. This will help the model to learn the pattern of the output sequence faster.

The encoder block processes a sequence of words by extracting the feature vector of each word. However, not every single word extracted by the encoder block has the same level of contribution to the sentence meaning. We implemented attention mechanism to approximate the importance of every single feature vector extracted by the encoder block. Not to mention, every single output word of the decoder will have different level of correspondence to every input word. Hence, the output of attention layer should be fed to every decoder cell. Our proposed model is illustrated in Fig. 4.

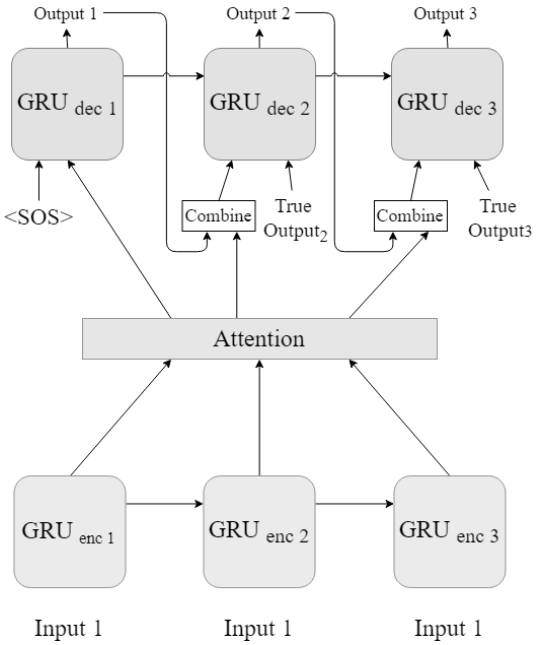

Figure 4 – Proposed Model

4. RESULTS AND DISCUSSION

In this chapter, we report the evaluation results of our argument generation system. We attempted to experiment with the Seq2Seq model with Negative Log Likelihood (NLL) loss. We present the loss progression and BLEU score of the model. The data that we extracted contains 10,226 unique words and 2,294 pairs of input-output (motion-claim). However, 10,226 unique words caused a curse of dimensionality problem. The softmax layer is the

generator network which has to compute 10,226 × output length and causes a memory problem. Therefore, we trimmed the dictionary of words in 2 ways. First, we limit the length of the motion and claim taken from the raw dataset source. Only the first 25 words of each motion are taken, and ten words of each claim are taken. Consecutively, we trimmed the words dictionary further by analyzing the word histogram plotting. We removed every word that occurs under five times in the entire corpus and finally come up with only 4,173 words. We further analyzed the impact of this mechanism by looking at the number of those rare words in comparison to the entire dataset, and it is under 5% of occurrence (4.38%). We used the "<UNK>" token to represent the rare words that are removed.

The main reason we used Seq2Seq is that the model allows us to manipulate the length of the output without having to follow the length of the input. The length of the input in every experiment is 20, and the output length is 10. We limit the length of the output quite small to prevent the vanishing gradient problem and exposure bias problem. As an enhancement, we implement scheduled sampling algorithm for each of the generator architecture. The model also implements Adam optimizer with a learning rate of 1e-4. Models are implemented from scratch using Pytorch.

4.1 LOSS PROJECTION

Loss function has one thing in common, which is the lower, the better. The following is the loss value projection over time for our model.

Figure 5 – Loss Graphs of Seq2Seq

As shown in Figure 5, Seq2Seq loss value progress is not stable. This is mostly due to the nature of RNN which is very hard to train. On top of that, the unconventional nature of the data makes it even harder for the model to learn. The data may present a plethora of output variations for one input. For instance, given an input "This house believes that abortions should be legal," the outputs are more than three. If the model successfully generates an argument that is very similar to the first output variation, then the model loss value will be small. But it is highly possible that the first output variation differs a lot from the third output (i.e., the third variation does not use the same word from the first variation). Therefore, it is hard for the model to learn the patterns, if there is any pattern.

4.2 BLEU SCORE

As shown in Table 1, the result of the BLEU score is the cumulative score of unigram, bigram, trigram, and 4-gram, while the unigram scores show how many words are in our generated claim in the real claim of a specific topic. With the seq2seq model, we achieve 0.664 BLEU score of Unigram; it shows that around 67% of the generated words are in the real claim dataset. The bigram, trigram, and 4-gram score are similar with the unigram, but with the difference that the bigram score compares the pair of words (such as that is, this house, or will ban) with the pair of words of the real claim dataset. The trigram and 4-gram will compare pairs of 3 words and four words respectively. The result shows that our model still struggles with generating sentences the right grammatical pattern like the real claim sentences. It is worth mentioning that the resulted data used to count the BLEU score is randomly selected. Therefore, replication results may vary.

