Exploring Discourse Structures for Argument Impact Classification

Xin Liu[1]* **Jiefu Ou**[1] **Yangqiu Song**[1] **Xin Jiang**[2]
[1]Department of CSE, the Hong Kong University of Science and Technology
[2]Huawei Noah's Ark Lab
xliucr@cse.ust.hk jouaa@connect.ust.hk
yqsong@cse.ust.hk jiang.xin@huawei.com

Abstract

Discourse relations among arguments reveal logical structures of a debate conversation. However, no prior work has explicitly studied how the sequence of discourse relations influence a claim's impact. This paper empirically shows that the discourse relations between two arguments along the context path are essential factors for identifying the persuasive power of an argument. We further propose DISCOC to inject and fuse the sentence-level structural discourse information with contextualized features derived from large-scale language models. Experimental results and extensive analysis show that the attention and gate mechanisms that explicitly model contexts and texts can indeed help the argument impact classification task defined by Durmus et al. (2019), and discourse structures among the context path of the claim to be classified can further boost the performance.

1 Introduction

It is an interesting natural language understanding problem to identify the impact and the persuasiveness of an argument in a conversation. Previous works have shown that many factors can affect the persuasiveness prediction, ranging from textual and argumentation features (Wei et al., 2016), style factors (Baff et al., 2020), to the traits of source or audience (Durmus and Cardie, 2018, 2019; Shmueli-Scheuer et al., 2019). Discourse relations, such as *Restatement* and *Instantiation*, among arguments reveal logical structures of a debate conversation. It is natural to consider using the discourse structure to study the argument impact.

Durmus et al. (2019) initiated a new study of the influence of discourse contexts on determining argument quality by constructing a new dataset *Kialo*.

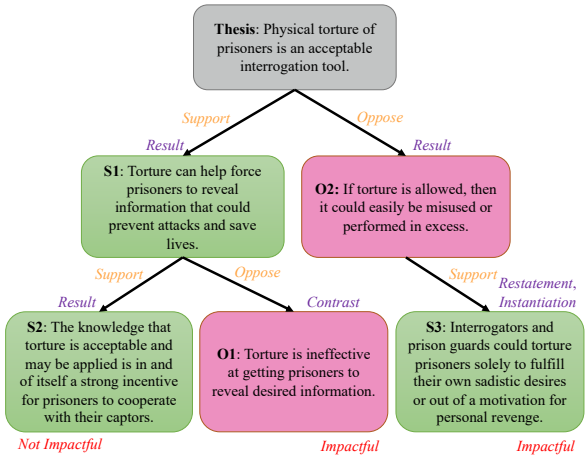

Figure 1: Example of an argument tree from *Kialo*. Stances, impact labels, and discourse relations are annotated in orange, red, and violet respectively.

As shown in Figure 1, it consists of arguments, impact labels, stances where every argument is located in an argument tree for a controversial topic. They argue contexts reflect the discourse of arguments and conduct experiments to utilize historical arguments. They find BERT with flat context concatenation is the best, but discourse structures are not easily captured by this method because it is difficult to reflect implicit discourse relations by the surface form of two arguments (Prasad et al., 2008; Lin et al., 2009; Xue et al., 2015; Lan et al., 2017; Varia et al., 2019). Therefore, there is still a gap to study how discourse relations and their sequential structures or patterns affect the argument impact and persuasiveness prediction.

In this paper, we acquire discourse relations for argument pairs with the state-of-the-art classifier for implicit discourse relations. Then we train a BiLSTM whose input is the sequence of discourse relations between two adjacent arguments to predict the last argument's impact, and the performance is comparable to that of a BiLSTM on raw text. This indicates that a sequence of discourse re-

* This work was done when Xin Liu was an intern at Huawei Noah's Ark Lab.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3958–3969
August 1–6, 2021.

lations is one of the essential factors for identifying the persuasive power of an argument. Based on this intuition, we further propose a new model called DISCOC (**Dis**course **C**ontext **O**riented **C**lassifier) to explicitly produce discourse-dependent contextualized representations, fuse context representations in long distances, and make predictions. By simple finetuning, our model beats the backbone RoBERTa (Liu et al., 2019) over 1.67% and previous best model BERT over 2.38%. Extensive experiments show that DISCOC results in steady increases when longer context paths with discourse structures, e.g., stances and discourse relations, are provided. On the contrary, encoders with full-range attentions are hard to capture such interactions, and narrow-range attentions cannot handle complex contexts and even become poisoned.

