Incorporating Argument-Level Interactions for Persuasion Comments Evaluation using Co-attention Model

Lu Ji[1] **, Zhongyu Wei**[2*]**, Xiangkun Hu**[1]**, Yang Liu**[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]Liulishuo Company
{17210240034,zywei,xkhu17,qi_zhang,xjhuang}@fudan.edu.cn
yang.liu@liulishuo.com

Abstract

In this paper, we investigate the issue of persuasiveness evaluation for argumentative comments. Most of the existing research explores different text features of reply comments on word level and ignores interactions between participants. In general, viewpoints are usually expressed by multiple arguments and exchanged on argument level. To better model the process of dialogical argumentation, we propose a novel co-attention mechanism based neural network to capture the interactions between participants on argument level. Experimental results on a publicly available dataset show that the proposed model significantly outperforms some state-of-the-art methods for persuasiveness evaluation. Further analysis reveals that attention weights computed in our model are able to extract interactive argument pairs from the original post and the reply.

1 Introduction

Computational argumentation is a growing field in natural language processing. Existing research covers argument unit detection (Al-Khatib et al., 2016), argument structure prediction (Peldszus and Stede, 2015; Stab and Gurevych, 2014), argumentation scheme classification (Feng et al., 2014), etc. Recently, the automatic assessment of argumentation quality has started gaining attention. It can be analyzed at two levels, namely monological argumentation and dialogical argumentation.

Monological argumentation refers to a composition of arguments on a certain issue (Wachsmuth et al., 2017). A typical example of quality evaluation for monological argumentation is automated essay scoring, which aims to process argumentative essays without human interference (Taghipour and Ng, 2016). It takes an essay as the input and outputs a numeric score, considering features of content, grammar, discourse structure and lexical richness (Burstein et al., 2013). Most of the efforts are made on the exploration for better document representation.

Dialogical argumentation refers to a series of interactive arguments related to a given topic, involving argument retraction, view exchange, and so on (Besnard et al., 2014). With the popularity of online debating forums like *convinceme* [1], *debatepedia* [2] and *change my view (CMV)* [3], researchers pay increasing attention to evaluating the quality of debating or persuasive content (Tan et al., 2016; Wei and Liu, 2016; Wei et al., 2016; Habernal and Gurevych, 2016a). Although some similarity features of content between the reply and the original post are used to evaluate the quality of the given reply, they are computed on word level without considering the exchange of opinions on the basis of arguments.

An example of dialogical argumentation is shown in Figure 1. There are three posts, one is original and the other two are replies. The repliers post to change the view of the original poster. And the *positive reply* is deemed to be more persuasive than the *negative reply*. We have three observations. First, viewpoints of the original poster and repliers are expressed via multiple arguments. Second, content of replies is organized in-line with arguments in the original post. Third, interactions between a reply and the

*Corresponding author
This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

[1]http://convinceme.net.
[2]http://debatepedia.idebate.org.
[3]https://reddit.com/r/changemyview

Proceedings of the 27th International Conference on Computational Linguistics, pages 3703–3714
Santa Fe, New Mexico, USA, August 20-26, 2018.

Original Post: Philosophy doesn't seem to have any practical applications. [What value does philosophy have in the modern age, right now, aside from contemplating things?] I have read the argument that it is impossible to argue that philosophy is useless without using philosophy. [What do you gain from studying philosophy that could not be gained from thoughtful introspection?]	
Positive Reply	**Negative Reply**
What do you gain from studying philosophy that could not be gained from thoughtful introspection? [Two answers. #1 rigor and #2 it saves us from reinventing the wheel.] [Why do you think we should start from scratch in all value decisions rather than seeking to understand the work that has been done in the past?]	What do you gain from studying philosophy that could not be gained from thoughtful introspection? [Ask yourself the same question about math.] Your argument seems to be that studying philosophy is a waste of time because it has no practical use.