We cannot make an apple-to-apple comparison to any previous work due to the different nature of the dataset. A research conducted before [8] focuses on the information retrieval technique instead of the generation technique. Hence, the dataset is also different. Their work used data that consists of motion sentences as the input and argument paragraphs as the output. On the other hand, our model uses motion sentences as the input and claim sentences as the output.

Table 1. BLEU Score Result for Seq2Seq

Model	Bleu Score	Unigram	Bigram	Trigram	4-Gram
Seq2Seq	0.258± 0.109	0.664± 0.164	0.253 ± 0.150	0.172 ± 0.108	0.175± 0.088

4.3 SUBJECTIVE EVALUATION

For our subjective evaluation, we use questionnaire made with Google Form. For our targeted groups, there are three main targets, first is the employees from several companies, second is students from certain university, and the third is a group chat of deep learning enthusiasts on Slack chat application. The questionnaire is comprised of ten claims; five claims are randomly picked from our dataset, and another five claims are picked from

our generated results. All motions and claims used in our questionnaire are shown in Table 2.

Table 2. List of Motions and Claims in Subjective Evaluation

No	Type	Motion	Claim
1	Human	This house would criminalize blasphemy	it's better to publish too much than not to have freedom
2	Human	This house believes that it is sometimes right for the government to restrict freedom of speech	there must be no constraints on the free flow of information and ideas
3	Human	This house believes that endangered species should be protected	the loss of native species as a loss to ecotourism
4	Human	This house believes that the Catholic Church is justified in forbidding the use of barrier methods of contraception	could open wide the way for marital infidelity
5	Human	This house would enforce term limits on the legislative branch of government	new lawmakers are more vulnerable to power by lobbyists
6	Machine	This house believes that male infant circumcision is tantamount to child abuse	circumcision without anesthetic is painful
7	Machine	This house would abolish the monarchy	the monarchy is an outdated and regressive
8	Machine	This house believes all nations have a right to nuclear weapons	nuclear weapons are intended to deter other states from attacking
9	Machine	This house believes the US is justified in using force to prevent states from acquiring nuclear weapons	nuclear proliferation may be beneficial for inducing stability
10	Machine	This house believes that Europe should weaken its austerity measures to guarantee its citizens greater social support	the harsh austerity measures have helped Greece

We share the questionnaire to 3 different target groups; first one includes fellow students on university level, second is for the working people, and last is to the chat group of deep learning community on a chat application. We got 40 responses for our subjective evaluation.

Our questionnaire requires the respondent to answer 2 questions for all 10 claims. First is to predict if the claim is human-made or machine generated. The second question asks whether the claim is in line with the motion given or not. From 40 respondents, the result we got for our model evaluation is shown in Table 3.

Table 3. Subjective Evaluation Results

Claim	Type	Prediction Result		In line with Motion	
		Human	Machine	Yes	No
1	Human	12	28	28	12
2	Human	13	27	22	18
3	Human	15	25	23	17
4	Human	18	22	24	16
5	Human	21	19	26	14
Human Total		**79 (39.50%)**	**121 (60.50%)**	**123 (61.50%)**	**77 (38.50%)**
6	Machine	13	27	25	15
7	Machine	16	24	25	15
8	Machine	18	22	25	15
9	Machine	19	21	33	7
10	Machine	21	19	30	10
Machine Total		**87 (43.50%)**	**113 (56.50%)**	**138 (69%)**	**62 (31%)**
Total		**166 (41.50%)**	**234 (58.50%)**	**261 (65.25%)**	**139 (34.75%)**

From the result, we conclude that our respondents have difficulties to differentiate the human claim from the generated claim, with almost all the prediction have balanced answers of true and false. Our respondents could predict correctly on generated claims (Type Machine predicted as machine in Table 3) as much as 47.5%, 67.5%, 52.5%, 55%, and 60% respectively. These results achieve 56.5% of average result. On the other hand, human-made claims achieve result of 45%, 52.5%, 30%, 32.5%, and 37.5% respectively, achieving average result of 39.5% of correct predicted result. Despite of our generated claim having higher correct prediction result, our model could successfully generate an adequate sentence that is hard to differentiate. This is because from total 400 answers of human or machine generated claims, 58.5% of the answers are generated claims. This means that our generated claims and the real claims from the dataset have similar sentence structure. And from our second question, we can conclude that our generated claim is in line with the motion given, with all claim have above half of the respondents say that the

claims are in line with the motion, specifically achieving average result of 69% across all generated claim.