Our contributions can be highlighted as follows:

1. To the best of our knowledge, we are the first to explicitly analyze the effect of discourse among contexts and an argument on the persuasiveness.

2. We propose a new model called DISCOC to utilize attentions to imitate recurrent networks for sentence-level contextual representation learning.

3. Fair and massive experiments demonstrate the significant improvement; detailed ablation studies prove the necessities of modules.

4. Last, we discover distinct discourse relation path patterns in a machine learning way and conduct consistent case studies.

Code is publicly released at https://github.com/HKUST-KnowComp/DisCOC.

2 Argument Tree Structure

2.1 Overview

Kialo dataset is collected by Durmus et al. (2019), which consists of 47,219 argument claim texts from *kialo.com* for 741 controversial topics and corresponding impact votes. Arguments are organized as tree structures, where a tree is rooted in an argument thesis, and each node corresponds to an argument claim. Along a path of an argument tree, every claim except the thesis was made to either support or oppose its parent claim and propose a viewpoint. As shown in Figure 1, an argument tree is rooted at the **thesis** "Physical torture of prisoners is an acceptable interrogation tool.". There is one claim to support this thesis (**S1** in green) and one to oppose it (**O2** in fuchsia). Moreover, **S1** is supported by its child claim **S2** and opposed by **O1**, and **S3** holds the same viewpoint of **O2**.

Stance / Impact	Train	Validation	Test
Pro	9,158	1,949	1,953
Con	8,695	1,873	1,891
Impactful	3,021	641	646
Medium Impact	1,023	215	207
Not Impactful	1,126	252	255

Table 1: Statistics of stances and impact labels in the training, validation, and test data.

2.2 Claim and Context Path

As each claim was put in view of all its ancestral claims and surrounding siblings, the audience evaluated the claim based on how timely and appropriate it is. Therefore, the context information is of most interest to be discussed and researched in the *Kialo* dataset. We define that a **claim** denoted as C is the argumentative and persuasive text to express an idea for the audience, and a **context path** of a claim of length l is the path from the ancestor claim to its parent claim, denoted as $(C^0, C^1, \cdots, C^{l-1})$ where C^{l-1} is the parent of C. For simplicity, we may use C^l instead of C without causing ambiguity. The longest path of C starts from the thesis. Statistically, the average length of the longest paths is 3.5.

2.3 Argument Stance

In a controversial topic, each argument claim except the thesis would have a stance, whether to support or oppose the argument thesis or its parent claim. In *Kialo*, users need to directly add a **stance** tag (*Pro* or *Con*) to show their agreement or disagreement about the chosen parent argument when they post their arguments. We use s^i to denote the stance whether C^i is to support or oppose its parent C^{i-1} when $i \geq 1$. The statistics of these stances are shown in Table 1.

2.4 Impact Label

After reading claims as well as the contexts, users may agree or disagree about these claims. The **impact vote** for each argument claim is provided by users who can choose from 1 to 5. Durmus et al. (2019) categorize votes into three impact classes (*Not Impactful*, *Medium Impact*, and *Impactful*) based on the agreement and the valid vote numbers to reduce noise. We can see the overall distribution from Table 1. The argument impact classification is defined to predict the impact label y of C given the claim text C and its corresponding context path $(C^0, C^1, \cdots, C^{l-1})$.

Discourse Relations	*Reason*	*Conjunction*	*Contrast*	*Restatement*	*Result*	*Instantiation*	*Chosen Alternative*
Numbers	6,559	6,421	5,718	5,343	1,355	99	23

Table 2: Statistics of predicted discourse relations.

3 Discourse Structure Analysis

3.1 Argument Impact from the Perspective of Discourse

As paths under a controversial topic are strongly related to *Comparison* (e.g., *Contrast*), *Contingency* (e.g., *Reason*), *Expansion* (e.g., *Restatement*), and *Temporal* (e.g., *Succession*) discourse relations (Prasad et al., 2008), we model the discourse structures from a view of discourse relations. The first step is to acquire discourse relation annotations. BMGF-RoBERTa (Liu et al., 2020) is the state-of-the-art model proposed to detect implicit discourse relations from raw text. In the following experiments, we use that as our annotation model to predict discourse relation distributions for each adjacent claim pair.