Figure 1: An example of dialogical argumentation consists of one original post and two persuasion replies from *change my view*, a sub-forum of *Reddit.com*. Different types of underlines are used to highlight the interactive relationship. The labels of the positive and negative replies are assigned by the original poster.

original post provide some indications for the persuasive comment identification. Inspired by the three findings, we aim to analyze dialogical argumentation on argument level and explore how argument-based interactions can help persuasiveness evaluation. Our datasets are collected from an on-line forum [4]. The content of posts is usually informal and not strictly grammatical. It is extremely difficult to parse the argument in a finer-grain with premise, conclusion and other components. Therefore, we treat each sentence as an argument for simplicity.

In this paper, we propose to incorporate argument-level interactions within dialogical argumentation for better persuasion comments quality evaluation. We propose a novel framework that includes three components, namely, argument representation, co-attention network and aggregation network. We first learn two different representations for each argument via a hierarchal neural network. Co-attention network captures the interactions between the reply and the original post on argument level via three kinds of attentions. Aggregation network combines the results of the co-attention network. Finally, a persuasiveness score is assigned to the target comment using a linear transformation. Experimental results on a benchmark dataset show that with the assistance of argument-level interactions, our proposed model can achieve much better performance than some state-of-the-art methods. In order to further understand how attention mechanism works for capturing the argument-level interactions, we formalize a task of interactive argument pair extraction. Experimental results on a self-constructed dataset show that our attention-based strategy significantly outperforms a word-overlap based strategy for the identification of interactive arguments.

2 Proposed Model

Given an original post and two corresponding replies, our task is to automatically identify which reply is more persuasive. In practice, we evaluate the quality of the two replies separately given the original post and treat the one with higher persuasiveness score as the winner. The overall architecture of our model is shown in Figure 2. It takes an original post and a reply as inputs, and outputs a real value as its persuasiveness score. It mainly consists of three components, namely, *Argument Representation*, *Co-attention Network* and *Aggregation Network*. We learn two representations for each single argument. One is based on its internal words and the other considers information of context arguments. What's more, three types of attentions are proposed to model the detailed interactions between the original post and the reply on argument level. The *Aggregation Network* integrates the results of *Co-attention Network*.

[4]Datasets are available at : https://github.com/lji0126/Persuasion-Comments-Evaluation

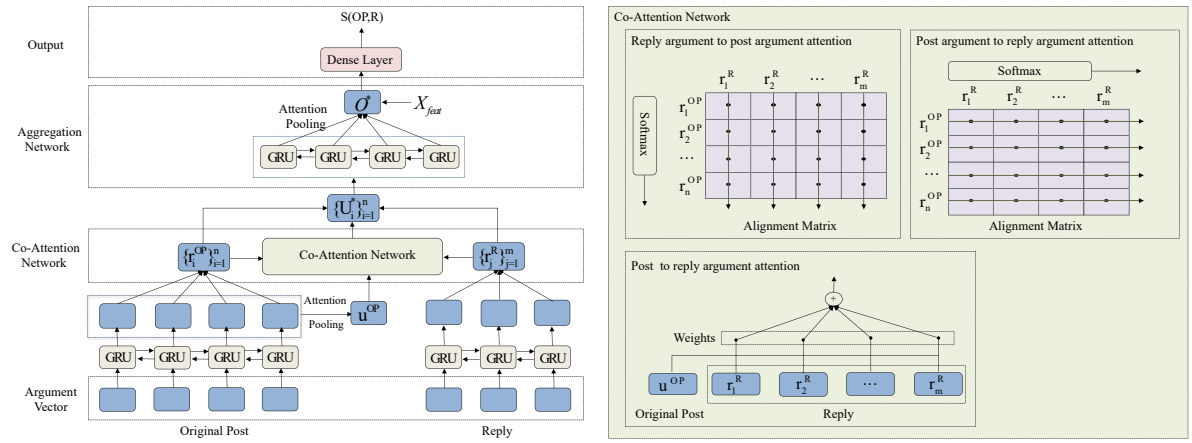

Figure 2: Overall architecture of the proposed model. The left part is the main framework of this work. The right part is the detailed structure of the co-attention network.

2.1 Argument Representation

Inspired by Dong et al. (2017), we employ a hierarchical architecture to obtain two different representations for each single argument. For simplicity, we consider each sentence as an argument.

