ArgDiver: Generating Sentential Arguments from Diverse Perspectives on Controversial Topic

ChaeHun Park **Wonsuk Yang** **Jong C. Park**[†]
School of Computing
Korea Advanced Institute of Science and Technology
{ddehun, derrick0511, park}@nlp.kaist.ac.kr

Abstract

Considering diverse aspects of an argumentative issue is an essential step for mitigating a biased opinion and making reasonable decisions. A related generation model can produce flexible results that cover a wide range of topics, compared to the retrieval-based method that may show unstable performance for unseen data. In this paper, we study the problem of generating sentential arguments from multiple perspectives, and propose a neural method to address this problem. Our model, ArgDiver (Argument generation model from Diverse perspectives), in a way a conversational system, successfully generates high-quality sentential arguments. At the same time, the automatically generated arguments by our model show a higher diversity than those generated by any other baseline models. We believe that our work provides evidence for the potential of a good generation model in providing diverse perspectives on a controversial topic.

1 Introduction

If one wants to address a potentially controversial issue, it is important to consider all of its aspects. When there are many such issues, some means of automating the process are called for. Automatically providing diverse aspects of an argumentative topic has thus received much attention. For instance, Wachsmuth et al. (2017) and Stab et al. (2018) developed a search engine for various arguments, while distinguishing the stance of each for a given claim. Ajjour et al. (2018) retrieved related arguments on a given topic, mapped the arguments to a topic space, and visualized such arguments within the topic space according to their distribution and their topical tendency.

These researches on a retrieval-based system have been very active, such as retrieving claims from documents (Levy et al., 2014; Lippi and Torroni, 2015, 2016) and discovering multiple viewpoints from an online debate (Trabelsi and Zaiane, 2018). As the outputs of these retrieval-based systems are based on sentences originally written by a human writer (as implied in the name "retrieval"), their outputs are often quite diverse and of high-quality.

However, a retrieval-based system does not have sufficient flexibility towards input with missing keywords or topics unseen to the database on which the system is based. Therefore, the performance of a retrieval-based system is bound by the coverage of the database. In response, a generation system has recently been looked into for argument mining. Wang and Ling (2016) summarized arguments to show only important contents in large text. Hua and Wang (2018) and Hua et al. (2019) generated counter-arguments for a given statement. Hidey and McKeown (2019) edited an original claim from the Reddit comments to generate contrastive claims. Online review generation, taking into account the personality of each e-commerce user, has also been actively studied (Ni and McAuley, 2018; Li et al., 2019). Well-trained generation-based systems could generate the results relatively independent of the coverage of the training data, since these systems could be generalized easily for an unseen dataset.

Still, a common problem that generation-based systems suffer from is that they often provide too generic output regardless of the input text (e.g., "I don't know.", "I don't agree with you."). Also, a popular sequence-to-sequence (Seq2Seq) framework (Sutskever et al., 2014) for various text generation tasks is designed to generate only one output from an input (*one-to-one*). Therefore, it is hard to model a *one-to-many* relationship, which is arguably more suitable for argument generation as a real-world argument may have multiple per-

[†] Corresponding author

Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 56–65
Hong Kong, China, November 4, 2019

Claim	This House believes university education should be free.
Sentential argument 1	Individuals have a right to the experience of higher education.
Sentential argument 2	The state benefits from the skills of a university educated populace.
Sentential argument 3	The cost to the state is far too great to sustain universal free education.
Sentential argument 4	State control of acceptance and curriculum criteria has negative effects.

Table 1: Example of a claim and its diverse sentential arguments.

spectives.

In this paper, we describe a model called ArgDiver, which stands for Argument generation model from Diverse perspectives, to overcome the limitations above of a generation-based argumentation system. For a given claim, ArgDiver generates multiple sentential arguments that cover diverse perspectives on the given claim. Table 1 shows an example[1] of the input and outputs of our system. More specifically, given a claim in favor of free university education, sentential arguments 1 and 2 support the claim, considering the right for higher education and benefits of the state, respectively. On the other hand, sentential arguments 3 and 4 are against the claim, considering the financial burden of the state and the negative effects of the intervention by the state, respectively. We understand that diverse perspectives of this kind should be provided with deep and varied stances, not only with a binary stance, towards given claims.

