Discrete Argument Representation Learning for Interactive Argument Pair Identification

Lu Ji[1] **, Zhongyu Wei**[2*]**, Jing Li**[3] **, Qi Zhang**[1] **, Xuanjing Huang**[1]
[1]School of Computer Science, Fudan University, Shanghai, China
[2]School of Data Science, Fudan University, China
[3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
{17210240034,zywei,qz,xjhuang}@fudan.edu.cn
jing-amelia.li@polyu.edu.hk

Abstract

In this paper, we focus on identifying interactive argument pairs from two posts with opposite stances to a certain topic. Considering opinions are exchanged from different perspectives of the discussing topic, we study the discrete representations for arguments to capture varying aspects in argumentation languages (e.g., the debate focus and the participant behavior). Moreover, we utilize hierarchical structure to model post-wise information incorporating contextual knowledge. Experimental results on the large-scale dataset collected from *CMV* show that our proposed framework can significantly outperform the competitive baselines. Further analyses reveal why our model yields superior performance and prove the usefulness of our learned representations.

1 Introduction

Arguments play a central role in decision making on social issues. Striving to automatically understand human arguments, computational argumentation becomes a growing field in natural language processing. It can be analyzed at two levels — monological argumentation and dialogical argumentation. Existing research on monological argumentation covers argument structure prediction (Stab and Gurevych, 2014), claims generation (Bilu and Slonim, 2016), essay scoring (Taghipour and Ng, 2016), etc. Recently, dialogical argumentation becomes an active topic.

In the process of dialogical arguments, participants exchange arguments on a given topic (Asterhan and Schwarz, 2007; Hunter, 2013). With the popularity of online debating forums, large volume of dialogical arguments are daily formed, concerning wide range of topics. A social media dialogical argumentation example from ChangeMyView subreddit is shown in Figure 1. There we show two posts holding opposite stances over the same topic. One is the original post and the other is reply. As can be seen, opinions from both sides are voiced with multiple arguments and the reply *post B* is organized in-line with *post A*'s arguments. Here we define an interactive argument pair formed with two arguments from both sides (with the same underline), which focuses on the same perspective of the discussion topic. The automatic identification of these pairs will be a fundamental step towards the understanding of dialogical argumentative structure. Moreover, it can benefit downstream tasks, such as debate summarization (Sanchan et al., 2017) and logical chain extraction in debates (Botschen et al., 2018).

CMV: The position of vice president of the USA should be eliminated from our government.
Post A: a1: [If the president is either killed or resigns, the vice president is a horrible choice to take over office.] **a2:** The speaker of the House would be more qualified for the position. **a3:** [I'm willing to bet that John Boehner would have an easier time dealing with congress as president than Joe Biden would due to his constant interaction with it.] **a4:** If Boehner took office, as a republican, would he do something to veto bills Obama supported?
Post B: b1:[Seriously, stop this hyperbole.] **b2:** [Do you think that have anything to do with the fact that Boehner is a republican, and republicans control congress?] **b3:** That argument has much less to do with the individuals than it does with the current party in control.

Figure 1: An example of dialogical argumentation consists of two posts from change my view, a sub-forum of Reddit.com. Different types of underlines are used to highlight the interactive argument pairs.

However, it is non-trivial to extract the interactive argument pairs holding opposite stances. Back to the example. Given argument **b1** with only four words contained, it is difficult, without richer contextual information, to understand why it has interactive relationship with **a1**. In addition, without modeling the debating focuses of arguments, it is likely for models to wrongly predict that **b2** has interactive relationship with **a4** for sharing more words. Motivated by these observations, we pro-

*Corresponding author

pose to explore discrete argument representations to capture varying aspects (e.g., the debate focus) in argumentation language and learn context-sensitive argumentative representations for the automatic identification of interactive argument pairs.

For argument representation learning, different from previous methods focusing on the modeling of continuous argument representations, we obtain discrete latent representations via discrete variational autoencoders and investigate their effects on the understanding of dialogical argumentative structure. For context representation modeling, we employ a hierarchical neural network to explore what content an argument conveys and how they interact with each other in the argumentative structure. To the best of our knowledge, we are the first to explore discrete representations on argumentative structure understanding. In model evaluation, we construct a dataset collected from *CMV* [1], which is built as part of our work and has been publicly released [2]. Experimental results show that our proposed model can significantly outperform the competitive baselines. Further analysis on discrete latent variables reveals why our model yields superior performance. At last, we show that the representations learned by our model can successfully boost the performance of argument persuasiveness evaluation.

2 Task Definition and Dataset Collection

In this section, we first define our task of interactive argument pair identification, followed by a description of how we collect the data for this task.

2.1 Task Definition

Given a argument q from the original post, a candidate set of replies consisting of one positive reply r^+, several negative replies $r_1^- \sim r_u^-$, and their corresponding argumentative contexts, our goal is to automatically identify which reply has interactive relationship with the quotation q.

We formulate the task of identifying interactive argument pairs as a pairwise ranking problem. In practice, we calculate the matching score $S(q, r)$ for each reply in the candidate set with the quotation q and treat the one with the highest matching score as the winner.