Representation based on internal words given an argument with words $w_1, w_2, ..., w_T$, we first map each word to a dense vector obtaining $x_1, x_2, ..., x_T$ correspondingly. We then employ a convolution layer to incorporate the context information on word level.

$$z_i = f(W_z \cdot [x_i : x_{i+h_w-1}] + b_z) \tag{1}$$

where W_z and b_z are weight matrix and bias vector. h_w is the window size in the convolution layer and z_i is the feature representation. Not all words contribute equally to the representation of the argument. Therefore, we conduct an attention pooling operation over all the words to get the contribution of each word.

$$m_i = \tanh(W_m \cdot z_i + b_m) \tag{2}$$

$$u_i = \frac{e^{W_{u \cdot m_i}}}{\sum_j e^{W_{u \cdot m_j}}} \tag{3}$$

$$a = \sum_i u_i \cdot z_i \tag{4}$$

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.

Representation incorporating context arguments in order to incorporate information of context arguments, we employ a bi-directional GRU (Cho et al., 2014) to get the representation of an argument under its context. A BiGRU consists of both forward and backward GRU that handle the sequence from the left and the right end, respectively. In practice, we concatenate the hidden states of two GRUs for each argument to get final argument representation. $r_i^{OP} = BiGRU(r_{i-1}^{OP}, a_i^{OP}) \in R^{n \times d}, i \in [1 \cdots n]$, $r_j^R = BiGRU(r_{j-1}^R, a_j^R) \in R^{m \times d}, j \in [1 \cdots m]$, where a_i^{OP}, a_j^R are representations of arguments in the original post and the reply, respectively. r_i^{OP}, r_j^R are the hidden state of the i-th argument in the original post and the j-th argument in the reply. d is the dimension of hidden units. n and m stand for the number of arguments for the original post and the reply, respectively.

2.2 Co-attention Network

Attention mechanism is a common way to link and fuse information from two content-related texts (Weston et al., 2014; Hill et al., 2016; Sordoni et al., 2016; Shen et al., 2017). In order to capture the

interactions between the original post and the reply, we propose a novel co-attention network that includes three kinds of attentions. In particular, *Post argument to reply argument attention* computes the relevance of each post argument with every reply argument and helps learn a series of new reply representations. *Reply argument to post argument attention* computes the relevance of each reply argument with every post argument and finally obtains a series of new post representations. *Post to reply argument attention* computes the relevance of each reply argument with the entire post argument which contributes to learn a new reply representation.

Post argument to reply argument attention We first compute the alignment matrix $A \in R^{n \times m}$ that contains similarity scores corresponding to all possible pairs of arguments between the original post and the reply via Equation 5.

$$A_{ij} = W_a^T [r_i^{OP}; r_j^R; r_i^{OP} \circ r_j^R] \tag{5}$$

where W_a is the weight parameter, $\circ$ is the Hadamard product. A_{ij} indicates the similarity between i-th argument in the original post and j-th argument in the reply.

For i-th argument in the original post, we could signify which arguments in the reply are relevant to it by this attention. Similar to in Seo et al. (2016), we normalize the alignment matrix A row-wise to produce the attention weights across the reply for each argument in the original post. The calculation is described in Equation 6.

$$V_i = softmax(A_{i:}), i \in [1 \cdots n] \tag{6}$$

Based on the attention probability V_i of i-th argument in the original post, the new representation of the reply can be calculated by Equation 7.

$$U_i^1 = \sum_t V_{it} \cdot r_t^R \tag{7}$$

Based on all the arguments in the original post, we could obtain a series of new reply representations, which constitute U^1.

Reply argument to post argument attention We normalize the alignment matrix A column-wise to produce the attention weights across the original post for each argument in the reply. The attention weights can be calculated by Equation 8.

$$Q_j = softmax(A_{:j}), j \in [1 \cdots m] \tag{8}$$

Subsequently, each attended argument representation of original post is shown in Equation 9.

$$U_j^2 = \sum_t Q_{jt} \cdot r_t^{OP} \tag{9}$$

Based on all the arguments in the reply, we could get a series of new post representations, which constitute U^2.