Our model adopts a Seq2Seq framework and introduces latent mechanisms based on the hypothesis that each latent mechanism may be matched with one perspective (Zhou et al., 2017, 2018; Tao et al., 2018; Gao et al., 2019; Chen et al., 2019a). We present a model that is trained by simply selecting a latent mechanism to optimize the model towards each target argument. Our model can avoid the generation of redundant outputs and be trained with a more accurate optimization strategy.

We use the PERSPECTRUM dataset proposed by Chen et al. (2019b). This dataset consists of pairs of one claim sentence (e.g., "Animals should have lawful rights.") and more than one cluster of sentential arguments (e.g., "Animals are equal to human beings.", "Animals have no interest or rationality."). Each cluster contains more than one sentential argument that share the same perspective within the cluster. In our research, we use a claim sentence as the input sequence of the model and each sentential argument as a target sequence of the model.

We evaluate our model with two measures, a) the quality of each of the generated sentential arguments, and b) their diversities. For the generation quality, we use BLEU score (Papineni et al., 2002) and three word embedding based metrics (Liu et al., 2016). For diversity, we use Dist-1/2 metric (Li et al., 2016) and a newly proposed metric. Experimental results show that our model generates sentential arguments of quality comparable to that of strong baseline models. Furthermore, our model generates more diverse sentential arguments than the baseline models.

The rest of this paper is organized as follows. We describe the related work in Section 2 and present our neural model in Section 3. We then describe the experimental settings and results in Sections 4 and 5, respectively. Finally, we conclude our work in Section 6.

2 Related Work

2.1 Argumentative Text Generation

Argumentative text generation is an active research area. Paul and Girju (2010) detected various contrastive viewpoints from an argumentative text by summarization. Le et al. (2018) proposed a chatbot to interact and debate with people with both retrieval-based and generation-based methods. Hua and Wang (2018) and Hua et al. (2019) generated counter-arguments given a statement on a controversial topic. They used an external knowledge (e.g., Wikipedia) to enrich their model. Hidey and McKeown (2019) edited the original claim semantically to generate a contrastive claim. Wachsmuth et al. (2018) and Khatib et al. (2017) discovered effective strategies and patterns that enhance persuasive argumentation. The most relevant work to the present research would be a retrieval-based system by Sato et al. (2015) that collects relevant sentences with frequently mentioned topics for debate (e.g., pollution, disease, poverty), and reorders them to offer related arguments. However, their system requires a pre-defined topic, a dictionary, and rules,

[1] https://idebate.org/debatabase

unlike ours.

2.2 Response Generation

Recently, neural generation models built upon a Seq2Seq framework (Sutskever et al., 2014) have been widely used in many text generation tasks, such as machine translation, document summarization and response generation (Bahdanau et al., 2015; Luong et al., 2015; Vinyals and Le, 2015; Nallapati et al., 2016; Xing et al., 2017). A few of them incorporate latent mechanisms to model the diversity of acceptable responses and one-to-many relationships. (Zhou et al., 2017, 2018) proposed an augmented Seq2Seq model with multiple latent mechanism embedding. Gao et al. (2019) used latent keywords as an additional factor to generate multiple responses and trained a model using a reinforcement learning algorithm. Tao et al. (2018) proposed a multi-head attention mechanism with a Seq2Seq model to attend various semantic aspects of an input text, using the heads to generate multiple responses. Chen et al. (2019a) claimed the importance of accurate optimization using a latent mechanism while proposing a posterior mapping selection that considers both the input text and target responses.

3 Method

3.1 Overview of ArgDiver

Our model is based on a neural Seq2Seq model with attention mechanism (Sutskever et al., 2014; Bahdanau et al., 2015). We extend this framework by inserting *N* different latent mechanisms to model the *one-to-many* relationship. Our model is trained to generate an independent sentential argument for a latent mechanism. In training, our model generates *N* different candidate arguments for a claim and uses only one of them using the minimum negative log-likelihood (NLL) for optimization. By this, our model can avoid general and redundant responses and each latent mechanism can help generate diverse arguments. In testing, each latent mechanism is utilized to generate a sentential argument. Our model may be understood as an extension of the model suggested by Zhou et al. (2017), in using latent mechanisms. Our model selects proper latent mechanisms to increase the diversity of the arguments that it generates.