[1] https://reddit.com/r/changemyview

[2] http://fudan-disc.com/data/arg-pairs-fudanU.zip

2.2 Dataset Collection

Our data collection is built on the *CMV* dataset released by Tan et al. (2016). In *CMV*, users submit posts to elaborate their perspectives on a specific topic and other users are invited to argue for the other side to change the posters' stances. The original dataset is crawled using Reddit API. Discussion threads from the period between January 2013 and May 2015 are collected as training set, besides, threads between May 2015 and September 2015 are considered as test set. In total, there are 18,363 and 2,263 discussion threads in training set and test set, respectively.

An observation on *CMV* shows that when users reply to a certain argument in the original post, they quote the argument first and write responsive argument directly, forming a quotation-reply pair. Figure 2 shows how quotation-reply pairs could be identified. Inspired by this finding, we decide to extract interactive argument pairs with the relation of quotation-reply. In general, the content of posts in *CMV* is informal, making it difficult to parse an argument in a finer-grain with premise, conclusion and other components. Therefore, following previous setting in Ji et al. (2018), we treat each sentence as an argument. Specifically, we only consider the quotation containing one argument and view the first sentence after the quotation as the reply. We treat the quotation-reply pairs extracted as positive samples and randomly select four replies from other posts that are also related to the original post to pair with the quotation as negative samples. In detail, each instance in our dataset includes the quotation, one positive reply, four negative replies, and the posts where they exist. The posts where they exist refer to argumentative contexts mentioned below. What's more, we remove quotations from argumentative contexts of replies.

Original Post: ... Strong family values in society lead to great results. *I want society to take positive aspects of the early Americans and implement that into society.* This would be a huge improvement than what we have now. ...
User Post: *> I want society to take positive aspects of the early Americans and implement that into society.* What do you believe those aspects to be? ...

Figure 2: An example illustrating the formation process of a quotation-reply pair in *CMV*.

We keep words with the frequency higher than 15 and this makes the word vocabulary with 20,692 distinct entries. In order to assure the quality of quotation-reply pairs, we only keep the instance where the number of words in the quotation and

	training set	test set
# of arg. per post	11.8±6.6	11.4±6.2
# of token per post	209.7±117.2	205.9±114.6
# of token per q	20.0±8.6	20.0±8.6
# of token per p_r	16.9±8.1	17.3±8.4
# of token per n_r	19.0±8.0	19.1±8.1
max # of q-p_r pairs	12	9
avg. # of q-p_r pairs	1.5±0.9	1.4±0.9

Table 1: Overview statistics of the constructed dataset (mean and standard deviation). *arg.*, q, p_r, n_r represent *argument*, *quotation*, *positive reply* and *negative reply* respectively. q-p_r represents the quotation-reply pair between posts.

replies range from 7 to 45. We regard the instances extracted from training set and test set in Tan et al. (2016) for training and test. The number of instances in training and test set is 11,565 and 1,481, respectively. We randomly select 10% of the training instances to form the development set. The statistic information of our dataset is shown in Table 1.

To further demonstrate that quotation-reply pairs have interactive relationships, we randomly select 100 instances from the test set and hire two trained annotators who are fluent English speakers to identify interactive argument pairs. The accuracy of the two annotators is 0.83 and 0.93, respectively. The inter-annotator agreement measured by Co-hens Kappa (Carletta, 1996) is 0.82. This confirms the quality of the constructed dataset.

3 Proposed Model

The overall architecture of our model is shown in Figure 3(a). It takes a quotation, a reply and their corresponding argumentative contexts as inputs, and outputs a real value as its matching score. It mainly consists of three components, namely, *Discrete Variational AutoEncoders* (*DVAE*, Figure 3(c)), *Argumentative Context Modeling* (Figure 3(b)) and *Argument Matching and Scoring*. We learn discrete argument representations via *DVAE* and employ a hierarchical architecture to obtain the argumentative context representations. The *Argument Matching and Scoring* integrates some semantic features between the quotation and the reply to calculate the matching score.

3.1 Discrete Variational AutoEncoders

We employ discrete variational autoencoders (Rolfe, 2017) to reconstruct arguments from auto-encoding and obtain argument representations based on discrete latent variables to capture different aspects of argumentation languages.

Encoder. Given an argument x with words $w_1, w_2, ..., w_T$, we first embed each word to a dense vector obtaining $w_1^{'}, w_2^{'}, ..., w_T^{'}$ correspondingly. Then we use a bi-directional GRU (Wang et al., 2018) to encode the argument.

$$h_t = BiGRU(w_t^{'}, h_{t-1}) \tag{1}$$

We obtain the hidden state for a given word $w_t^{'}$ by concatenating the forward hidden state and backward hidden state. Finally, we consider the last hidden state h^T as the continuous representation of the argument.

Discrete Latent Variables. We introduce z as a set of K-way categorical variables $z = \{z_1, z_2, ..., z_M\}$, where M is the number of variables. Here, each z_i is independent and we can easily extend the calculation process below to every latent variables. Firstly, we calculate the logits l_i as follows.

$$l_i = W_l h_i^T + b_l \tag{2}$$

where $W_l \in R^{K \times E}$ stands for the weight matrix, E is the dimension of hidden units in encoder, while b_l is a weight vector.