Post to reply argument attention In order to evaluate the importance of arguments in the reply, we propose this attention. Firstly, we learn a representation u^{OP} for the original post via applying the attention pooling operation over all its hidden states $r_i^{OP}, i \in [1 \cdots n]$. We then compute attention weights with the original post based on u^{OP} for each argument in the reply. In practice, we conduct dot product between u^{OP} and each hidden state representation r_j^R in the reply and a softmax layer is used to obtain an attention distribution. The calculation process is shown in Equation 10.

$$v_j = softmax(u^{OP} \cdot r_j^R) \tag{10}$$

Based on the attention probability v_j of the j-th argument in the reply, the new representation of the reply can then be constructed as Equation 11.

$$u^3 = \sum_j v_j \cdot r_j^R \tag{11}$$

Finally, we put all of the attention representations in a linear function to get the integrated information. The detail is illustrated in Equation 12.

$$U = f(U^1, U^2, U^3, \{r_i^{OP}\}_{i=1}^n, \{r_j^R\}_{j=1}^m) \tag{12}$$

where f is a simple linear function, U^3 is a matrix that is tiled n times by u^3.

2.3 Aggregation Network

After acquiring the local alignment representation by the co-attention network, we employ a filtration gate to hold the interactive information. Then, we fuse the interactive information via a bi-directional GRU and compute the persuasiveness score.

Filtration Gate We utilize the filtration gate (Wang et al., 2017b) to hold the information that helps to understand the argument-level interactions between the original post and the reply. The formulas are in Equation 13 and 14.

$$gt = sigmoid(W_g U + b) \tag{13}$$

$$U^* = gt \odot U \tag{14}$$

We fuse the interactive information reserved by the filtration gate via a bi-directional GRU. The calculation is described in Equation 15.

$$O_t = BiGRU(O_{t-1}, U_t^*) \tag{15}$$

Then, we use an attention pooling operation over the whole hidden states of this BiGRU to summarize the interactive features into a dense vector O^*.

Scoring Tay et al. (2017) prove that adding some manual features such as word-overlap to models is helpful for improving performance. We incorporate some word-overlap features X_{feat} to the proposed model, i.e. similarities between original post and the reply in terms of word-overlap. Then, we use two fully connected layers to obtain a higher-level representation r. Finally, the persuasiveness score S is obtained by a linear transformation via Equation 16 and 17 .

$$r = f(W_r[O^*; X_{feat}] + b_r) \tag{16}$$

$$S = W_s r + b_s \tag{17}$$

where W_r and W_s stand for the weight matrices, while b_r and b_s are weight vectors.

2.4 Loss Function and Training

Given an original post and two corresponding replies, we want to automatically identify which reply is more persuasive. We formalize this issue as a ranking task and utilize a pairwise hinge loss for training. Given a triple (OP, R^+, R^-), where R^+ and R^- respectively denote the positive and the negative reply for OP. The loss function is defined in Equation 18.

$$L = max(0, 1 - S(OP, R^+) + S(OP, R^-)) \tag{18}$$

where $S(OP, R^+)$ and $S(OP, R^-)$ are the corresponding persuasiveness scores.

The model is trained by stochastic gradient descent on 105 epochs, and evaluated on the development set at every epoch to select the best model. Dropout (Srivastava et al., 2014) has proved to be an effective method and is used in our work. We use Glove (Pennington et al., 2014) word embeddings, which are 50-dimension word vectors trained with a crawled large corpus with 840 billion tokens. Embeddings for words not present are randomly initialized with sampled numbers from a uniform distribution [-0.25,0.25]. We set initial learning rate to 0.1, batch size to 20, filter sizes to 5, filter numbers to 100 and the hidden unit of BiGRU to 200. Early stopping was used with a patience of 15 epochs. We implemented our model using TensorFlow. The model converged in 23 hours on an NVIDIA Titan X machine.

	Training Set				Test Set			
	Ave_w	Var_w	Ave_p	Var_p	Ave_w	Var_w	Ave_p	Var_p
Original post	10	49.5	14	163.7	11	53.2	15	133.7
Positive reply	10	46.3	14	125.0	10	44.1	13	123.8
Negative reply	10	39.2	11	82.0	10	44.7	10	69.5

Table 1: Statistics of the evaluation dataset. Ave_w represents the average number of words per argument. Ave_p represents the average number of arguments per post. Var_w indicates the variance of the number of words per argument. Var_p indicates the variance of the number of arguments per post.

