HARGAN: Heterogeneous Argument Attention Network for Persuasiveness Prediction

Kuo-Yu Huang[1], Hen-Hsen Huang[2,3], Hsin-Hsi Chen[1,3]
[1] Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[2] Department of Computer Science, National Chengchi University, Taiwan
[3] MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
kyhuang@nlg.csie.ntu.edu.tw, hhhuang@nccu.edu.tw, hhchen@ntu.edu.tw

Abstract

Argument structure elaborates the relation among claims and premises. Previous works in persuasiveness prediction seldom consider this relation in their architectures. To take argument structure information into account, this paper proposes an approach to persuasiveness prediction with a novel graph-based neural network model, called **h**eterogeneous **arg**ument **a**ttention **n**etwork (HARGAN). By jointly training on the persuasiveness and stance of the replies, our model achieves the state-of-the-art performance on the ChangeMyView (CMV) dataset for the persuasiveness prediction task. Experimental results show that the graph setting enables our model to aggregate information across multiple paragraphs effectively. In the meanwhile, our stance prediction auxiliary task enables our model to identify the viewpoint of each party, and helps our model perform better on the persuasiveness prediction.

Introduction

Argument mining is an attractive topic where nearly all aspects of natural language processing tasks are explored, including sequence labeling (argumentative discourse unit detection), structure parsing (argument structure parsing), knowledge retrieval (claim/premise retrieval), paragraph generation (argument generation), and text classification (persuasiveness prediction). Although the above tasks seem very different, they all share an ultimate goal—persuasiveness. For example, the main purpose behind the claim/premise retrieval is to find a persuasive resource. The goal of argument generation is to generate more persuasive text. A model for persuasiveness prediction cannot only measure the persuasiveness, but also serve as an auxiliary task for other tasks in the argument mining. However, the state-of-the-art persuasiveness prediction model still leaves much room to be improved.

Several research has been conducted on persuasiveness prediction. Tan et al. (2016) use a set of handcrafted features to predict persuasiveness. Toledo et al. (2019) and Gleize et al. (2019) predict persuasiveness in a single sentence. However, the argumentation process in real world is usually involved with more than one sentence. The relation among multiple sentences should be a more important factor that affects persuasiveness. Jo et al. (2018) and Ji et al. (2018) consider the sentence-level attention between the paragraphs in the original post and in the root reply. However, their model limits the argumentation process into a single turn so that it cannot deal with multi-turn debates. Hidey and McKeown (2018) and Zeng et al. (2020) use the attention mechanism to encode paragraph-level information. We point out that none of these models consider the sentence-level relation across several paragraphs in the conversation data and these models might only encode local information. To consider the sentence-level attention across whole the conversation, argument structure, which is a tree structure defining the relation among claims and premise, is used in our model. In summary, our model leverages the information of argument structure to predict the persuasiveness in the conversation.

This paper presents a new graph neural network (GNN) based model called **H**eterogeneous **Arg**ument **A**ttention **N**etwork (HARGAN) that jointly learns persuasiveness prediction and stance prediction in multi-turn conversation. The advantage of our model is that it utilizes a GNN module to exploit argument structure information. The argument structure is a tree structure defining the relation among argumentative discourse units (ADUs). In argument mining, an ADU, an elementary unit expressing a point, might be a claim or a premise. Participants in the debate often use claims/premises to support their viewpoints or attack other's viewpoints, thus the argument structure reveals the structure of a debate. Since an argument structure constitutes a tree, which is a special kind of graph, it can be modeled with a GNN that learns the relation between nodes in a graph. In addition, by using stance prediction as an auxiliary task, our model can better recognize the stance of each anonymous online user and make better decisions. Experimental results show that our methodology achieves the state-of-the-art performance in persuasiveness prediction on the ChangeMyView dataset.

Our contributions are threefold listed as follows:

- We propose a novel hierarchical graph setting to represent argument structure and show the advantage of incorporating the argument structure into persuasiveness prediction.
- We introduce the stance prediction as an auxiliary task to help our model recognize the viewpoint of each party.
- We showcase our idea with a system that parses the argument structure from raw data and then leverages its information to predict the persuasiveness.

Related Work

Persuasiveness in NLP

Persuasiveness has been discussed in several aspects. Some research focuses on finding the key factor that persuades others. Durmus and Cardie (2018) study the effects of prior belief on persuasion. Wang et al. (2019b) provide a dataset to study sentence-level persuasion strategy in a dialogue. Atkinson, Srinivasan, and Tan (2019) use the explanation provided by those being persuaded to identify key points. Some research focuses on training a neural network model to predict persuasiveness or argument quality in a sentence. Beigman Klebanov et al. (2016) investigate the relationship between argumentation structures, argument content, and the quality of the essay. Habernal and Gurevych (2016b,a), Toledo et al. (2019), and Gleize et al. (2019) provide datasets with baseline results. Some research focuses on predicting whether the original poster will be persuaded by others in the conversation. Tan et al. (2016) explore a set of handcrafted features. Jo et al. (2018) and Ji et al. (2018) consider co-attention between the paragraphs in the original post and the root reply. Hidey and McKeown (2018) use the hierarchical attention mechanism to encode each paragraph. Zeng et al. (2020) encode the paragraph representation by using discourse and topic information with the memory mechanism. Li, Durmus, and Cardie (2020) use LSTM to encode several predefined argument structure features.