Specifically, for a given argument claim C^l and its context path $(C^0, C^1, \cdots, C^{l-1})$, we denote $p_{\text{disco}}(C^l) = (r^1, r^2, \cdots, r^l)$ as a **discourse relation path** such that $r^i \in \mathcal{R}$ indicates the discourse relation between C^{i-1} and C^i when $i \geq 1$. In this work, we adopt the 14 discourse relation senses in CoNLL2015 Shared Task (Xue et al., 2015) as $\mathcal{R}$. And we also define the corresponding **distributed discourse relation path** to be $p_{\text{dist}}(C^l) = (\boldsymbol{d}^1, \boldsymbol{d}^2, \cdots, \boldsymbol{d}^l)$ such that $\boldsymbol{d}^i = F(C^{i-1}, C^i)$ is the predicted discourse relation distribution between claims C^{i-1} and C^i ($i \geq 1$) by a predictive model $\mathcal{F}$. In experiments, $\mathcal{F}$ is BMGF-RoBERTa[1]. 8 out of 14 relations appear in the predictions, and the statistics of 7 frequent predictions are shown in Table 2.

As discourse contexts would affect the persuasive power of claims, we first discover the correlations between impacts and stances as well as correlations between impacts and discourse relations, illustrated in Figure 2. From the label distribution and correlations, we find there are some clear trends: 1) Stances have little influence on argument impact, but discourse relations do. Correlations indicate that it is the contents instead of standpoints that contribute to potential impacts; 2) It is a smart choice to show some examples to convince others

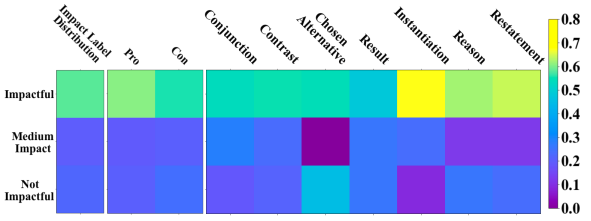

Figure 2: Impact label distributions, the correlations between labels and stances, and the correlations between labels and discourse relations. Normalization is applied to the columns.

because *Instantiation* is more relevant to *Impactful* than any other relations; 3) Similarly, explaining is also helpful to make voices outstanding; 4) *Restatement* is also positively correlated with *Impactful* so that we can also share our opinions by paraphrasing others' viewpoints to command more attention. On the contrary, *Chosen Alternative* is a risky method because the audience may object.

To investigate the role of discourse relations in impact analysis, we design a simple experiment that a single-layer BiLSTM followed by a 2-layer MLP with batch normalization predicts the impact by utilizing the distributed discourse relation path $p_{\text{dist}}(C^l)$. For the purposes of comparison and analysis, we build another BiLSTM on the raw text. Each claim has [BOS] and [EOS] tokens to clarify boundaries and we use 300-dim pretrained GloVe word embeddings (Pennington et al., 2014) and remain them fixed. We set different thresholds for context path lengths so that we can control how many discourse relations or contexts are provided. From Figure 3, discourse features can result in comparable performance, especially when longer discourse paths are provided. Instead, the model with raw text gets stuck in complex contexts.

3.2 Discourse Context Oriented Classifier

It is generally agreed that the informative context can help understand the text to be classified. However, it is still unclear how to determine whether a context is helpful. One drawback of a broader context is the increasing ambiguity, especially in the scenario of the argument context path from different users like the results shown in Figure 3. Take claims in Figure 1 for example, **S1** and **O2** give two different consequences to support or oppose

[1]The official open-source code is at https://github.com/HKUST-KnowComp/BMGF-RoBERTa. We train such a classifier on CoNLL2015 Shared Task training data, and achieve 57.57% accuracy on the test set.