3 Experiments

3.1 Dataset and Metric

We use the same dataset as Tan et al. (2016) for evaluation, which focuses on arguments from root reply. The dataset is collected from the */r/ChangeMyView* subreddit (CMV). 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' opinions. Users can give *delta* to a reply if it changes their original mind about the topic. In this dataset, for the same original post, the reply with *delta* is treated as *positive reply*, otherwise it is chosen as the *negative reply*.

The whole dataset consists of 3,456 training instances and 807 testing instances, where each instance includes an original post with one positive and one negative reply respectively. We randomly select 10% of the training instances to form the development set. In preprocessing, we use NLTK [5] for tokenization and lowercase conversion. We also filter out stop words and low frequency words. The constructed word vocabulary contains 15,767 distinct words. The detailed statistics are shown in Table 1.

Since we treat this task as a pairwise ranking problem, pairwise accuracy is conducted as the evaluation metric, which also mentioned in Tan et al. (2016).

3.2 Models for Comparing

We compare our model with the previous state-of-the-art model and the variant models of our model.
– **Tan et al. (2016)**: Tan et al. (2016) regard this task as a binary classification problem and use logistic regression model to classify replies based on some manually designed features, including interplay features, argument-related features and text style features. Because there is no source code published, we directly present the result reported in their paper for comparison.
– **Word-level BiGRU (WB)**: This model employs BiGRU to encode the original post and the corresponding reply on word level. Both representations of the original post and the reply are then concatenated to compute its persuasiveness score using a fully connected layer.
– **CNN + BiGRU (CB)**: This model encodes the original post and the corresponding reply via a hierarchical neural network. All the hidden states from the BiGRU are input into the aggregation network to compute the persuasiveness score. This is a part of our model without the co-attention network and the word-overlap features.
– **Word Overlap Features (WOF)**: This model directly uses the word-overlap features to evaluate the quality of arguments. More concretely, word-overlap features contain Jaccard similarity and some scores based on common words between the original post and the reply.
– **CNN+BiGRU+Co-Att (CBCA)**: This model is a part of our model without the word-overlap features. We input argument representations of the original post and the corresponding reply into the co-attention network and then obtain the persuasiveness score via the results of aggregation network.
– **CNN + BiGRU + Word Overlap Features (CBWOF)**: This model is a part of our model without the co-attention network. After obtaining argument representations of the original post and the corresponding reply, we input the hidden states of BiGRU into the aggregation network. We then concatenate the result of aggregation network with the word-overlap features to get the persuasiveness score.

[5] http://www.nltk.org/

Model	Pairwise accuracy
Tan et al. (2016)	<u>65.70</u>
Word-level BiGRU (WB)	61.22
CNN+BiGRU (CB)	63.34
Word Overlap Features (WOF)	63.59
CNN+BiGRU+Co-Att (CBCA)	66.96‡
CNN+BiGRU+Word Overlap Features (CBWOF)	68.08‡
CNN+BiGRU+Att_III+Word Overlap Features (CBAWOF_III)	69.95‡
CNN+BiGRU+Att_I+Word Overlap Features (CBAWOF_I)	70.07‡
CNN+BiGRU+Att_II+Word Overlap Features (CBAWOF_II)	70.20‡
CNN+BiGRU+Co-Att+Word Overlap Features (CBCAWOF)	**70.45**‡*

Table 2: The performance of different approaches on our datasets. The model underlined is the state-of-the-art method. The models that outperform the state-of-the-art method are highlighted with ‡. Our model that significantly outperforms the state-of-the-art method is marked with * (p<0.01, Student's paired t-test and Wilcoxon signed rank test). Best result is in **bold**.

– **CNN+BiGRU+Att_I+Word Overlap Features (CBAWOF_I)**: This model is a part of our model with the post argument to reply argument attention in the co-attention network.
– **CNN+BiGRU+Att_II+Word Overlap Features (CBAWOF_II)**: This model is a part of our model with the reply argument to post argument attention in the co-attention network.
– **CNN+BiGRU+Att_III+Word Overlap Features (CBAWOF_III)**: This model is a part of our model with the post to reply argument attention in the co-attention network.
– **CNN + BiGRU + Co-Att+Word Overlap Features (CBCAWOF)**: This is our proposed model.