Argument Structure Parsing

Neural networks have recently been used for argument structure parsing. Eger, Daxenberger, and Gurevych (2017) build an end-to-end model to label the ADU and predict each link at a token-level. Most other works still partition this task into argumentative discourse unit (ADU) recognition and dependency structure parsing. For ADU recognition, Ajjour et al. (2017) compare the performance of different models on the benchmark dataset in this task. Trautmann et al. (2020) present a dataset of arguments from various sources. Chakrabarty, Hidey, and McKeown (2019) fine tuning a language model using a Reddit corpus of 5.5 million opinionated claims, which are collected by finding the internet acronyms IMO/IMHO. For dependency structure parsing, Potash, Romanov, and Rumshisky (2017) use pointer network to solve this problem. The ADUs are first input into the encoder, and then the decoder is asked to predict the parent ADU for i-th ADU at timestep i. Kuribayashi et al. (2019) leverage information from argument markers, such as "however" and "therefore", in front of each ADU to help predict the higher-order relations. Chakrabarty et al. (2019) consider link prediction as a sequence pair classification problem. Also, they further increase the performance by pruning out some ADUs using a summary model.

Graph Neural Network

In NLP, more and more research incorporates graph neural networks into their models. GNN is used to capture dependency information between words (Marcheggiani and Titov 2017; Zhang, Qi, and Manning 2018; Zhang, Li, and Song 2019). In multi-hop question answering, some work (**?**Tu et al. 2019; Qiu et al. 2019) considers named entities that appear in different sentences to build links between sentences or entities and use a GNN to aggregate multiple sentences information. In text summarization, some work (Yasunaga et al. 2017; Xu et al. 2020) considers the similarity or discourse relation between sentences to build the link between sentences and use a GNN to aggregate the information from multiple sentences.

Dataset

Our study is conducted on the ChangeMyView (CMV) dataset (Tan et al. 2016). CMV is a subreddit where users join multi-turn discussions and try to change the opinion holder's mind in several aspects. Hence a post could have multiple root replies. By considering each reply as a node, the whole post will become a tree-like structure. The original post (OP) represents the root and the replies directly responding to the OP are called root reply. The system asks the opinion holder to denote the reply that changes his/her mind and provide explanation as to why.

We perform depth-first search on a discussion thread for constructing the root-to-leaf path. A path consists of multiple replies, and a reply consists of one or more paragraphs. Winning paths are truncated at the winning reply and the others are truncated at the last non-OP reply. Our goal is to predict whether a path is persuasive enough to convince the opinion holder to change their view. Our setting is based on Tan et al. (2016) and Zeng et al. (2020) with several modifications. Firstly, in order to prevent some special topics, we remove posts with less than 10 authors. Secondly, since the difficulty of being persuaded by others varies accordingly, we examine persuasiveness under the pairwise comparison. Pairs are chosen under the following three constraints.

1. We only compare the paths under the same post.
2. The candidates in the same pair should come from different root replies.
3. The Jaccard similarity between positive and negative should be greater than 0.5.

We also filter the data by a length limit. Unlike Zeng et al. (2020), we do not remove original poster's replies from the path because the interaction between OP and replier is a key feature for prediction. Table 1 shows the statistic after preprocessing.

Methodology

This section will shows how to build the argument structure in the given path, and how our model leverages the argument structure to predict persuasiveness. Figure 1 depicts our overall system, which consists of two stages (1) ar-

	Train	Dev	Test
# trees	969	241	311
# pairs	14922	3504	5013
Avg. turns	2.87	2.96	2.83

Table 1: Statistic of changemyview.

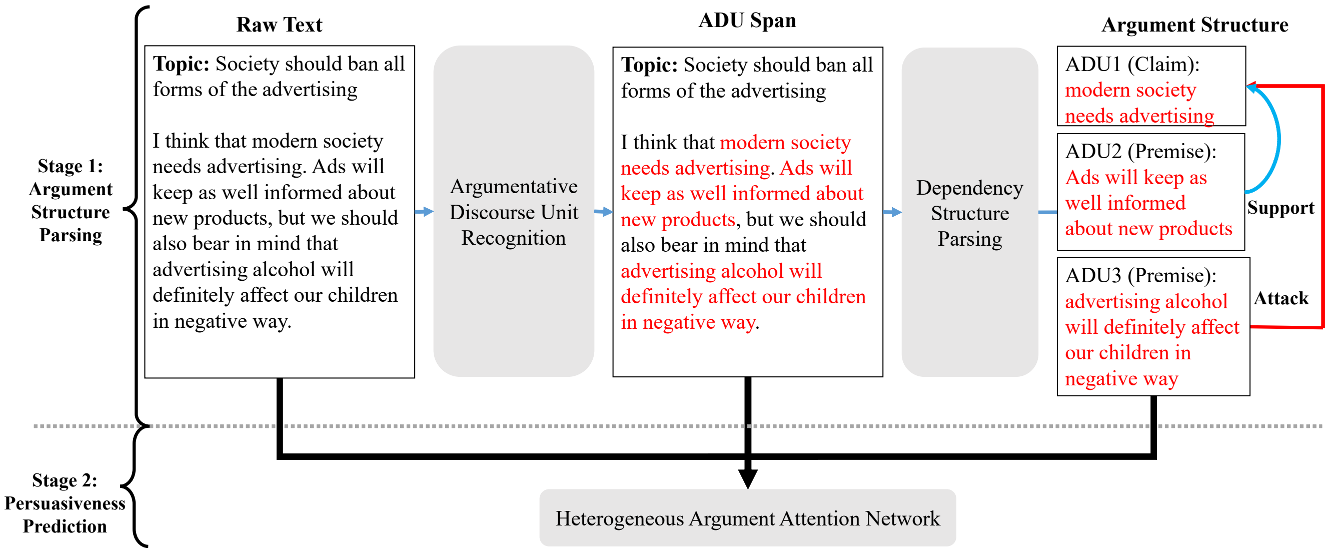

Stage 1: Argument Structure Parsing
Raw Text
Topic: Society should ban all forms of the advertising
I think that modern society needs advertising. Ads will keep as well informed about new products, but we should also bear in mind that advertising alcohol will definitely affect our children in negative way.
Argumentative Discourse Unit Recognition
ADU Span
Topic: Society should ban all forms of the advertising
I think that modern society needs advertising. Ads will keep as well informed about new products, but we should also bear in mind that advertising alcohol will definitely affect our children in negative way.
Dependency Structure Parsing
Argument Structure
ADU1 (Claim): modern society needs advertising
ADU2 (Premise): Ads will keep as well informed about new products
ADU3 (Premise): advertising alcohol will definitely affect our children in negative way
Support
Attack
Stage 2: Persuasiveness Prediction
Heterogeneous Argument Attention Network