After obtaining the logits l_i, we can calculate the posterior distribution and discrete code of z_i.

$$q(z_i|x) = Softmax(l_i) \tag{3}$$

$$Z_{code}(i) = \underset{k \in [1,2,...,K]}{\arg\max} (l_{ik}) \tag{4}$$

However, using discrete latent variables is challenging when training models end-to-end. To alleviate this problem, we use the recently proposed Gumbel-Softmax trick (Lu et al., 2017) to create a differentiable estimator for categorical variables. During training we draw samples $g_1, g_2, ..., g_K$ from the Gumbel distribution: $g_k \sim$-log(-log(u)), where $u \sim U(0, 1)$ are uniform samples. Then, we compute the log-softmax of l_i to get $\omega_i \in R^K$:

$$\omega_{ik} = \frac{exp((l_{ik} + g_k)/\tau)}{\sum_k exp((l_{ik} + g_k)/\tau)} \tag{5}$$

τ is a hyper-parameter. With low temperature τ, this vector ω_i is close to the one-hot vector representing the maximum index of l_i. But with higher temperature, this vector ω_i is smoother.

Then we map the latent samples to the initial state of the decoder as follows:

$$h_{dec}^0 = \sum_{i=1}^{M} W_{ei} \omega_i \tag{6}$$

where $W_{ei} \in R^{K \times D}$ is the embedding matrix, D is the dimension of hidden units in decoder. Finally, we use a GRU as the decoder to reconstruct the

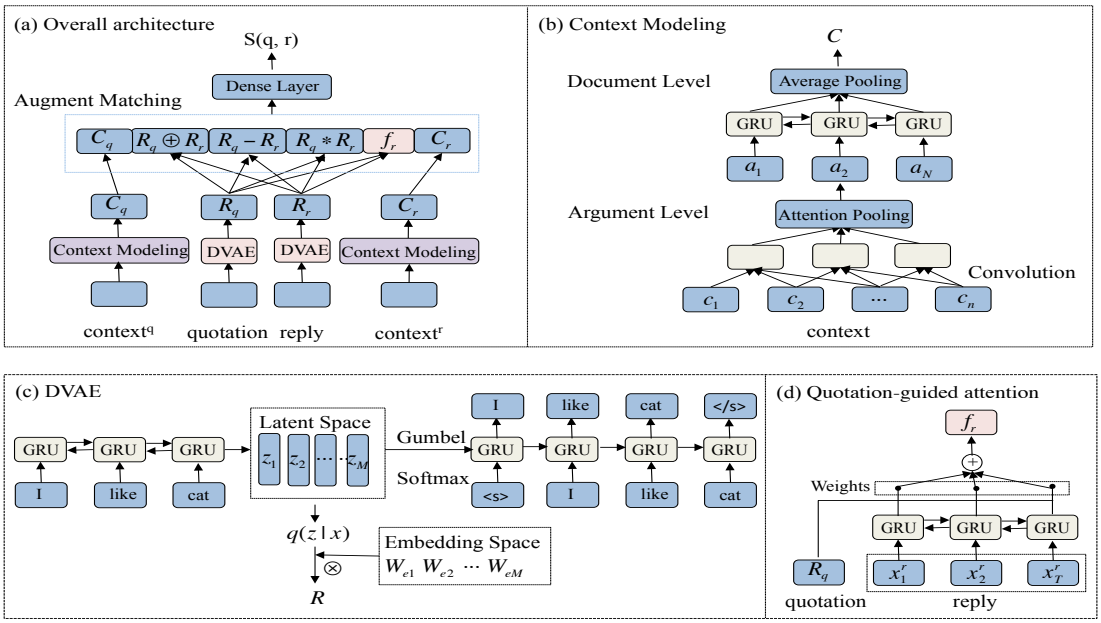

Figure 3: (a) Overall architecture of the proposed model. (b) Hierarchical architecture for argumentative context modeling. (c) Detailed structure of the discrete variational autoencoders (*DVAE*). (d) Structure of the quotation-guided attention in argument matching.

argument given h^0_{dec}.

Discrete Argument Representations. Through the process of auto-encoding mentioned above, we can reconstruct the argument. The representation that we want to find can capture varying aspects in argumentation languages and contain salient features of the argument. $q(z_i|x)$ shows the probability distribution of z_i over K categories, which contains salient features of the argument on varying aspects. Therefore, we obtain the discrete argument representation by the posterior distribution of discrete latent variables z.

$$R = \sum_{i=1}^{M} W_{ei} q(z_i|x) \tag{7}$$

3.2 Argumentative Context Modeling

Here, we introduce contextual information of the quotation and the reply to help identify the interactive argument pairs. The argumentative context contains a list of arguments. Following previous setting in Ji et al. (2018), we consider each sentence as an argument in the context. Inspired by Dong et al. (2017), we employ a hierarchical architecture to obtain argumentative context representations.