3.2 Proposed Model

Assume a claim X and a group of related arguments P_1, P_2, P_3. Our proposed model takes a sequence of tokens within the claim $X = (x_1, x_2 \ldots, x_{|X|})$ as input, where x_i is a token at timestep i and $|X|$ is the length of the claim. Each token is passed to the word embedding layer and transformed into a fixed size word embedding vector $e(x_i)$. Each word embedding vector is then transformed into a hidden state h_i by one-layer bidirectional GRU (bi-GRU) encoder (Cho et al., 2014) as follows:

$$h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}] \tag{1}$$

$$\overrightarrow{h_i} = GRU(\overrightarrow{h_{i-1}}, e(x_i)) \tag{2}$$

$$\overleftarrow{h_i} = GRU(\overleftarrow{h_{i+1}}, e(x_i)) \tag{3}$$

where $[\overrightarrow{h_i}; \overleftarrow{h_i}]$ denotes the concatenation of forward and backward hidden states at timestep i, $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ are the forward and backward hidden states at timestep i, respectively. The last hidden states of both directions are then concatenated into $h = [\overrightarrow{h_{|x|}}; \overleftarrow{h_1}]$. This vector is used as the final semantic representation of the input claim.

Our model uses one-layer unidirectional GRU as the decoder. The semantic representation of the claim is concatenated with randomly initialized N different latent mechanisms M=$(m_1, m_2 \ldots, m_N)$, to make N different semantic representations H=$([h; m_1], [h; m_2], \ldots, [h; m_N])$. These concatenated representations are then used independently as N different initial states of the decoder.

The hidden state of the decoder is updated by an attention mechanism as proposed by Bahdanau et al. (2015):

$$\mathrm{s_{kt}} = GRU(\mathrm{s_{kt-1}}, \mathrm{c_{kt-1}}, e(\mathrm{y_{t-1}})); \mathrm{s_{k1}} = h_k \tag{4}$$

$$\mathrm{c_{kt}} = \sum_{i=1}^{|X|} \mathrm{a_{kti}} h_i \tag{5}$$

$$\mathrm{a_{kti}} = \frac{\exp(\mathrm{e_{kti}})}{\sum_{j=1}^{|X|} \exp(\mathrm{e_{ktj}})} \tag{6}$$