Figure 1: Overview of our system. The upper part of the figure illustrates the input and the output of each step in stage 1. After the first stage, the raw text, the span of ADU, and the argument structure are given to our Heterogeneous Argument Attention Network for persuasiveness prediction.

gument structure parsing and (2) persuasiveness prediction. The argument structure parsing stage is composed of ADU recognition and dependency structure parsing. For persuasiveness prediction, we propose a graph-based model to incorporate argument structure information.

Argument Structure Parsing

The upper part of Figure 1 illustrates the two steps in the stage 1. First, raw text is input into the ADU recognition model for tagging the spans of ADUs. Then, the ADUs are input into the dependency structure parsing model to construct the tree structure.

Argumentative Discourse Unit Recognition The first step to the argument structure parsing is to find all ADUs in a paragraph. The boundary of an ADU is not limited to the punctuation. That is, an ADU may be a clause, a complete sentence, multiple sentences, or something in between. Therefore, a word-level sequence labeling model is needed. We build a model using BERT (Devlin et al. 2019) plus a conditional random field (CRF) layer. Due to the unclear ending signal of an ADU, we further find that our model can be improved by using the BIEO, where B, I, E, and O denote Beginning, Inside, Ending, and Outside, tagging scheme (Borthwick 1999).

Dependency Structure Parsing In the stage of dependency structure parsing, we modify the model proposed by Kuribayashi et al. (2019) to tackle multiple paragraphs. The information of an ADU emphasized by adding argumentative markers (AMs), such as “however” and “therefore”, in front of each ADU. Hereafter, we will refer to the original ADU as argumentative component (AC), and the new ADU is the combination of AM and AC.

The input of the dependency structure parser is paragraphs marked with start and end points of each AC and AM. Each AC/AM is represented by utilizing ELMO (Peters et al. 2018). For each word in the paragraph, the word embedding is input into a bidirectional long short-term memory (Bi-LSTM) (Hochreiter and Schmidhuber 1997) to encode the contextual embedding. Then, representations of AC and AM are encoded by LSTM-Minus (Wang and Chang 2016). Let Z_n^{AC} and Z_n^{AM} denote the representations of n-th AC and n-th AM, respectively. Assuming Z_n^{AC} spans from i-th to j-th words. It is defined by the following equations:

$$Z_n^{AC} = \text{diff}(W, i, j) \tag{1}$$

$$\text{diff}(W, i, j) = [\overrightarrow{W_j} - \overrightarrow{W_{i-1}} \| \overleftarrow{W_i} - \overleftarrow{W_{j+1}} \| \overrightarrow{W_{i-1}} \| \overleftarrow{W_{j+1}}] \tag{2}$$

where W is the output from the Bi-LSTM. Then, ADU Z_n is represented by the concatenation of its AC and AM representations and the position embedding ϕ, that is,

$$Z_n = [Z_n^{AC} \parallel Z_n^{AM} \| \phi(n)] \tag{3}$$

Finally, the child-parent score $score_{link}(n, m)$ for the m-th ADU to be the parent of the n-th ADU is defined as follows:

$$\text{score}_{link}(n, m) = \text{softmax}_m(\alpha_{nm}) \tag{4}$$

$$\alpha_{nm} = W_{link}[Z_n \parallel Z_m \parallel Z_n \odot Z_m \| \theta(n - m)] \tag{5}$$

and the link type for this link is computed as follows:

$$P_{type}^{nm}(Attack | Z_n, Z_m) = \text{sigmoid}(\beta_{nm}) \tag{6}$$

$$\beta_{nm} = W_{type}[Z_n \parallel Z_m \parallel Z_n \odot Z_m \| \theta(n - m)] \tag{7}$$

where W_{link} and W_{type} are two trainable matrices, $\odot$ denotes the pointwise product between two vectors, and θ is

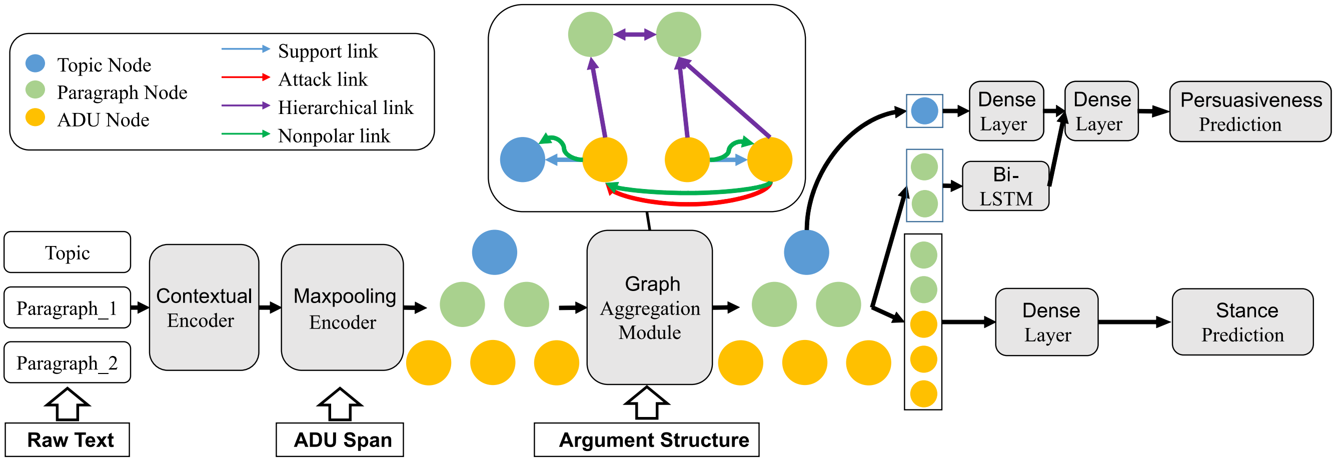


Figure 2: Overview of our model. The middle of the upper part gives an example of the graph (the self-link for each node is omitted for clarity).

the relative position embedding.