Argument-level CNN. Given an argument and their embedding forms $\{e_1, e_2, ..., e_n\}$, we employ a convolution layer to incorporate the context information on word level.

$$s_i = f(W_s \cdot [e_i : e_{i+ws-1}] + b_s) \tag{8}$$

where W_s and b_s are weight matrix and bias vector. ws is the window size in the convolution layer and s_i is the feature representation. Then, we conduct an attention pooling operation over all the words to get argument embedding vectors.

$$m_i = \tanh(W_m \cdot s_i + b_m) \tag{9}$$

$$u_i = \frac{e^{W_u \cdot m_i}}{\sum_j e^{W_u \cdot m_j}} \tag{10}$$

$$a = \sum_i u_i \cdot s_i \tag{11}$$

where W_m and W_u are weight matrix and vector, b_m is the bias vector, m_i and u_i are attention vector and attention weight of the i-th word. a is the argument representation.

Document-level BiGRU. Given the argument embedding $\{a_1, a_2, ..., a_N\}$, we employ a bi-directional GRU to incorporate the contextual information on argument level.

$$h^c_i = BiGRU(a_i, h^c_{i-1}) \tag{12}$$

Finally, we employ an average pooling over arguments to obtain the context representation C.

3.3 Argument Matching and Scoring

Once representations of the quotation and the reply are generated, three matching methods are applied to analyze relevance between the two arguments. We conduct element-wise product and

element-wise difference to get the semantic features $f_p = R_q * R_r$ and $f_d = R_q - R_r$. Furthermore, to evaluate the relevance between each word in the reply and the discrete representation of the quotation, we propose the quotation-guided attention and obtain a new representation of the reply.

Quotation-Guided Attention. We conduct dot product between R_q and each hidden state representation h_j^r in the reply. Then, a softmax layer is used to obtain an attention distribution.

$$v_j = softmax(R_q \cdot h_j^r) \quad (13)$$

Based on the attention probability v_j of the j-th word in the reply, the new representation of the reply can then be constructed as follows:

$$f_r = \sum_j v_j \cdot h_j^r \quad (14)$$

After obtaining the discrete representations, argumentative context representations and some semantic matching features f_p, f_d, f_r of the quotation and the reply, we use two fully connected layers to obtain a higher-level representation H. Finally, the matching score S is obtained by a linear transformation.

$$f_m = [f_p; f_d; f_r] \quad (15)$$

$$H = f(W_H[R_q; R_r; C_q; C_r; f_m] + b_H) \quad (16)$$

$$S = W_s H + b_s \quad (17)$$

where W_H and W_S stand for the weight matrices, while b_H and b_S are weight vectors.

3.4 Joint Learning

The proposed model contains three modules, i.e., the *DVAE*, argumentative context modeling and argument matching, which are trained jointly. We define the loss function of the overall framework to combine the two effects.

$$L = L_{DVAE} + \lambda L_m \quad (18)$$

where λ is a hyper-parameter to balance the two loss terms. The first loss term is defined on the *DVAE* and cross entropy loss is defined as the reconstruction loss. We apply the regularization on *KL* cost term to solve posterior collapse issue. Due to the space limitation, we leave out the derivation details and refer the readers to Zhao et al. (2018).

$$L_{DVAE} = E_{q(z|x)}[\log p(x|z)] - KL(q(z|x)||p(z)) \quad (19)$$

The second loss term is defined on the argument matching. We formalize this issue as a ranking task and utilize hinge loss for training.

$$L_m = \sum_{i=1}^{u} max(0, \gamma - S(q, r^+) + S(q, r_i^-)) \quad (20)$$

where u is the number of negative replies in each instance. γ is a margin parameter, $S(q, r^+)$ is the matching score of the positive pair and $S(q, r_i^-)$ is the matching score of the i-th negative pair.

4 Experiment Setup

4.1 Training Details

We use Glove (Pennington et al., 2014) word embeddings with dimension of 50. The number of discrete latent variables *M* is 5 and the number of categories for each latent variable is also 5. What's more, the hidden units of GRU cell in encoder are 200 while that for the decoder is 400. We set batch size to 32, filter sizes to 5, filter numbers to 100, dropout with probability of 0.5, temperature τ to 1. The hyper-parameters in loss function are set as γ= 10 for max margin and λ= 1 for controlling the effects of discrete argument representation learning and argument matching.

The proposed model is optimized by SGD and applied the strategy of learning rate decay with initial learning rate of 0.1. We evaluate our model on development set at every epoch to select the best model. During training, we run our model for 200 epochs with early-stop (Caruana et al., 2000).