3.3 Results and Discussion

The overall result of the comparison is shown in Table 2. We have following findings.
– The model proposed by Tan et al. (2016) achieves an accuracy of 65.70 % on the dataset. The performance actually shows the effectiveness of human generated features on this task. However, some writing style features used are very difficult to obtain, limiting its ability to generalize. The authors also explore to use interactive features between the original post and the reply, however, only word-level text similarity is considered.
– The performance of *CB* is much better than that of *WB*. This proves the effectiveness of representing posts on argument level instead of word level.
– The performance of *WOF* is comparable to that of *CB*. This indicates that correlation between the original post and the reply is an important feature for persuasiveness evaluation.
– The performance of *CBCA* is much better than that of *CB*. This indicates the effectiveness of the co-attention network.
– By combining both argument representations and word-overlap features, the performance of *CBWOF* is significantly better than that of *CB* and *WOF*.
– The performance of *CBAWOF_I*, *CBAWOF_II*, *CBAWOF_III* is better than that of all models except *CBCAWOF*. This proves the effectiveness of the three kinds of attentions separately.
– Our proposed model *CBCAWOF* generates the best performance among all the models. This confirms the effectiveness of our proposed co-attention model.

In order to further prove the effectiveness of our model, we conduct the *Student's paired t-test* [6] and

[6]https://en.wikipedia.org/wiki/Student's_t-test.

Original Post: [I'm talking about making the human race smarter, forever.] [We could use the IQ scale (for want of a better intelligence measure) to determine the number of offspring a person should be able to genetically contribute to.] [A man and a woman with average IQ can have two children and average IQs of 125-174 can contribute towards 3 children.] This would make human more likely to survive.

Positive Reply
[When the proposal comes up I am reminded of why it's a bad idea: not because we couldn't do it, but because we don't know how to do it right.][Why would a more intelligent society automatically be a better one?] We don't know enough about the biology. [Simply, we are not ready for eugenics.]

Negative Reply:
[As you put it, the entire goal of eugenics is to make the human race to advance and survive, but the irony of eugenics is that one of the best ways we can guarantee our survival is to maximize the size of our gene pool.] However, it's dangerous to call certain genes good and other genes bad.

Figure 3: A sample consists of one original post with two persuasion replies. Interactive argument pairs are highlighted with the same kind of underline. Pairs are extracted via co-attention network.

the *Wilcoxon signed rank test* [7]. Because Tan et al. (2016) don't publish their source code, we are unable to obtain detail results of their model. In Table 2, we find that the performance of CBWOF is better than that of Tan et al. (2016), so we carry out the significance tests between the results of our model and CBWOF. The p-value of the two significance tests is less than 0.01 respectively, which proves that our model significantly outperforms the state-of-the-art method.

4 Further Analysis on Co-attention Network

In order to further understand the capability of our co-attention network for capturing interactions between the original post and the reply on argument level, we perform an additional experiment to evaluate the effectiveness of the attention weights for the identification of interactive argument pairs. Therefore, we propose a novel task of extracting the interactive argument pairs between the original post and the corresponding reply. We first build an evaluation dataset and then compare the extraction performance of attention-based extraction strategy with a word-overlap based strategy on this dataset.

4.1 Dataset for Interactive Argument Pair Extraction

We sample 50 triples in the form of (original post, positive reply, negative reply) from the training set and split these into 100 original post-reply pairs in the form of (original post, positive reply) and (original post, negative reply). Given two collections of arguments from the original post and the reply, namely, $OP = \{op_1, op_2, op_3, ..., op_n\}$ and $R = \{r_1, r_2, r_3, ..., r_m\}$, for each argument r_j in the reply, we aim to identify arguments in the original post that interact with it.

Two annotators are hired to annotate the dataset independently and a third annotator is asked to solve the conflict between the two annotators. Two annotators identify 371 and 355 pairs of interactive arguments respectively. The inter-annotator agreement measured by Co-hens Kappa (Carletta et al., 1996) is 91.83%. With the final decision from the third annotator, we obtain 365 pairs in total. In detail, 234 interactive argument pairs come from positive replies, and the other 131 pairs are generated by negative replies. This re-confirms that the degree of interaction is a good indicator for persuasive reply identification.