5. CONCLUSION

Our seq2seq model with scheduled sampling synthesizes proper and motion-aware sentences. To quantitatively assess the model, we implemented BLEU score as the output quality metrics. Our model manages to achieve 0.175 ± 0.088 BLEU-4 score. To accompany the BLEU score that only quantitatively scores the grammatical structure of the outputs, we used the questionnaire to assess the quality of the output. We attempted to imitate the process to assess our model's output qualitatively. We asked the respondents to predict which one of the real claims or synthetic claims at the same time randomly is more "human-like" and relatable to the motion. The questionnaire results statistically present that the respondents cannot easily distinguish the machine-generated claims from the human-generated claims. For the future work, we will further improve our model capable of generating claim and use it as the generator block of Generative Adversarial Network (GAN) in hopes that we can achieve better results.

ACKNOWLEDGMENTS

We would like to thank Universitas Indonesia for grant "Hibah Tugas Akhir Mahasiswa Doktor" year 2018 numbered 1263/UN2.R3.1/HKP.05.00/2018 which support our research.

6. REFERENCES

[1] T. Govier, *A Practical Study of Argument*, Cengage Learning, 2013.

[2] S. Parsons, N. Oren, Reed, C. and F. Cerutti, *Computational Models of Argument*, Ios Press, 2014.

[3] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim, "Context dependent claim detection," *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers COLING'2014*, 2014, pp. 1489-1500.

[4] R. Rinott, L. Dankin, C.A. Perez, M.M. Khapra, E. Aharoni, and N. Slonim, "Show me your evidence – an automatic method for context dependent evidence detection," *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 17-21 September 2015, pp. 440–450.

[5] R. Bar-Haim, I. Bhattacharya, F. Dinuzzo, A. Saha, and N. Slonim, "Stance classification of context-dependent claims," *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1, 2017, pp. 251-261.

[6] D. Suhartono, A.P. Gema, S. Winton, T. David, M.I. Fanany, and A.M. Arymurthy, "Hierarchical attention network with XGBoost for recognizing insufficiently supported argument," In: S. Phon-Amnuaisuk, S-P. Ang, & S-Y. Lee (Eds.), *Multi-disciplinary Trends in Artificial Intelligence - 11th International Workshop, MIWAI 2017, Proceedings* (pp. 174-188). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 10607 LNAI), 2017, pp. 174-188. Springer Verlag. https://doi.org/10.1007/978-3-319-69456-6_15

[7] A.P. Gema, S. Winton, T. David, D. Suhartono, M. Shodiq, and W. Gazali, "It takes two to tango: Modification of Siamese long short term memory network with attention mechanism in recognizing argumentative relations in persuasive essay," *Procedia Computer Science*, vol. 116, pp. 449-459, 2017.

[8] M. Sato, K. Yanai, T. Miyoshi, T. Yanase, M. Iwayama, Q. Sun, and Y. Niwa, "End-to-end argument generation system in debating," *Proceedings of the ACL-IJCNLP 2015 System Demonstrations*, 2015, pp. 109-1142015.

[9] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," *Interspeech*, vol. 2, pp. 3, 2010.

[10] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," arXiv preprint arXiv:1409.2329, 2014.

[11] Y. Kim, Y. Jernite, D. Sontag, and A.M. Rush, "Character-aware neural language models," *AAAI*, pp. 2741-2749, 2016.

[12] Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen, and L. Carin, "Adversarial feature matching for text generation," In: D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 4006-4015, Sydney, Australia, August, 2017.

[13] K. Cho, B.V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[14] I. Sutskever, O. Vinyals, and Q.V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, pp. 3104-3112, 2014.