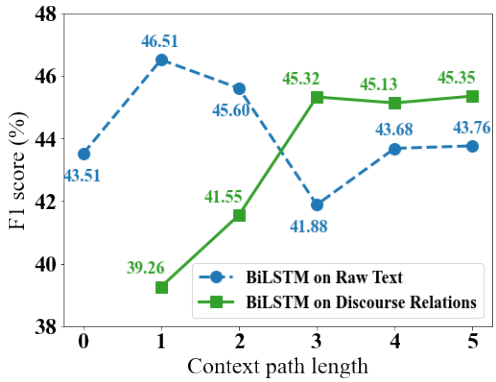

Figure 3: Performance of BiLSTM on discourse relations and BiLSTM on raw text.

the **thesis**. And **O1** objects **S1** by a contrast conclusion. It is hard to build a connection between the **thesis** and **O1** if **S1** is not given because it is challenging to build a connection between "reveal desired information" with "interrogation tool" without a precondition "Torture can help force prisoners to reveal information". On the contrary, **thesis** and **S2** are still compatible as **S2** is also a kind of result. Hence, a recurrent model with the gating mechanism that depicts pair-wise relations and passes to the following texts makes more sense.

LSTM has gates to decide whether to remember or forget during encoding, but it cannot handle long-range information with limited memory. Recently, transformer-based encoders have shown remarkable performance in various complicated tasks. These models regard sequences as fully connected graphs to learn the correlations and representations for each token. People assume that transformers can learn whether two tokens are relevant and how strong the correlation is by back-propagation. Table 3 illustrates different possible ways to aggregation context information. Transformer (Vaswani et al., 2017) and BERT (Devlin et al., 2019) adopt full-range attentions while TransformerXL (Dai et al., 2019) and XLNet (Yang et al., 2019) regard historical encoded representations as memories to reuse hidden states. SparseTransformer (Child et al., 2019), in the opposite direction, stacks hundreds of layers by narrow the attention scope by sparse factorization. Information can still spread after propagations in several layers. Inspired by these observations, we design DISCOC (**Dis**course **C**ontext **O**riented **C**lassifier) to capture contextualized features by localized attentions and imitate recurrent models to reduce noises from long distance context. As shown in Figure 4, DISCOC predicts the argument impact through three steps.

Attention	Representative	Query	Key & Value
Full	BERT	C^i	$C^0, \cdots, C^l$
Memory	XLNet	C^i	$(C^0, \cdots, C^{i-1})$
Context	SparseTransformer	C^i	C^{i-1}

Table 3: Different attention mechanisms. The Memory attention freezes the historical representations so that gradients of C^i would not propagate to the memory $(C^0, \cdots, C^{i-1})$.

3.2.1 Adjacent Claim Pair Encoding

A difficult problem in such an argument claim tree is the noise in irrelevant contexts. A claim is connected to its parent claim because of a supporting or opposing stance, but claims in long distances are not high-correlated. Based on this observation, DISCOC conduct word-level representations by encoding claim pairs instead of the whole contexts.

Given a claim C^l and its context path $(C^0, C^1, \cdots, C^{l-1})$, all adjacent pairs are coupled together, i.e., $(C^0, C^1), \cdots, (C^{l-1}, C^l)$. We can observe that each claim appears twice except the first and the last. Next, each pair (C^{i-1}, C^i) is fed into the RoBERTa encoder to get the contextualized word representations. C^0 and C^l are also encoded separately so that each claim has been encoded twice. We use $\overrightarrow{\boldsymbol{H}}^i$ to denote the encoded word representations of C^i when this claim is encoded with its parent C^{i-1}, or when it is computed alone as C^0. Similarly, $\overleftarrow{\boldsymbol{H}}^i$ is the representations when encoding (C^i, C^{i+1}), or when it is fed as C^l.

The encoding runs in parallel but we still use the term *phase* to demonstrate for better understanding. In 0-th phase, RoBERTa outputs $\overrightarrow{\boldsymbol{H}}^0$. One particular relationship between a parent-child pair is the stance, and we insert the one special token [Pro] or [Con] between them. It makes the sentiment and viewpoint of the child claim more accurate. On the other hand, discourse relations can also influence impact prediction, as reported in Section 3.1. However, discourse relations are not mutually exclusive, let alone predictions from BMGF-RoBERTa are not precise. Thus, we use the relation distributions as weights to get sense-related embeddings over 14 relations. We add additional $\boldsymbol{W}^1\boldsymbol{d}^i$ for the parent and $\boldsymbol{W}^2\boldsymbol{d}^i$ for the child except position embeddings and segment embeddings, where $\boldsymbol{d}^i$ is predicted discourse relation distribution for (C^{i-1}, C^i), $\boldsymbol{W}^1$ and $\boldsymbol{W}^2$ are trainable transformations for parents and children. Hence, RoBERTa outputs $\overleftarrow{\boldsymbol{H}}^{i-1}$ and $\overrightarrow{\boldsymbol{H}}^i$ with the concatenation of two claims, [CTX] C^{i-1} [SEP] [CLS] s^i C^i [SEP] in the i-th phase