$$\mathrm{e_{kti}} = v^T \tanh(W_h[\mathrm{s_{kt}}; h_i]) \tag{7}$$

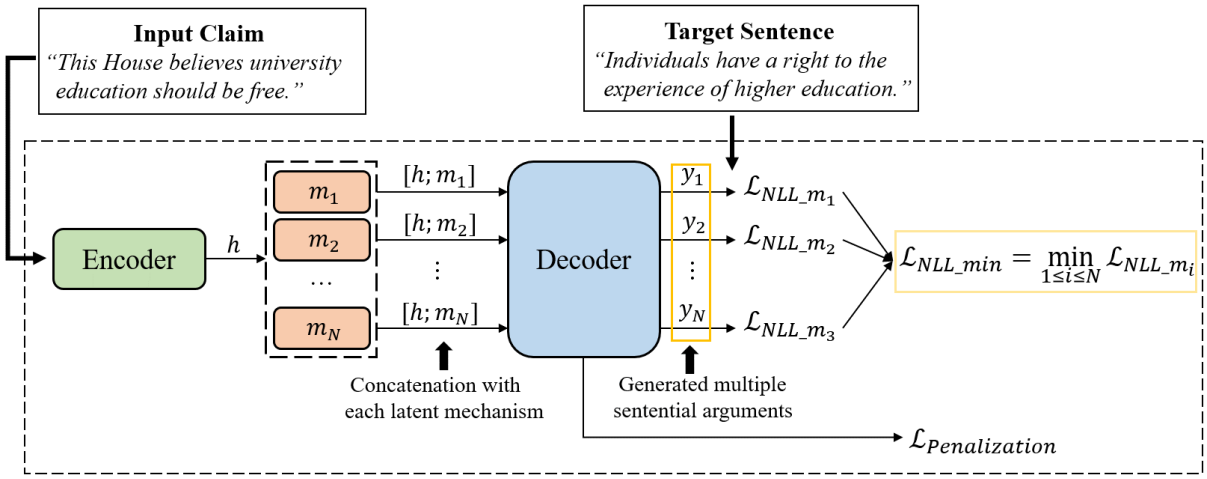

Figure 1: Overview of our sentential arguments generation model.

where s_{kt} denotes the hidden state at timestep t with the kth latent mechanism, and W_h and v^T are learnable parameters. $e(y_{t-1})$ is the word embedding vector of the target token at timestep $t-1$. c_{kt-1} is the context vector at timestep $t-1$ with the kth latent mechanism, which is the weighted sum of the hidden states of the encoder.

3.3 Objective Function

The remaining part of the model architecture is choosing the proper objective function to train our model for one target and multiple generated results. A general and typical approach in this case is calculating all losses of each generated argument and averaging them:

$$\mathcal{L}_{\text{NLL\_avg}} = -\frac{1}{N}\sum_{i=1}^{N} \log P(Y|X, m_i) \quad (8)$$

where NLL means negative log-likelihood and $P(Y|X, m_i)$ is the conditional probability that the model generates the target argument Y when input claim X and latent mechanism m_i are given. However, a naïve and rough optimization that does not select the appropriate latent mechanism to generate the given target argument may result in poor and redundant performance (Gao et al., 2019; Chen et al., 2019a). To avoid this, we select only one generated argument that shows minimum NLL for the given target argument to optimize our model, following Gao et al. (2019):

$$\mathcal{L}_{\text{NLL\_min}} = \min(\{-\log P(Y|X, m_1), \ldots, -\log P(Y|X, m_N)\}) \quad (9)$$

This is based on the hypothesis that the most appropriate latent mechanism to generate the target sentential argument would generate the best result with target (minimum NLL), compared with other generated results using other latent mechanisms. We compare the impacts of two different objective functions on performance in Section 5.2.

3.4 Penalty Term

We introduce an additional penalty term into the objective function, to regularize each latent mechanism to attend different semantic aspects of the input claim and avoid redundant outcomes within different latent mechanisms. We follow the work by Lin et al. (2017) and Tao et al. (2018), to encourage each latent mechanism to focus consistently on different and diverse semantic aspects of the input text. We accumulate the attention distribution of the decoder for each decoder timestep per latent mechanism, and normalize it by the length of the target sequence. We then concatenate them to make an $N \times |X|$ dimension matrix as follows:

$$A_k = \frac{\sum_{i=1}^{|Y|} a_{kti}}{|Y|} \in \mathbb{R}^{1\times|X|} \quad (10)$$

$$A = \{A_1||A_2||\ldots||A_N\} \in \mathbb{R}^{N\times|X|} \quad (11)$$

where A_k is the result of mean pooling across the decoding timestep, where $\sum A_k$ is 1. We then introduce a Frobenius norm after dot product between A and A^T, and subtract an identity matrix from it:

$$\mathcal{L}_{\text{penalization}} = \left\|AA^T - I\right\|_F^2 \quad (12)$$

where $\|\cdot\|_F^2$ is the square after standard Frobenius norm and I is an identity matrix. Note that each

element $AA^T[i,j]$ is the summation after element-wise product of the two attention distributions A_i and A_j. To minimize the term above, the diagonal elements and other elements of AA^T should be approximated to 1 and 0, respectively. This makes two attention distributions by different latent mechanisms to become more orthogonal to each other on the semantic space, encouraging each attention distribution to become more sparse.