4.2 Comparison Models

For baselines, we consider simple models that rank argument pairs with cosine similarity measured with two types of word vectors: TF-IDF scores (henceforth TF-IDF) and the pre-trained word embeddings from word2vec corpus (henceforth WORD2VEC). Also, we compare with the neural models from related areas: MALSTM (Mueller and Thyagarajan, 2016), the popular method for sentence-level semantic matching, and CBCA-WOF (Ji et al., 2018), the state-of-the-art model to evaluate the persuasiveness of argumentative comments, which is tailored to fit our task. In addition, we compare with some ablations to study the contribution from our components. Here we first consider MATCH$_{rnn}$, which uses BiGRU to learn argument representations and explore the match of arguments without modeling the context therein. Then we compare with other ablations that adopt varying argument context modeling methods. Here we consider BiGRU (henceforth MATCH$_{rnn}$+C$_b$), which

Models	P@1	MRR
Cosine Similarity based		
TF-IDF	28.36*	51.66*
WORD2VEC	28.70*	52.03*
Neural-Network based		
MALSTM (Mueller and Thyagarajan, 2016)	31.26*	52.97*
CBCAWOF (Ji et al., 2018)	56.04*	73.03*
Ablation Study		
MATCH$_{rnn}$	51.52*	70.57*
MATCH$_{rnn}$+C$_b$	55.98*	73.20*
MATCH$_{rnn}$+C$_h$	57.46*	73.72*
MATCH$_{ae}$+C$_h$	58.27‡	74.16*
MATCH$_{vae}$+C$_h$	58.61‡	74.66‡
Our model	**61.17**	**76.16**

Table 2: The performances of different models on our dataset in terms of Mean Reciprocal Rank (*MRR*) and Precision at 1 (denoted as *P@1*). The proposed model significantly outperforms all the comparison methods marked with * or ‡ (*: p<0.01; ‡: p<0.05, Wilcoxon signed rank test). Best results are in **bold**.

focuses on words in argument context and ignores the argument interaction structure. We also consider a hierarchical neural network ablation (henceforth MATCH$_{rnn}$+C$_h$), which models argument interactions with BiGRU and the words therein with CNN. In addition, we compare with MATCH$_{ae}$+C$_h$ and MATCH$_{vae}$+C$_h$, employing auto-encoder (AE) and variational AE (VAE), respectively, to take the duty of the *DVAE* module of our full model.

5 Results and Discussions

To evaluate the performance of different models, we first show the overall performance of different models for argument pair identification. Then, we conduct three analyses including *hyper-parameters sensitivity analysis*, *discrete latent variables analysis* and *error analysis* to study the impact of hyper-parameters, explain why *DVAE* performs well on interactive argument pair identification and analyze the major causes of errors. Finally, we apply our model to a downstream task to investigate the usefulness of discrete argument representations.

5.1 Overall Performance Comparison

The overall results of different models are shown in Table 2. Mean Reciprocal Rank (*MRR*) and Precision at 1 (denoted as *P@1*) are used for evaluation metrics. We have following findings.

- Our model significantly outperforms all comparison models in terms of both evaluation metrics. This proves the effectiveness of our model.
- Neural network models perform better than TFIDF and WORD2VEC. This observation shows the effectiveness of argument representation learning in neural networks.
- By modeling context representations, MATCH$_{rnn}$+C$_b$ and MATCH$_{rnn}$+C$_h$ significantly outperform MATCH$_{rnn}$. This proves that contextual information is helpful for identifying interactive argument pairs.
- Argumentative contexts often contain a list of arguments. In comparison of MATCH$_{rnn}$+C$_b$ and MATCH$_{rnn}$+C$_h$, we find that MATCH$_{rnn}$+C$_h$ achieve much better results than MATCH$_{rnn}$+C$_b$. This demonstrates the effectiveness of representing argumentative contexts on argument level instead of word level.
- By using autoencoders for argument representation learning, our model, MATCH$_{vae}$+C$_h$ and MATCH$_{ae}$+C$_h$ outperform MATCH$_{rnn}$+C$_h$. This indicates the effectiveness of argument representation learning.

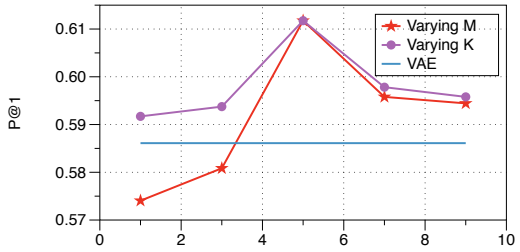

Figure 4: The impact of varying the number of discrete latent variables *M* and categories for each latent variable *K* on *P@1*. We find that our model still outperforms *VAE* which is the most competitive baseline.

5.2 Hyper-Parameter Sensitivity Analysis

We investigate the impact of two hyper-parameters on our model, namely the number of discrete latent variables *M* and the number of categories for each latent variable *K* in *DVAE*. For studying the impact of *M* and *K*, we set them as 1, 3, 5, 7, 9 respectively while keep other hyper-parameters the same as our best model. We report *P@1* of different settings.

As shown in Figure 4, we observe that curves obtained by changing the two parameters follow similar pattern. When the number increases, *P@1* first gradually grows, reaching the highest at position 5 and drops gradually after that. When *K* and *M* are relatively high, say larger than 3, our model can always outperform *VAE* which is the most competitive baseline, indicating the effectiveness of the discrete representation for interactive arguments identification.