HARGAN

Figure 2 illustrates the architecture of our model. First, the contextual encoder is used to compute the contextual embedding for each word. Second, the representations of ADUs, the topic, and paragraphs are encoded by using the span information of ADUs, the topic length, and paragraph lengths respectively. Third, these representations are updated according to the hierarchical graph by the graph aggregation module. Finally, the updated representations are used for two tasks: stance prediction and persuasiveness prediction.

Contextual Encoder The ELMO-based encoder is also employed as the contextual encoder in argument structure parsing. Since some paragraphs contain no ADU, the encoder here also encodes the representations of all paragraphs. In addition, we find that the max-pooling performs better than LSTM-Minus in encoding the representation of ADUs and paragraphs in persuasive prediction, so the max-pooling is selected.

Heterogeneous Graph Aggregation Module

Heterogeneous Graph. In the original setting of argument structure parsing, ADUs are combined into a tree. In this way, each ADU links to a parent ADU, which has the highest child-parent score from the dependency structure parsing stage. However, the prediction might contain errors. To prevent error propagation from the previous stage, we loosen the constraint for the structure to be a graph. That is to say, each ADU could link to more than one parent ADU. We compute the child-parent score between the current ADU and each candidate using Equation (4), and choose the top-k candidates to link with. In our experiment, k is set to three. To construct the graph for the graph aggregation module, four kinds of relations are defined according to argument structure.

1. **Support**: The support relation contains all the links whose link type is predicted as support in dependency structure parsing.
2. **Attack**: The attack relation contains all the links whose link type is predicted as attack in dependency structure parsing.
3. **Non-polar**: To prevent the error propagation from the structure parsing stage, the non-polar relation contains both attack and support links.
4. **Hierarchical**: The hierarchical relation contains links from an ADU to the paragraph it belongs to, from a paragraph to other paragraphs in the same reply, and from a paragraph to paragraphs in the consecutive reply.

Among these relations, **support**, **attack**, and **non-polar** are for the ADU level, and **hierarchical** is for the paragraph level. Also, following the GAT's setting, each node contains a self-link for all types of relationships. The middle of the upper part in Figure 2 illustrates an example of our graph.

Graph Aggregation Model. Due to the heterogeneity of our link, our model utilizes the RNN version of the heterogeneous graph attention network (HAN) (Wang et al. 2019a) to aggregate node's information.

First, the node-level attention is applied to fuse information inside each relation. Recall there are four kinds of relations in our graph demonstrated in Section . That is, $R \equiv \{Support, Attack, Non-polar, Hierarchical\}$. Let X_i denote the i-th node's representation. For $r \in R$, the equation of the attention weight α_{ij}^r for X_i on X_j is defined as follows:

$$\begin{aligned} \alpha_{ij}^r &= \mathrm{softmax}_j(e_{ij}^r) \\ &= \frac{\exp(\sigma(W_r^T \cdot [X_i' \parallel X_j']))}{\sum_{k \in N_i^r} \exp(\sigma(W_r^T \cdot [X_i' \parallel X_k']))} \end{aligned} \tag{8}$$

where X_i' is a linear transformation of X_i, Leaky ReLU (Maas, Hannun, and Ng 2013) is used for the activation function σ and W_r^T is a trainable matrix.

Following the setting in HAN, the multi-head attention is used in our model. The representation after this attention is shown below:

$$X_i^r = \overset{M}{\underset{m=1}{\Vert}} \delta\Big(\sum_{j \in N_i^r} \alpha_{ij}^r \cdot X_j'\Big) \tag{9}$$

where the activation function is elu (Clevert, Unterthiner, and Hochreiter 2016) and M, which is the number of attention head, is chosen to be 4 in our experiment.

Second, the semantic level attention is applied to fuse information between different relations. The output X_i is determined by the following equation:

$$X_i = \sum_{r \in R} \beta_i^r \cdot X_i^r \tag{10}$$

$$\begin{aligned} \beta_i^r &= \mathrm{softmax}_r(e_i^r) \\ &= \frac{\exp(q^T \cdot \tanh(W \cdot X_i^r + b))}{\sum_{r \in R} \exp(q^T \cdot \tanh(W \cdot X_i^r + b))} \end{aligned} \tag{11}$$

where W is a trainable matrix. Finally, an LSTM is utilized to collect the output of every layer.

Stance Prediction. We consider only positive or negative stances; stance prediction is treated as a binary classification task. Since the stance of each individual reply on the forum CMV is unlabeled, we train a model for stance prediction with distant-supervision and semi-supervised. Pseudo labels are made by regarding the OP and the label giver as the positive side. On the other hand, the root replier are automatically labeled as the negative side. The stance prediction loss is only measured on those ADUs and paragraphs of labeled users by using binary cross entropy.

$$\begin{aligned} Loss_{stance} = \frac{1}{|C_L|} \sum_{l \in C_L} (L_l \log(\sigma(Y_l^{stance}) + \\ (1 - L_l) \log(\sigma(Y_l^{stance})) \end{aligned} \tag{12}$$

where L_l and Y_l^{stance} are the label and the prediction of node l, and C_L is the set of nodes whose stance could be identified.