The final objective function of our model is defined as:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{NLL\_min}} + (1-\lambda)\mathcal{L}_{\text{penalization}} \quad (13)$$

where $\mathcal{L}_{\text{NLL\_min}}$ is negative log likelihood that is defined in Equation 9 and $\mathcal{L}_{\text{penalization}}$ is defined in Equation 12. λ is the hyperparameter that controls the weight of two loss terms.

4 Experiments

4.1 Dataset

We use the PERSPECTRUM dataset (Chen et al., 2019b), which consists of a sentence that corresponds to a claim (e.g., "Animals should have lawful rights.") and more than one group of sentential arguments. Each argument group contains diverse sentential arguments regarding the claim (e.g., "Animals are equal to human beings.", "Animals have no interest or rationality."), and sentences in the same group share the same perspective towards the claim. We use the claim sentence as an input sequence and each sentence of every sentence group as the target sequence of our model. The dataset contains 907 claims and 11,164 related sentential arguments. We split the dataset into 541, 139, and 227 claims (and the corresponding sentential arguments) for training, validation, and testing, respectively. We use the split guidelines from Chen et al. (2019b), making sure that claims on the same topic are in the same partition. The split guidelines are to prevent the model from overfitting to a fixed set of keywords.

4.2 Compared Method

We compare our proposed model with several neural response generation models.

Seq2Seq + attention (Bahdanau et al., 2015): The standard sequence-to-sequence architecture with soft attention mechanism.

MMI-bidi (Li et al., 2016): Beam search using Maximum Mutual Information (MMI) to generate diverse outputs, by using both input sequence to output sequence and vice versa. We train another Seq2Seq model that generates input sequence from output sequence. We used the hyperparameters of λ=0.5, γ=1 and beam size=100.

MARM (Zhou et al., 2017): This model augments the Seq2Seq model with latent mechanism embedding to model the diversity of responding mechanisms. The number of latent mechanisms is set to 5.

CMHAM (Tao et al., 2018): This model uses multi-head attention with a Seq2Seq architecture and introduces a penalty term to encourage diverse attentions over different heads. We used 5 heads in our experiments.

MMPMS (Chen et al., 2019a): This model maps the semantic representation of the input text into multiple semantic spaces, and selects an appropriate mapping using both the input text and a target response. We set the number of mappings to 12.

ArgDiver: We use a model that is trained with the objective function in Equation 9 as our proposed model (**ArgDiver**). In addition, we compare our model with a variant that is trained with the objective function in Equation 8 (**ArgDiver**$_{\text{avg}}$) as described in Section 5.2.

4.3 Evaluation

We evaluate the models with two critics, the quality and the diversity of the generated sentential arguments for each.

For the quality, we use the following metrics. For the evaluation of a multiple argument generation system, we measure the score of each generated argument and report their average score.

BLEU-1/2 (Papineni et al., 2002): A widely used metric for the text generation task by measuring n-gram precision. We regard the target arguments that correspond to an input claim as the multiple references to calculate the score.

Embedding Average/Greedy/Extreme (Liu et al., 2016): These metrics evaluate results based on the semantic similarity between hypothesis and references, using a semantic representation by word embedding. These metrics take into account the diversity of a possible hypothesis and have been adopted for the evaluation of a conversation system (Xu et al., 2017; Tao et al., 2018).

For the diversity, we use the following metrics.

Dist-1/2 (Li et al., 2016): The number of unique unigrams/bigrams within a sentence normalized

Method	BLEU-1	BLEU-2	Embedding Average	Embedding Greedy	Embedding Extreme
Seq2Seq	0.3189	0.0947	0.8489	0.6198	0.4142
MMI-bidi	0.2263	0.0755	**0.8660**	0.6507	0.3971
MARM	0.2352	0.0099	0.7875	**0.6707**	**0.4497**
CMHAM	0.3227	**0.1009**	0.8334	0.6192	0.4069
MMPMS	0.2676	0.0725	0.8162	0.6256	0.4186
ArgDiver	**0.3268**	0.0964	0.8107	0.6002	0.4146

Table 2: Automatic evaluation results on generation quality. The highest and second highest scores are highlighted by bold and underline, respectively, for each metric.