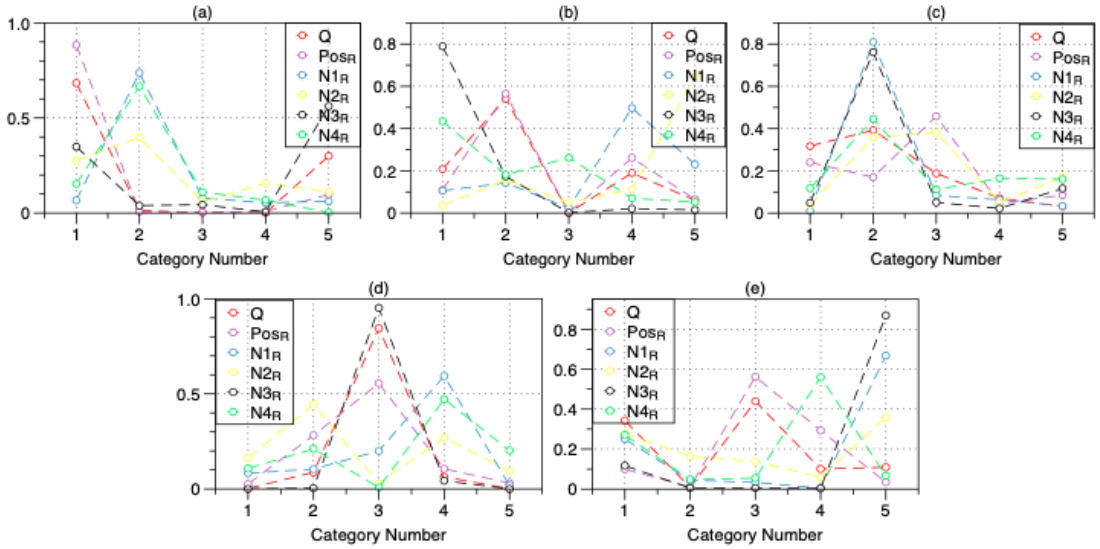

Figure 5: Visualization of posterior distributions of discrete latent variables $z_1 \sim z_5$ respectively. We find that the posterior distributions of $z_1 \sim z_5$ of *Positive reply* is more similar to those of *Quotation* compared to other *Negative replies*.

5.3 Discrete Latent Variables Analysis

Here, we try to find out why *DVAE* performs best on interactive argument pair identification. Given an argument, we set *M*=5, *K*=5 and learn the corresponding discrete code set $Z_{code}(1) \sim Z_{code}(5)$. We use the best model to select correct instances for argument matching in the dataset and cluster all quotations and corresponding replies according to the same discrete code set. We get 2,272 clusters, of which 119 clusters have more than 100 arguments and we find that arguments with the same discrete code set are semantically related.

To show the reason why *DVAE* performs well on our task more intuitively, we select a case from our dataset shown in Table 3 and employ *DVAE* to learn discrete representations for arguments to capture varying aspects $z_1 \sim z_5$. The posterior distributions of discrete latent variables $z_1 \sim z_5$ for the quotation and replies are shown in Figure 5.

As shown in Figure 5, each subgraph shows the distribution of z_i on *K* categories of the quotation and corresponding replies. We can find that the posterior distributions of $z_1 \sim z_5$ of *Positive reply* are more similar to those of *Quotation* compared to other *Negative replies*. This finding proves that if the two arguments are more semantically related, their posterior distribution on each aspect z_i should be more similar. This further interprets why *Positive reply* has interactive relationship with *Quotation* and why *DVAE* performs well on interactive argument pair identification.

Quotation: I bet that John Boehner would deal with congress as president more easily than Joe Biden due to his constant interaction with it.
Positive reply: Do you think that have anything to do with the fact that Boehner is a republican, and congress is controlled by republicans? **Negative reply 1:** I would propose that the title of vice president be kept, but to remove their right to succession for presidency. **Negative reply 2:** Does Biden have the same level of respect from foreign nations needed to guide the country? **Negative reply 3:** He did lose however, so perhaps people do put weight into the vp choice. **Negative reply 4:** I don't know why you think this can be ignored.

Table 3: A case selected from our dataset.

5.4 Error Analysis

Here, we inspect outputs of our model to identify major causes of errors. Here are two major issues.

- The number of *M* and *K* may not cover the latent space of all arguments in the dataset. Natural language is complex and diverse. If the size of the latent space doesn't fully contain semantic information of the arguments, it will cause the failure of our model. Considering the number of aspects may vary for different topics, it is not perfect to use a universal setting of *K* and *M*.
- Attention Error. In our model, we employ a quotation-guided attention to evaluate the relevance between each word in the reply and the discrete representation of the quotation. If the attention focuses on unimportant words, it causes errors. It might be useful to utilize discrete representation to further regulate the attention procedure.

Models	Pairwise accuracy
Tan et al. (2016)	65.70
Ji et al. (2018)	70.45
Our model	84.50

Table 4: The performances of different models for the task of argumentative comments persuasiveness evaluation on the dataset in Tan et al. (2016). Numbers for the two comparative models are copied from their original papers.