Persuasiveness Prediction. We follow the setting in Zeng et al. (2020) and consider relative persuasiveness given a pair of paths. Our goal is to make the winning path scored higher than the failed path. A Bi-LSTM is utilized to collect the representation of paragraphs from the heterogeneous graph aggregation module. Next, the outputs of the Bi-LSTM at each timestep are summarized to get the representation of the whole path by max-pooling, mean-pooling, and attention. Finally, the concatenation of the representations of the topic node and path is passed into a fully connected layer to predict its persuasiveness. The pairwise cross-entropy loss is defined as follows to measure the margin of Y_{pers}^+ and Y_{pers}^- for the winning path and the failed path:

$$Loss_{pers} = \log(1 + \exp(Y_{pers}^- - Y_{pers}^+)) \tag{13}$$

Model	Acc (%)	F1 (%)
BERT+CRF with BIO	92.3	85.0
BERT+CRF with BIEO	93.0	88.4

Table 2: Results of ADU recognition.

Task	Metric	Performance
Link	Acc	39.08%
Link	MRR	57.39%
Link Type	F-score	52.86%

Table 3: Results of dependency structure parsing.

Objective Function The overall objective function is the summation of stance prediction loss $Loss_{stance}^{para}$ and $Loss_{stance}^{ADU}$ on paragraph and ADU, and persuasiveness prediction loss $Loss_{pers}$:

$$Loss = Loss_{pers} + Loss_{stance}^{para} + \alpha Loss_{stance}^{ADU} \tag{14}$$

where α is set to be 0.01 in our experiment.

Experiments

Data Preprocessing

We randomly split the training set into two parts, 80% of instances for training and 20% of instances for validation. For preprocessing, we take the following steps. First, quotations, user name, edit tag, and links were respectively replaced with generic tags "<cite>", "<user>", "<edit>", and "<url>". Next, we utilized the natural language toolkit (NLTK) (Loper and Bird 2002) for tokenization. Afterwards, all letters were converted to lowercase.

Results of ADU Recognition

Our ADU recognition model is trained based on datasets by Egawa, Morio, and Fujita (2019) and Stab and Gurevych (2014). Performance is evaluated by accuracy and F score, where accuracy is calculated by exact matching of words, and F-score is macro F1. Table 2 presents the results under the BIO and the BIEO two tagging schemes. The BIEO tagging scheme improves the accuracy by 0.7% and F-score by 3.4%, showing our model could catch the ending signal more effectively by using the ending tag.

Results of Dependency Structure Parsing

The dependency structure parsing model is trained based on the dataset by Egawa, Morio, and Fujita (2020). Three metrics, accuracy on link prediction, mean reciprocal rank (MRR) on link prediction, and F-score of the link type, are employed for evaluation.

Table 3 shows that for link prediction, we achieve an accuracy of 39.08% when choosing the top candidate to be our target. However, the MRR indicates that the correct parent exists around top-2 or top-3 in most cases. This is the reason why we choose k to be 3 in our graph setting. For link type prediction, we achieve an F-score of 52.86%, indicating that

	Persuasiveness		ADU-wide stance		Paragraph-wide stance	
Model	**Acc (%)**	**F1 (%)**	**Acc (%)**	**F1 (%)**	**Acc (%)**	**F1 (%)**
TFIDF-LR	62.64	62.64	-	-	-	-
HN-ATT	72.15	72.15	-	-	-	-
DTDMN	76.38	76.38	-	-	-	-
HARGAN-RGCN	82.23*	82.23*	96.23	95.20	93.41	90.97
HARGAN-R-RGCN	83.90*	83.90*	96.83	96.01	91.84	88.38
HARGAN-HAN	83.96*	83.96*	**98.24**	**97.79**	96.22	94.84
HARGAN-RHAN	**85.14***	**85.14***	97.97	97.42	**96.81**	**95.63**

Table 4: Results of persuasiveness prediction and stance prediction. The * symbol denotes the model significantly outperforms DTDMN.

the model might be misled if only the support and attack relations are considered. This is why we need the non-polar relation in our graph.

Results of Persuasiveness & Stance Prediction

The proposed HARGAN is compared with three previous approaches on persuasiveness prediction. Meanwhile, three variants of heterogeneous graph neural network are used to verify the effectiveness of our graph setting. Since our model uses R-HAN as GNN, we term it HARGAN-RHAN.

- TFIDF-LR (Tan et al. 2016) employs logistic regression with bag-of-words feature to predict pervasiveness.
- HN-ATT (Yang et al. 2016) employs word and sentence-level attention to encode paragraph information and use paragraph level attention to encode the path information.
- DTDMN (Zeng et al. 2020), the state-of-the-art persuasiveness prediction, employs topic and discourse information as the weight of the memory network to encode the paragraph information.
- RGCN (Schlichtkrull et al. 2018), a variant of heterogeneous graph neural networks, separately employs the graph convolutional network on different relation links and aggregates them by average. We term this variant HARGAN-RGCN.
- R-RGCN (Huang and Carley 2019), a variant of heterogeneous graph neural network, solves the problem that GNN could not be very deep by adding an RNN layer at each layer's output. We term this variant HARGAN-R-RGCN.
- HAN (Wang et al. 2019a), a variant of heterogeneous graph neural networks, separately employs the graph attention network on different relation links and aggregates them by attention. We term this variant HARGAN-HAN.

Table 4 shows the results on persuasiveness prediction and stance prediction. For persuasiveness prediction, it shows that all of our models significantly outperform the state-of-the-art model, DTDMN, by using McNemar's test (McNemar 1947) (p<0.01). On the other hand, among four different graph settings the graph-attention-based model performs better than graph-convolutional-based model on both persuasiveness prediction and stance prediction. We guess that it is because out task is an inductive learning problem and it is impossible to see all kinds of argument structures during training, the graph-attention-based models, which are good at inductive learning, perform better than graph-convolutional-based models.