Method	Dist-1	Dist-2	Dist-1-within	Dist-2-within
Seq2Seq	0.1230	0.2697	0.1624	0.2903
MMI-bidi	0.0707	0.2014	0.0868	0.1757
MARM	0.0456	0.0753	0.0377	0.1200
CMHAM	0.1418	**0.3236**	0.3222	0.5412
MMPMS	0.0650	0.1376	0.1485	0.3389
ArgDiver	**0.1585**	0.2909	**0.3645**	**0.6134**

Table 3: Automatic evaluation results on diversity of generation. The highest and second highest scores are highlighted by bold and underline, respectively, for each metric.

by the total number of unigrams/bigrams.

Dist-1/2-within: To the best of our knowledge, there has been no widely used metric to measure the diversity among multiple generated texts. We propose a simple metric to measure the diversity within the generated texts from the given input text, namely, Dist-1/2-within. To this end, this metric is calculated by *(The sum of the numbers of unique n-grams for each result that does not occur in other results) / (The sum of all generated numbers of unigrams/bigrams).*

4.4 Implementation Details

We use a Tensorflow framework (Abadi et al., 2016) to implement our model and baselines. We adopt the pre-trained 300-dimensional Glove word embedding (Pennington et al., 2014) for the word embedding layer of each model. The vocabulary size is the same for all models and set as 50K. Stanford CoreNLP (Manning et al., 2014) is used to tokenize our dataset. We use 256-dimensional hidden states for encoder and 384-dimensional hidden states for decoder. We use a dropout on the GRU cells with a probability of 0.2 (Srivastava et al., 2014), and apply gradient clipping (Pascanu et al., 2013) with a maximum norm of 3. The maximum numbers of tokens for encoder and decoder are set both to 50 and the batch size is set to 16 for all models. We use Adam optimizer (Kingma and Ba, 2015), with the initial learning rate set to 0.0005. In our model, the number and the dimension of the latent mechanism are set to 5 and 128, respectively. We initialized each of the vectors that represent latent mechanisms to a uniform distribution over [-0.001, 0.001]. We use beam search for generation, where the beam size is set to 10, except for the MMI-bidi model. We pre-train the weights of our encoder and decoder with the Wikitext 103 dataset proposed by Merity et al. (2017), and use it to initialize the weights of all baseline models and ours. We set λ in Equation 13 as 0.5.

5 Results

5.1 Overall Performance

Table 2 shows the evaluation results of each model in terms of generation quality using BLEU score and word embedding based metrics. We can see that our model achieves competitive performance in nearly all metrics. In BLEU score, our model ArgDiver and CMHAM outperform other baseline models. For the word embedding metrics, however, the two models show relatively low performance.

The evaluation results about the diversity of the generation are shown in Table 3. We see that ArgDiver achieves the best performance in three metrics (Dist-1, Dist-1/2-within), and the second performance in one metric (Dist-2). Except for our model, CMHAM outperforms other baselines in all metrics. By this, we can see that our model can generate diverse and multiple arguments to ex-

Method	BLEU-1	BLEU-2	Embedding Average	Embedding Greedy	Embedding Extreme
ArgDiver$_{\mathrm{avg}}$	0.3376	0.1100	0.8561	0.6335	0.4270
ArgDiver	0.3268	0.0964	0.8107	0.6002	0.4146

Table 4: Automatic evaluation results on generation quality with different objective functions.

Method	Dist-1	Dist-2	Dist-1-within	Dist-2-within
ArgDiver$_{\mathrm{avg}}$	0.0976	0.1611	0.0159	0.0261
ArgDiver	0.1585	0.2909	0.3645	0.6134

Table 5: Automatic evaluation results on diversity of generation with different objective functions.

amine diverse aspects of a given claim.