4.2 Automatic Argument Extraction

We use two methods to extract argument pairs automatically. One is based on the attention weights computed in our proposed model and the other extracts pairs based on word-overlap similarity.

[7] https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test.

	P@1	P@2	P@3	P@4	P@5	MRR
WS	17.53	30.41	39.72	48.49	53.97	31.33
CN	22.19	36.71	43.84	49.86	54.24	39.41

Table 3: Experimental results of *WS* and *CN* for interactive argument pair extraction on the self-constructed dataset.

Co-attention Network (CN): We extract interactive argument pairs based on the results of our co-attention network. Because not every argument in the reply has interactive relationship, we choose 10 arguments in each reply based on the top-10 weights of the *post to reply argument attention* (this attention vector computes the importance of arguments in the reply in the perspective of the original post). Secondly, we choose the top-5 arguments in the original post for the 10 arguments in reply respectively in terms of the *reply argument to post argument attention* weights.

Word-overlap Similarity (WS): The extraction of arguments in the reply is the same as *CN*. We use the number of common words to identify interactive arguments in the original post for each argument in the reply. Top-5 arguments obtaining most common words with the argument in the reply are kept.

4.3 Results and Analysis

For each argument from the reply in the gold pair, we will see whether its corresponding argument is ranked in top k by automatic models. We report Mean Reciprocal Rank (MRR), precision at position 1, 2, 3, 4 and 5 as evaluation metrics. Table 3 shows the experiment results of *WS* and *CN*. We can see great differences between the two methods. In detail, the performance of CN in terms of P@1$\sim$P@5 and MRR is higher than those of WS by 4.66%, 6.30%, 4.12%, 1.37%, 0.27% and 8.08% respectively. This confirms the effectiveness of our co-attention network for capturing interactions between the original post and the reply. However, the overall performance of both *CN* and *WS* are relatively low. This indicates that the task is difficult in nature. A sample of interacting argument pairs extracted by the attention-based approach can be seen in Figure 3.

5 Related Work

Two major areas related to our work are argumentation quality evaluation and attention mechanism.

5.1 Argumentation quality evaluation

Computational argumentation is a growing sub-field of natural language processing in which arguments are analyzed in various respects. Previous works in computational argumentation mainly focus on the methods for argument mining, which arms to determine the argumentative structure in texts. Recently, argumentation quality evaluation has become an active topic in this field.

There have been several attempts to address tasks related to argumentation quality evaluation. Habernal and Gurevych (2016b) propose a new task of predicting which argument from an argument pair is more convincing and use SVM and bidirectional LSTM to experiment on their annotated datasets. Tan et al. (2016) construct datasets from the ChangeMyView subreddit. They study factors affecting whether a challenger can successfully change the view of a commenter expressed in the original post and employ logistic regression to predict which reply in the pair is more persuasive.

In addition, Wei and Liu (2016) acquire discussion threads from the ChangeMyView subreddit to study the mechanisms behind persuasion. They propose and evaluate a set of features to predict the persuasiveness of debate posts, including textual features and social interaction related features. Wei et al. (2016) propose a task for quality evaluation of disputing argument. They manually annotate a real dataset collected from an online debating forum and analyze the correlation between disputing quality and different disputation behaviors. Wang et al. (2017a) use linguistic features of arguments, latent persuasive strengths of different topics and the interactions of debate comments to predict the debate outcome. Persing and Ng (2017) study the persuasiveness by designing five respects of error that have

negative impacts on persuasiveness. They not only focus on determining how persuasive an argument is, but also tell us why an argument is unpersuasive.

From the brief descriptions given above, we can find that most of the existing research focuses on the interactions among debate comments only from the perspective of text similarity. The interactions among argument pairs are ignored. In this work, we evaluate the quality of debate comments through the interactions among them on argument level.

5.2 Attention mechanism

Attention mechanism allows models to focus on specific parts of inputs at each step of a task. Moreover, attention mechanism has been proved to be significantly effective in natural language processing tasks.