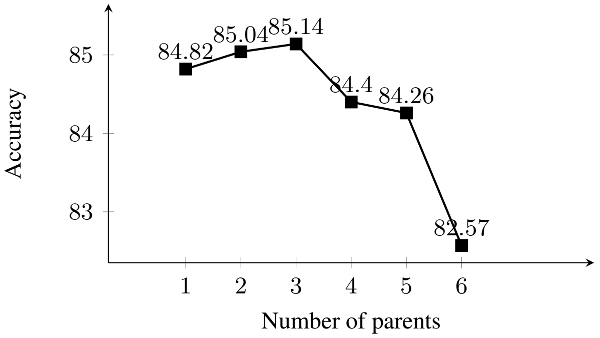


Figure 3: Accuracy of various number of parents of an ADU.

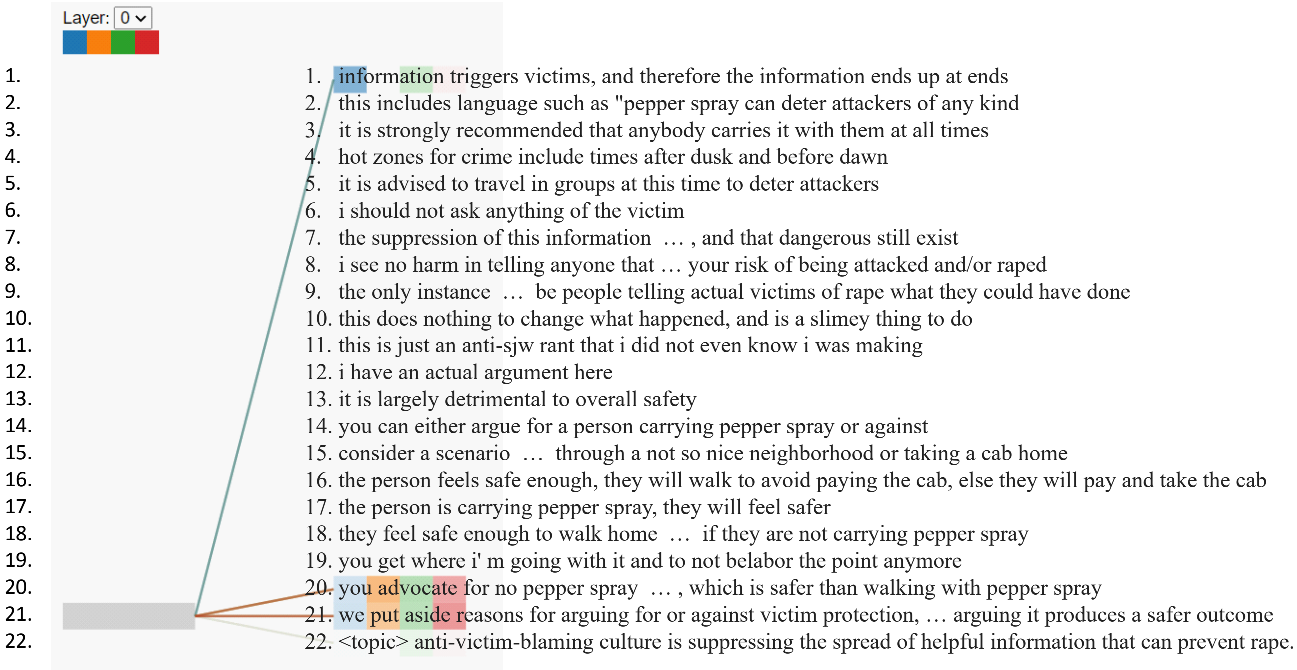
Discussion

Ablation Analysis on HARGAN

To study the effectiveness of our approach, we ablate the model with the following factors:

- *No heterogeneous graph*, which removes the **attack**, **support** and **non-polar** relation in the graph and leaves only **hierarchical** relation;
- *Random structure*, which uses randomly generated argument structure on training and testing in persuasiveness prediction. It is noted that **hierarchical** relation is same as the HARGAN-RHAN in this setting;
- *No stance prediction loss*, which leaves our loss with only the persuasiveness prediction loss;
- *No heterogeneous graph aggregation module*, which encodes the paragraph representation by doing max-pooling over words in paragraph;

Table 5 indicates three insights: (1) **The heterogeneous graph makes the result better.** Removing the attack, support, and non-polar relations drops the accuracy by 0.8%. Therefore, we conclude that by using the argument structure, our model can understand the support and attack relation between ADUs and obtain better embedding. (2) **Wrong argument structure will hurt performance.** Replace the predicted argument structure with random one



1. information triggers victims, and therefore the information ends up at ends
2. this includes language such as "pepper spray can deter attackers of any kind
3. it is strongly recommended that anybody carries it with them at all times
4. hot zones for crime include times after dusk and before dawn
5. it is advised to travel in groups at this time to deter attackers
6. i should not ask anything of the victim
7. the suppression of this information … , and that dangerous still exist
8. i see no harm in telling anyone that … your risk of being attacked and/or raped
9. the only instance … be people telling actual victims of rape what they could have done
10. this does nothing to change what happened, and is a slimey thing to do
11. this is just an anti-sjw rant that i did not even know i was making
12. i have an actual argument here
13. it is largely detrimental to overall safety
14. you can either argue for a person carrying pepper spray or against
15. consider a scenario … through a not so nice neighborhood or taking a cab home
16. the person feels safe enough, they will walk to avoid paying the cab, else they will pay and take the cab
17. the person is carrying pepper spray, they will feel safer
18. they feel safe enough to walk home … if they are not carrying pepper spray
19. you get where i' m going with it and to not belabor the point anymore
20. you advocate for no pepper spray … , which is safer than walking with pepper spray
21. we put aside reasons for arguing for or against victim protection, … arguing it produces a safer outcome
22. <topic> anti-victim-blaming culture is suppressing the spread of helpful information that can prevent rape.