[15] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," arXiv preprint arXiv:1511.06732, 2015.
[16] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," *Advances in Neural Information Processing Systems*, pp. 1171-1179, 2015.
[17] K. Cho, B.V. Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," arXiv preprint arXiv:1409.1259, 2014.
[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.
[19] Habernal, and I. Gurevych, "Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2016, pp. 1589-1599.
[20] C. Stab and I. Gurevych, "Recognizing insufficiently supported arguments in argumentative essays," *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1, 2017, pp. 980-990.
[21] R. Manurung, G. Ritchie, and H. Thompson, "Using genetic algorithms to create meaningful poetic text," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 24, issue 1, pp. 43-64, 2012.
[22] G. Ritchie, R. Manurung, H. Pain, A. Waller, R. Black, and D. Omara, "A practical application of computational humour," *Proceedings of the 4th International Joint Conference on Computational Creativity*, 2007, pp. 91-98.
[23] Y. Zhang, Z. Gan, and L. Carin, Generating text via adversarial training, Proceedings of the Workshop on Adversarial Training, NIPS 2016, Barcelona, Spain, 2016, pp. 1-6.
[24] Z. Hu, Z. Yang, R. Salakhutdinov, and E.P. Xing, "On unifying deep generative models," arXiv preprint arXiv:1706.00550, 2017.
[25] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, issue 1, pp. 2096-2030, 2016.
[26] D.P. Kingma, and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.
[27] C. Napoles, M. Gormley, and B.V. Durme, "Annotated gigaword," *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, Association for Computational Linguistics*, 2012, pp. 95-100.
[28] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT press, Cambridge, 2016.
[29] J. Li, M.T. Luong, and D. Jurafsky, "A hierarchical neural autoencoder for paragraphs and documents," arXiv preprint arXiv:1506.01057, 2015.

Derwin Suhartono *is faculty member of Bina Nusantara University, Indonesia. He got his PhD in computer science from Universitas Indonesia in 2018. His research fields are natural language processing. Recently, he is doing research in argumentation mining and personality recognition. He actively involves in Indonesia Association of Computational Linguistics (INACL), a national scientific association in Indonesia. He has professional memberships in ACM, INSTICC, and IACT. He also takes role as reviewer in many international conferences and journals.*

Aryo Pradipta Gema *has bachelor's degree from Bina Nusantara University majoring Computer Science with Intelligent Systems specialty. He is also one of many awardees of best student in the university. In his final year of study, he undergoes an enrichment program provided by his university as a junior researcher. Main topic that he is interested with is deep learning. He writes and presents widely on argumentation mining research, a subfield of natural language processing field of research as well as several image processing tasks.*

Suhendro Winton *has bachelor's degree from Bina Nusantara University majoring Computer Science with Intelligent Systems specialty. In his final year of study, he undergoes an enrichment program provided by his university as a junior researcher. Main topic that he is interested with is deep learning. He writes and presents widely on argumentation mining research, a subfield of natural language processing field of research.*

Theodorus David *has bachelor's degree from Bina Nusantara University majoring Computer Science with Intelligent Systems specialty. In his final year of study, he undergoes an enrichment program provided by his university as a junior researcher. Main topic that he is interested with is deep learning. He writes and presents widely on argumentation mining research, a subfield of natural language processing field of research.*

Mohamad Ivan Fanany *a researcher and lecturer at Faculty of Computer Science, Universitas Indonesia. His research interests include machine learning, data science, and combining vision and graphics, remote sensing, climate modeling, biomedical engineering. Before joining the faculty, he worked at Future Project Div. Toyota Motor Corp, Japan, as a member of middleware development and recognition team; NHK ES Inc., as a researcher of IT21 Millennium Project on Advanced High Resolution and Highly Sensible Presence 3D Content Creation funded by NICT Japan; and a JSPS Fellow and Research Assistant at Imaging Science and Engineering, Tokyo Institute of Technology (Tokyo Tech). He served as the Chairman of Titech IEEE student branch 2002-2003. Currently a member of IEEE Consumers Electronics and IEEE Geoscience and Remote Sensing.*

Aniati Murni Arymurthy *is a Professor in computer science with specialty in computer vision and image processing. She got her MSc from Computer and Information Sciences Department in The Ohio State University (OSU), Columbus, Ohio, USA. She got PhD from Universitas Indonesia with sandwich program in Pattern Recognition and Image Processing Lab (PRIP Lab), Department of Computer Science, Michigan State University (MSU), East Lansing, Michigan, USA. Currently, she is active as lecturer in Faculty of Computer Science, Universitas Indonesia. Her research interests include pattern recognition, image processing, and spatial data.*