5.5 Effectiveness on Argumentative Comments Persuasiveness Evaluation

To further investigate the usefulness of our learned representations, we apply them to a downstream task: persuasiveness evaluation for argumentative comments (Tan et al., 2016; Ji et al., 2018). It takes two arguments as input (one is original and another is a reply) and output a score to evaluate the quality of the reply. The reasons for choosing this task are two fold. First, both tasks focus on dialogical arguments. Second, both tasks can be formulated as a pairwise ranking problem. The performance of different models are shown in Table 4. Note that we use the original *CMV* dataset and follow the previous setup in Tan et al. (2016); Ji et al. (2018). We find that our model outperforms the state-of-the-art method (Ji et al., 2018) by a large margin, which indicates that our learned representation can well help downstream tasks.

6 Related Work

In this section, we will introduce two major areas related to our work, which are dialogical argumentation and argument representation learning.

6.1 Dialogical Argumentation

Computational argumentation is a growing subfield of natural language processing in which arguments are analyzed in various respects. Previous works mainly focus on analyzing the argumentative structure in texts. Recently, the dialogical argumentation has become an active topic.

Dialogical argumentation refers to a series of interactive arguments related to a given topic, involving argument retraction, view exchange, and so on. Existing research covers discourse structure prediction (Liu et al., 2018), dialog summarization (Hsueh and Moore, 2007), etc. There are several attempts to address tasks related to analyzing the relationship between arguments (Wang and Cardie, 2014; Persing and Ng, 2017) and evaluating the quality of persuasive arguments (Habernal and Gurevych, 2016).

Gottipati et al. (2013) use sentiment lexicons as a preprocessing step and propose a probabilistic graphical model to predict stance of arguments in their dataset. Park et al. (2011) design several argumentation-motivated features to finish the debate stance classification in Korean newswire discourse. Sridhar et al. (2015) consider the joint stance classification of arguments and relations among them and find a multi-level model will work better. for a combination of post-level and author-level collective modeling of both stance and disagreement could bring further improvements in performance.

Wang and Cardie (2014) create a dispute corpus from Wikipedia and use a sentiment analysis to predict the dispute label of arguments. Wei et al. (2016) collect a dataset from *CMV* and analyze the correlation between disputing quality and disputation behaviors. analyze the disputation action in the online debate. Given an original argument and an argument disputing it, they aims to evaluate the quality of a disputing comment based on the original argument and the discussed topic. Habernal and Gurevych (2016) crowdsource the UKPConvArg1 corpus to study what makes an informal social media argument convincing. They crowdsource the UKPConvArg1 corpus and use SVM and bidirectional LSTM to experiment on their annotated datasets. Tan et al. (2016) pay attention to belief change in the ChangeMyView subreddit, in which an original poster challenges others to change his/her opinion. They construct datasets from CMV and employ logistic regression to predict which reply in the pair is more persuasive. In addition, Persing and Ng (2017) annotate a corpus with persuasiveness scores and the errors they contain to analyze why arguments are unpersuasive.

Previous work mainly focuses on analyzing interactions between two arguments in debate. However, there is limited research on the interactions between posts. In this work, we propose a novel task of identifying interactive argument pairs from argumentative posts to further understand the interactions between posts. Our work is also related with some similar tasks, such as question answering and sentence alignment. They focus on the design of attention mechanism to learn sentence representations (Wang et al., 2017a) and their relations with others (Wang et al., 2017b). Our task is inherently

different from theirs because our target arguments naturally occur in the complex interaction context of dialogues, which requires additional efforts for understanding the discourse structure therein.

6.2 Argument Representation Learning

Argument representation learning for natural language has been studied widely in the past few years. Previous work discuss prior approaches to learning argument representations from labelled and unlabelled data.

There have been attempts to use labeled/structured data to learn argument representations. Wieting et al. (2016) and Wieting and Gimpel (2017) introduce a large sentential paraphrase dataset and use paraphrase data to learn an encoder that maps synonymous phrases to similar embeddings. Wieting et al. (2017) explore the use of machine translation to obtain more paraphrase data via back-translation of bilingual argument pairs for learning paraphrastic embeddings. They show how neural backtranslation could be used to generate paraphrases. Hermann and Blunsom (2013) explore a language-specific encoder applied to each argument and represent the argument by the mean vector of the words involved. They consider minimizing the inner product between paired arguments in different languages as the training objective and do not rely on word alignments. Conneau et al. (2017) propose a model called InferSent, which is used as the baseline as it served as the inspiration for the inclusion of the SNLI task in the multitask model. They prove that NLI is an effective task for pre-training and transfer learning in obtaining generic argument representations. They train argument encoders from identifying one of three relationships between two given arguments - entailment, neutral and contradiction. Results prove that the argument representations learned by this task perform strongly on downstream transfer tasks.

Due to the availability of practically unlimited textual data, learning argument representations via unsupervised methods is an attractive proposition. Kiros et al. (2015) present the model called Skip Thought for learning representations by predicting the previous and next argument, which is a generalization of the skip-gram model (Mikolov et al., 2013). Exploiting the relatedness inherent in adjacent arguments, the model is trained by using the encoder to encode a particular argument and then using the decoder to decode words in adjacent arguments. Bowman et al. (2016) introduce variational autoencoders to incorporate distributed latent representations of entire arguments. In addition, Hill et al. (2016) propose the FastSent model, using bag-of-words of arguments to predict the adjacent arguments. Logeswaran and Lee (2018) propose the Quick Thoughts to exploit the closeness of adjacent arguments. They formulate the argument representation learning as a classification problem.