Figure 4: Visualization of the attention graph for an ADU.

	Persuasiveness	
Model	**Acc (%)**	**Macro F1 (%)**
HARGAN-RHAN	85.14	85.14
w/o Argument structure	84.34*	84.34*
w Random structure	83.66**	83.66**
w/o Stance prediction	83.60**	83.60**
w/o GNN	69.56**	69.55**

Table 5: Ablation analysis on HARGAN. The * and ** symbol denote the model significantly underperforms HARGAN-RHAN at $p < 0.05$ and $p < 0.001$.

drops the accuracy by 1.48%, which is even worse than that removed support, attack and relation link. Therefore, we conclude that wrong argument will make the performance even worse. (3) **Stance information helps our model understand which party a paragraph belongs to.** Removing the stance prediction loss significantly drops the accuracy by 1.54%. Therefore, we conclude that it can help our model distinguish the viewpoint of each party. (4) **The graph aggregation module is the most important in our model.**

Removing the graph aggregation module significantly drops the accuracy by 15.58%, which is even worse than our baseline model HN-ATT. Thus, we conclude that our graph aggregation module can aggregate information from different ADUs and give a better result than using max-pooling.

Ablation Analysis on Graph Relation

To study the effectiveness of our graph setting, we compare the performance of different graph settings. We ablate the graph with:

- *No support*, which removes the support relation;
- *No attack*, which removes the attack relation;
- *No attack and support*, which removes the attack and support relation;
- *No non-polar*, which removes the non-polar relation;
- *No hierarchical*, which removes the hierarchical relation.

It is important to note that the paragraph representation would not be updated in the graph aggregate model if we remove the hierarchical relation. As a result, the comparison might be unfair. To close the gap between each setting, the input for the Bi-LSTM for persuasiveness prediction is modified to be the ADU's representations from the output of the graph aggregation module for no hierarchical setting.

Table 6 shows the result by removing different relations. First, removing support and attack relation significantly drops the accuracy significantly by 1.76%, which is larger than the accuracy drops of removing the support and the attack relations. Therefore, we conclude that adding support and attack relationships allows the model to learn more features. Second, removing non-polar relation significantly drops the accuracy significantly by 0.76%, which is larger than the accuracy drops of removing the support and the attack relations. Therefore, we conclude that adding the non-polar relation without the link type gives our model a chance to fix up the error in the argument structure parsing stage. Finally, removing the hierarchical relation significantly drops the accuracy by 0.82%. Therefore, we conclude that adding the hierarchical relation helps us keep the information of those paragraphs without ADUs.

	Persuasiveness	
Model	**Acc (%)**	**Macro F1 (%)**
HARGAN-RHAN	85.14	85.14
w/o Support	84.60	84.60
w/o Attack	84.78	84.78
w/o Attack & Support	83.38**	83.38**
w/o Non-polar	84.38*	84.38*
w/o Hierarchical	84.32*	84.32*

Table 6: Ablation analysis on different graph relation. The * and ** symbol denote the model significantly underperforms HARGAN-RHAN at $p < 0.05$ and $p < 0.001$.

Performance of Pruned Graph

To study the effectiveness of the threshold on our links in the argument structure, we compare the performance under different numbers of parents that a node points to. Figure 3 illustrates that although the attention mechanism could learn the relation between two nodes, the performance still drops if we consider too many links. We assume that by increasing the numbers of the parents for each node, some noise is induced in our model and causes the bad performance.

Case Study

Here, we visualize the *Non-polar* relation it learns in a CMV conversation [1] as a case study. Figure 4 demonstrates the result. Different colors denote the multi-head attention, and the attention score is higher if the color is darker. In this conversation, the replier first agrees with the author's point of view, and then gives a negative example of this point. We can see the links from sentence 21 to sentences 1 and 20 because sentence 21 illustrates how the replier attacks the original poster, while sentences 1 is the claim from original poster and sentence 20 is the summarization of the author's idea. The attention mechanism builds the relation among ADU and helps us understand each claim and premise. Finally the replier successfully convinced the original poster.

Conclusion

In this paper, we introduce a hierarchical graph setting to argument mining and propose a novel graph-based model leveraging argument structure to jointly predict the stance and the persuasiveness. The experiment demonstrates that our graph setting and graph aggregation module are important for persuasiveness prediction. In the future, our model can be used to recommend retrieved passages for the topic. On the other hand, our model can be extended to other tasks in argument mining. For example, our model can be used to encode argument history and generate attractive arguments for debate or salesman use and other novel applications. Also, the score of our model can serve as an auxiliary task for other tasks in the argument mining to optimize. Our code is publicly available for the research community.[2]

Acknowledgments

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-109-2634-F-002-040-, MOST-109-2634-F-002-034-, MOST-109-2218-E-009-014-, and by Academia Sinica, Taiwan, under grant AS-TP-107-M05.

References

Ajjour, Y.; Chen, W.-F.; Kiesel, J.; Wachsmuth, H.; and Stein, B. 2017. Unit Segmentation of Argumentative Texts. In *Proceedings of the 4th Workshop on Argument Mining*, 118–128.

Atkinson, D.; Srinivasan, K. B.; and Tan, C. 2019. What Gets Echoed? Understanding the "Pointers" in Explanations of Persuasive Arguments. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2911–2921.

Beigman Klebanov, B.; Stab, C.; Burstein, J.; Song, Y.; Gyawali, B.; and Gurevych, I. 2016. Argumentation: Content, Structure, and Relationship with Essay Quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, 70–75.

Borthwick, A. E. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, USA. AAI9945252.