5.2 Effect of Objective Function

As we described in Section 3.3, we compare the impact on performance of two different objective functions. Table 4 and Table 5 show the evaluation results of our models in terms of quality and diversity of generated text, respectively. In terms of the generation quality, ArgDiver$_{\mathrm{avg}}$ shows similar but slightly better performance than ArgDiver. Meanwhile, ArgDiver shows more promising results than ArgDiver$_{\mathrm{avg}}$ against the diversity metric. In particular, we see that each latent mechanism generates exactly the same texts to the given claim about 74% for ArgDiver$_{\mathrm{avg}}$, though only about 6% for ArgDiver. These results indicate that ArgDiver$_{\mathrm{avg}}$ fails to utilize the full capacity of latent mechanisms, and goes back to the vanilla Seq2Seq model. By this, we postulate that the accurate optimization of a model considering the difference of each latent mechanism is the key for generating truly diverse arguments.

5.3 Case Study

The sample generated sentential arguments by each model and by a human are displayed in Table 6. The human-generated arguments are from the PERSPECTRUM dataset. The results of Seq2Seq model begin with the same phrase, and make a difference by selecting different words at the ending steps of decoding. In case of the MMPMS model, some of the mappings generate meaningless and repeated results. This may be due to the absence of a posterior mapping selection as it requires the target argument for the generation to proceed, which is absent in the testing scenario. CMHAM model and ArgDiver generate diverse and high quality multiple arguments. Including the CMHAM model and our proposed model, exactly the same texts with different latent mechanisms are often found in the results. This may point out the limitation of a small size of the dataset and the necessity of advanced approaches, which is left for future work.

5.4 Limitations and Future Work

In this subsection, we discuss the limitations of the current work and possible ways to improve our proposal as future work.

For the prior distribution of latent mechanisms, our current model uses all latent mechanisms to generate individual sentential arguments for all kinds of claim. It is yet reasonable to posit that the appropriate degree of each latent mechanism for its use in generation may depend on the topic of the given claim. As future work, we plan to devise a model which considers the probability by which each latent mechanism would be used to generate sentential arguments with the given claim.

For the low interpretability of latent mechanism, ideal results of our model would be that there exist shared characteristics in the generated sentential arguments with the same latent mechanism and a different input claim. However, it is hard to observe these characteristics within the generated results of our model. In addition, the latent mechanism sometimes tends to generate the output by memorizing some of the frequent phrases in the dataset (e.g., "This is the right of (. . .).", "There is no need for compulsion."). One of the possible reasons is that each latent mechanism focuses on the syntactic difference of each sentential argument, rather than semantic differences such as topics or characteristics.

As future work, we plan to present an improved model to distinguish the semantic and syntactic factors of each perspective. One possibility is to model the latent personality in the sentential arguments. For instance, the person who is interested in environmental issues is more likely to have a

Claim	We should fear the power of government over the internet.
Human	Internet regulation is necessary to ensure a safe internet.
	Internet regulation is a euphemism for censorship.
	Internet governance is necessary to combat heinous crimes committed via the internet.
	Internet regulation is an attempt by big interest groups to regulate the internet in their favour.
Seq2Seq	There is no reason to have the negative impact on nationalist sentiment.
	There is no reason to have the negative impact on them.
	There is no reason to have the negative impact on politics.
	There is no reason to have the problems in the environment.
	There is no reason to have the negative impact on nationalist footprint.
CMHAM	Everyone should be allowed free speech.
	It is clear to impose their religion!
	The American people would be more accountable for the council.
	The American people would be more accountable for the council.
	This is a part of a crime and should not be the state.
MMPMS	The result of all should have the rights to have the right to have the right to all their own decisions.
	Domestic protect the vote.
	Make these equal off taken off against equal off countries would make all these rights as illegal as as as as as (. . .)
	The freedom of the economy would have the freedom of the freedom of the freedom of the freedom of the (. . .)
	It would have a negative impact .
ArgDiver	National sovereignty would result in a government's freedom of expression.
	The government should not be celebrated.
	It is a necessary for national security.
	It's conceivable to the wrong hands.
	The government is a best way to have a universal right to have a universal right to practice.

Table 6: Sample arguments of a claim generated by human and models.

relatively predictable and specific perspective on certain topics than those who are not. The generation model considering these aspects could provide more human-like arguments with a wide coverage of many persons' characteristics.