Previous work focuses on learning continuous argument representations with no interpretability. In this work, we study the discrete argument representations, capturing varying aspects in argumentation languages.

7 Conclusion and Future Work

In this paper, we propose a novel task of interactive argument pair identification from two posts with opposite stances on a certain topic. We examine contexts of arguments and induce latent representations via discrete variational autoencoders. Experimental results on the dataset show that our model significantly outperforms the competitive baselines. Further analyses reveal why our model yields superior performance and prove the usefulness of discrete argument representations.

The future work will be carried out in two directions. First, we will study the usage of our model for applying to other dialogical argumentation related tasks, such as debate summarization. Second, we will utilize neural topic model for learning discrete argument representations to further improve the interpretability of representations.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (No.71991471), Science and Technology Commission of Shanghai Municipality Grant (No.20dz1200600). Jing Li is supported by CCF-Tencent Rhino-Bird Young Faculty Open Research Fund (R-ZDCJ), the Hong Kong Polytechnic University internal funds (1-BE2W and 1-ZVRH), and NSFC Young Scientists Fund 62006203.

References

Christa SC Asterhan and Baruch B Schwarz. 2007. The effects of monological and dialogical argumentation on concept learning in evolutionary theory. *Journal of educational psychology*, 99(3):626.

Yonatan Bilu and Noam Slonim. 2016. Claim synthesis via predicate recycling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 525–530, Berlin, Germany. Association for Computational Linguistics.

Teresa Botschen, Daniil Sorokin, and Iryna Gurevych. 2018. Frame- and entity-based knowledge for common-sense argumentative reasoning. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 90–96. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 10–21. ACL.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Rich Caruana, Steve Lawrence, and C. Lee Giles. 2000. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 402–408. MIT Press.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 670–680. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 153–162. Association for Computational Linguistics.

Swapna Gottipati, Minghui Qiu, Yanchuan Sim, Jing Jiang, and Noah A. Smith. 2013. Learning topics and positions from debatepedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1858–1868. ACL.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Karl Moritz Hermann and Phil Blunsom. 2013. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377. Association for Computational Linguistics.

Pei-yun Hsueh and Johanna D. Moore. 2007. Automatic decision detection in meeting speech. In *Machine Learning for Multimodal Interaction , 4th International Workshop, MLMI 2007, Brno, Czech Republic, June 28-30, 2007, Revised Selected Papers*, volume 4892 of *Lecture Notes in Computer Science*, pages 168–179. Springer.

Anthony Hunter. 2013. Analysis of dialogical argumentation via finite state machines. In *International Conference on Scalable Uncertainty Management*, pages 1–14. Springer.

Lu Ji, Zhongyu Wei, Xiangkun Hu, Yang Liu, Qi Zhang, and Xuanjing Huang. 2018. Incorporating argument-level interactions for persuasion comments evaluation using co-attention model. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3703–3714. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3294–3302.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2018. Discourse representation structure parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 429–439. Association for Computational Linguistics.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*, pages 314–324.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2786–2792. AAAI Press.

Souneil Park, Kyung Soon Lee, and Junehwa Song. 2011. Contrasting opposing views of news articles on contentious issues. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 340–349. The Association for Computer Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4082–4088. ijcai.org.

Jason Tyler Rolfe. 2017. Discrete variational autoencoders. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Nattapong Sanchan, Ahmet Aker, and Kalina Bontcheva. 2017. Automatic summarization of online debates. In *Proceedings of the 1st Workshop on Natural Language Processing and Information Retrieval associated with RANLP 2017*, pages 19–27. INCOMA Inc.

Dhanya Sridhar, James R. Foulds, Bert Huang, Lise Getoor, and Marilyn A. Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 116–125. The Association for Computer Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1882–1891. The Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 613–624. ACM.

Lu Wang and Claire Cardie. 2014. A piece of my mind: A sentiment analysis approach for online dispute detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–699. Association for Computational Linguistics.

Nan Wang, Jin Wang, and Xuejie Zhang. 2018. Ynuhpcc at semeval-2018 task 2: Multi-ensemble bi-gru model with attention mechanism for multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 459–465. Association for Computational Linguistics.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017a. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 189–198. Association for Computational Linguistics.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017b. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4144–4150. ijcai.org.

Zhongyu Wei, Yandi Xia, Chen Li, Yang Liu, Zachary Stallbohm, Yi Li, and Yang Jin. 2016. A preliminary study of disputation behavior in online debating forum. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 166–171, Berlin, Germany. Association for Computational Linguistics.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

John Wieting and Kevin Gimpel. 2017. Revisiting recurrent networks for paraphrastic sentence embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 2078–2088. Association for Computational Linguistics.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 274–285. Association for Computational Linguistics.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1098–1107. Association for Computational Linguistics.