Chakrabarty, T.; Hidey, C.; and McKeown, K. 2019. IMHO Fine-Tuning Improves Claim Detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 558–563.

Chakrabarty, T.; Hidey, C.; Muresan, S.; McKeown, K.; and Hwang, A. 2019. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2933–2943.

Clevert, D.; Unterthiner, T.; and Hochreiter, S. 2016. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). In Bengio, Y.; and LeCun, Y., eds., *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Durmus, E.; and Cardie, C. 2018. Exploring the Role of Prior Beliefs for Argument Persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1035–1045.

Egawa, R.; Morio, G.; and Fujita, K. 2019. Annotating and Analyzing Semantic Role of Elementary Units and Relations in Online Persuasive Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 422–428.

[1] https://www.reddit.com/r/changemyview/comments/3fvhjh/cmv_antivictimblaming_culture_is_suppressing_the/ctsfvej/?context=8&depth=9

[2] https://github.com/seasa2016/Heterogeneous_Argument_Attention_Network

Egawa, R.; Morio, G.; and Fujita, K. 2020. Corpus for Modeling User Interactions in Online Persuasive Discussions. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 1135–1141. ISBN 979-10-95546-34-4.

Eger, S.; Daxenberger, J.; and Gurevych, I. 2017. Neural End-to-End Learning for Computational Argumentation Mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11–22.

Gleize, M.; Shnarch, E.; Choshen, L.; Dankin, L.; Moshkowich, G.; Aharonov, R.; and Slonim, N. 2019. Are You Convinced? Choosing the More Convincing Evidence with a Siamese Network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 967–976.

Habernal, I.; and Gurevych, I. 2016a. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1214–1223.

Habernal, I.; and Gurevych, I. 2016b. 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)*, 1589–1599.

Hidey, C.; and McKeown, K. R. 2018. Persuasive Influence Detection: The Role of Argument Sequencing. In McIlraith, S. A.; and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 5173–5180.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.* 9(8): 1735–1780. ISSN 0899-7667.

Huang, B.; and Carley, K. M. 2019. Inductive Graph Representation Learning with Recurrent Graph Neural Networks. *CoRR* abs/1904.08035.

Ji, L.; Wei, Z.; Hu, X.; Liu, Y.; Zhang, Q.; and Huang, X. 2018. Incorporating Argument-Level Interactions for Persuasion Comments Evaluation using Co-attention Model. In *Proceedings of the 27th International Conference on Computational Linguistics*, 3703–3714.

Jo, Y.; Poddar, S.; Jeon, B.; Shen, Q.; Rosé, C.; and Neubig, G. 2018. Attentive Interaction Model: Modeling Changes in View in Argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 103–116.

Kuribayashi, T.; Ouchi, H.; Inoue, N.; Reisert, P.; Miyoshi, T.; Suzuki, J.; and Inui, K. 2019. An Empirical Study of Span Representations in Argumentation Structure Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4691–4698.

Li, J.; Durmus, E.; and Cardie, C. 2020. Exploring the Role of Argument Structure in Online Debate Persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8905–8912.

Loper, E.; and Bird, S. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, 63–70.

Maas, A.; Hannun, A.; and Ng, A. 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *Proceedings of the International Conference on Machine Learning*. Atlanta, Georgia.

Marcheggiani, D.; and Titov, I. 2017. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1506–1515.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2): 153–157.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237.

Potash, P.; Romanov, A.; and Rumshisky, A. 2017. Here's My Point: Joint Pointer Architecture for Argument Mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1364–1373.

Qiu, L.; Xiao, Y.; Qu, Y.; Zhou, H.; Li, L.; Zhang, W.; and Yu, Y. 2019. Dynamically Fused Graph Network for Multi-hop Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6140–6150.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, 593–607. Springer.

Stab, C.; and Gurevych, I. 2014. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1501–1510.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 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 '16, 613–624. ISBN 9781450341431.

Toledo, A.; Gretz, S.; Cohen-Karlik, E.; Friedman, R.; Venezian, E.; Lahav, D.; Jacovi, M.; Aharonov, R.; and Slonim, N. 2019. Automatic Argument Quality Assessment - New Datasets and Methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5625–5635.

Trautmann, D.; Daxenberger, J.; Stab, C.; Schütze, H.; and Gurevych, I. 2020. Fine-Grained Argument Unit Recognition and Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(05): 9048–9056.

Tu, M.; Wang, G.; Huang, J.; Tang, Y.; He, X.; and Zhou, B. 2019. Multi-hop Reading Comprehension across Multiple Documents by Reasoning over Heterogeneous Graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2704–2713.

Wang, W.; and Chang, B. 2016. Graph-based Dependency Parsing with Bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2306–2315.

Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; and Yu, P. S. 2019a. Heterogeneous Graph Attention Network. In *The*

World Wide Web Conference, WWW '19, 2022–2032. ISBN 9781450366748.

Wang, X.; Shi, W.; Kim, R.; Oh, Y.; Yang, S.; Zhang, J.; and Yu, Z. 2019b. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5635–5649.

Xu, J.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Discourse-Aware Neural Extractive Text Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5021–5031.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489.

Yasunaga, M.; Zhang, R.; Meelu, K.; Pareek, A.; Srinivasan, K.; and Radev, D. 2017. Graph-based Neural Multi-Document Summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 452–462.

Zeng, J.; Li, J.; He, Y.; Gao, C.; Lyu, M.; and King, I. 2020. What Changed Your Mind: The Roles of Dynamic Topics and Discourse in Argumentation Process. In *Proceedings of The Web Conference 2020*, WWW '20, 1502–1513. ISBN 9781450370233.

Zhang, C.; Li, Q.; and Song, D. 2019. Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4568–4578.

Zhang, Y.; Qi, P.; and Manning, C. D. 2018. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2205–2215.