Co-attention mechanism has recently attracted lots of research interest in the fields of machine translation (Bahdanau et al., 2014), question answering (Wu et al., 2017), text generation (Li et al., 2015), etc. It is computed as an alignment matrix based on two inputs, which can model complex interactions between the two inputs. Xiong et al. (2016) present a co-attention encoder to focus on relevant parts of the representations of the question and document and use a dynamic pointing decoder to locate the answer. Cui et al. (2016) propose a two-way attention mechanism to encode the passage and question mutually and induce attended attention for final answer predictions.

Self-attention mechanism is an attention mechanism aiming at aligning the sequence with itself, which has been successfully used in a variety of tasks. In Cheng et al. (2016), both encoder and decoder are modeled as LSTMs with self-attention for extractive summarization of documents. In Lin et al. (2017), the authors conduct a self-attention over the hidden states of a BiLSTM to extract the sentence embedding. Instead of sentence vector, they use a 2-D matrix to represent the embedding, with each row of the matrix attending on a different part of the sentence.

In this work, we employ a co-attention mechanism to capture the interactions between the original post and the reply on argument level. What's more, we use a self-attention mechanism to obtain the argument representation, which is called attention pooling in the previous sections.

6 Conclusions and Future Work

In this paper, we propose to incorporate argument-level interactions for dialogical argumentation. A novel co-attention network is proposed to capture the detailed interactions between the original post and the reply on argument level for better persuasiveness evaluation. Experimental results on a benchmark dataset show that the proposed model can achieve much better performance than the previous state-of-the-art method. Further analysis of extracting interactive argument pairs from the original post and the reply also proves the effectiveness of our co-attention network.

The future work will be carried out in three directions. First, we will fully investigate the usage of our model for applying to other dialogical argumentation related tasks, such as debate summarization. Second, we will enlarge the annotated dataset for interactive argument pair identification and explore more effective methods to generate argument pairs automatically. Third, we will explore to utilize topic information for the quality evaluation of persuasion comments.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. The work is partially supported by National Natural Science Foundation of China (Grant No. 61702106), Shanghai Science and Technology Commission (Grant No. 17JC1420200, Grant No. 17YF1427600 and Grant No.16JC1420401).

References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 16)*, pages 1395–1404. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 2015 International Conference on Learning Representations*.

Philippe Besnard, Alejandro Garcia, Anthony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Simari, and Francesca Toni. 2014. Introduction to structured argumentation. 5(1):1–4.

Jill Burstein, Joel Tetreault, and Nitin Madnani. 2013. The e-rater automated essay scoring system, *Handbook of automated essay evaluation: current applications and new directions*. pages 55–67, New York, NY: Routledge

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. In *Computational linguistics*. 22(2):249–254.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2016. Attention-over-attention neural networks for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 593–602.

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)*, pages 153–162.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223.

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. Association for Computational Linguistics.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of the 2016 International Conference on Learning Representations*.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1106-1115.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the 2017 International Conference on Learning Representations*.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal, September. Association for Computational 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)*, pages 1532–1543.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4082–4088.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. In *Proceedings of the 2017 International Conference on Learning Representations*.

Yelong Shen, Po Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. Reasonet: Learning to stop reading in machine comprehension. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1047–1055.

Alessandro Sordoni, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2016. Iterative alternating neural attention for machine reading. *arXiv preprint arXiv:1606.02245*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. In *The Journal of Machine Learning Research*. 15(1):1929–1958.

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 *Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891.

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*, pages 613–624.

Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. 2017. Learning to rank question answer pairs with holographic dual lstm architecture. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 695–704.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Meeting of the Association for Computational Linguistics*, pages 189–198.

Zhongyu Wei and Yang Liu. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the Third Workshop on Argument Mining*, pages 166–171.

Zhongyu Wei, Yandi Xia, Chen LiYang Liu, Zachary Stallbohm, Yi Li, and Yang Jin. 2016. A Preliminary Study of Disputation Behavior in Online Debating Forum. In *Meeting of the Association for Computational Linguistics*, pages 195–200.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. In *Proceedings of the 2015 International Conference on Learning Representations*.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding* , 163: 21–40.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. In *Proceedings of the 2017 International Conference on Learning Representations*.