Another possibility would be for our model to incorporate the background knowledge to generate the arguments. We believe that such an explicit provision of the background knowledge to the model can increase the informativeness and the relevance of the generated arguments to the input claim.

6 Conclusion

In this work, we looked into a new task that generates diverse and multiple sentential arguments with the given claim on a controversial topic. To address this task, we introduced a new model based on the Seq2Seq framework, called ArgDiver, to optimize each latent mechanism more properly and generate diverse outputs. Experimental results confirm that diverse sentential arguments could be generated with high quality, and that our model shows higher diversity than any other baseline models.

Acknowledgments

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government MSIT) (No. 2018-0-00582-002, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection).

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283.

Yamen Ajjour, Henning Wachsmuth, Dora Kiesel, Patrick Riehmann, Fan Fan, Giuliano Castiglia, Rosemary Adejoh, Bernd Fröhlich, and Benno Stein. 2018. Visualization of the topic space of argument search results in args. me. In *Proceedings of the EMNLP 2018: System Demonstrations*, pages 60–65.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly

learning to align and translate. In *Proceedings of the 3rd ICLR*.

Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 2019a. Generating multiple diverse responses with multi-mapping and posterior mapping selection. In *Proceedings of the 28th IJCAI*, pages 4918–4924.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019b. Seeing things from a different angle: Discovering diverse perspectives about claims. In *Proceedings of the 17th NAACL-HLT*, pages 542–557.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Qatar, 25 October 2014*, pages 103–111.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. 2019. Generating multiple diverse responses for short-text conversation. In *Proceedings of the 33rd AAAI*, volume 33, pages 6383–6390.

Christopher Hidey and Kathy McKeown. 2019. Fixed that for you: Generating contrastive claims with semantic edits. In *Proceedings of the 17th NAACL-HLT*, pages 1756–1767.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th ACL*, pages 2661–2672, Florence, Italy.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th ACL*, pages 219–230.

Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the EMNLP 2017*, pages 1351–1357.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Dieu-Thu Le, Cam Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of the 25th COLING: Technical Papers*, pages 1489–1500.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT, San Diego California, USA, June 12-17*, pages 110–119.

Piji Li, Zihao Wang, Lidong Bing, and Wai Lam. 2019. Persona-aware tips generation. In *Proceedings of the WWW 2019*, pages 1006–1016. ACM.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *5th ICLR, Toulon, France, April 24-26, 2017*.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the 24th IJCAI*.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the 30th AAAI*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the EMNLP 2016*, pages 2122–2132.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the EMNLP 2015*, pages 1412–1421.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations*, pages 55–60.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th ICLR 2017, Toulon, France, April 24-26*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL*, pages 280–290.

Jianmo Ni and Julian McAuley. 2018. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *Proceedings of the 56th ACL*, pages 706–711.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, pages 311–318. Association for Computational Linguistics.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *ICML*, pages 1310–1318.

Michael Paul and Roxana Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proceedings of the 24th AAAI*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the EMNLP 2014, October 25-29, 2014, Qatar*, pages 1532–1543.

Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proceedings of the ACL-IJCNLP 2015: System Demonstrations*, pages 109–114.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. Argumentext: Searching for arguments in heterogeneous sources. In *Proceedings of the 16th NAACL: Demonstrations*, pages 21–25.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th NIPS*, pages 3104–3112.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *Proceedings of the 27th IJCAI*, pages 4418–4424.

Amine Trabelsi and Osmar R Zaiane. 2018. Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions. In *Proceedings of the 12th ICWSM*.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th ICCL*, pages 3753–3765.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 15th NAACL-HLT 2016*, pages 47–57.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the 31st AAAI*.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, Xiaolong Wang, Zhuoran Wang, and Chao Qi. 2017. Neural response generation via GAN with an approximate embedding layer. In *Proceedings of the EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 617–626.

Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In *Proceedings of the 31st AAAI*.

Ganbin Zhou, Ping Luo, Yijun Xiao, Fen Lin, Bo Chen, and Qing He. 2018. Elastic responding machine for dialog generation with dynamically mechanism selecting. In *Proceedings of the 32nd AAAI*.