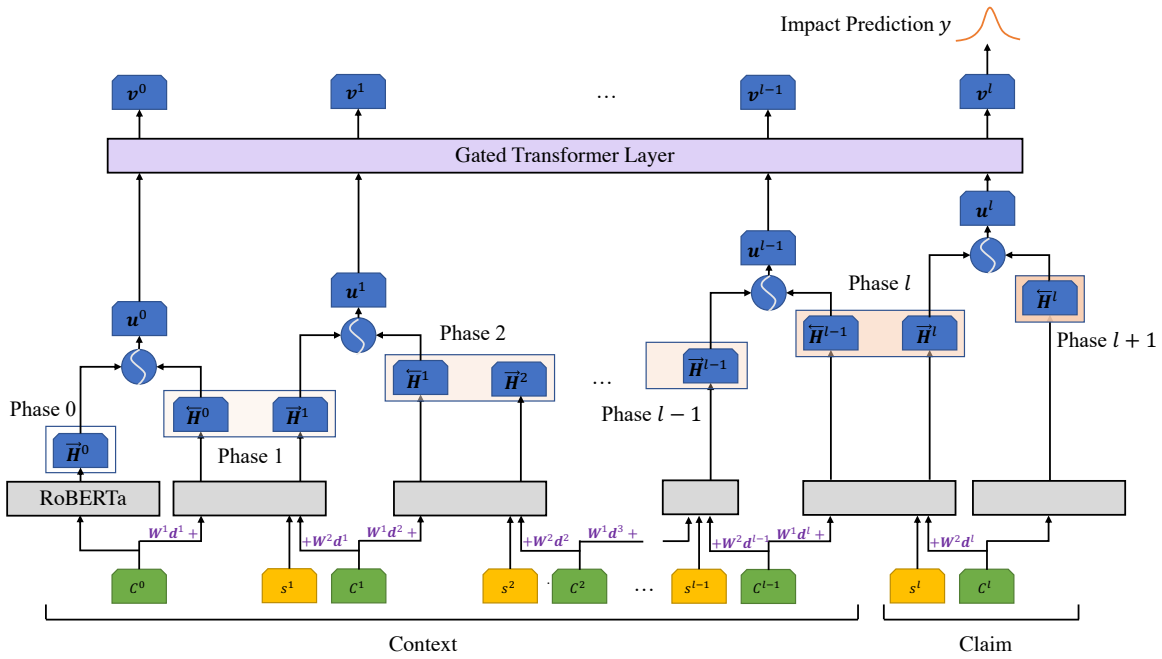

Figure 4: The architecture of DISCOC. s^i refers to the stance between C^{i-1} and C^i, $\boldsymbol{d}^i$ is the discourse relation distribution obtained from $\mathcal{F}(C^{i-1}, C^i)$. Gray boxes represent the RoBERTa encoder and the violet is a gated transformer layer. [CTX], [CLS], and [SEP] are omitted in this figure.

($i \in \{1, 2, \cdots, l\}$), where [CTX] is a special token to indicate the parent claim and distinguish from [CLS]. Its embedding is initialized as a copy embedding of [CLS] but able to update by itself. And $\overleftarrow{\boldsymbol{H}}^l$ is computed by self-attention with no context in the last phase. In the end, each claim C^i has two contextualized representations $\overleftarrow{\boldsymbol{H}}^i$ and $\overrightarrow{\boldsymbol{H}}^i$ with limited surrounding context information.

3.2.2 Bidirectional Representation Fusion

As claim representations $\{\overleftarrow{\boldsymbol{H}}^i\}$ and $\{\overrightarrow{\boldsymbol{H}}^i\}$ from RoBERTa are not bidirectional, we need to combine them and control which of them matters more. The gated fusion (Liu et al., 2020) has been shown of a better mixture than the combination of multi-head attention and layer normalization. We use it to maintain the powerful representative features and carry useful historical context information:

$$\hat{\boldsymbol{H}}^i = \text{MultiHead}(\overleftarrow{\boldsymbol{H}}^i, \overrightarrow{\boldsymbol{H}}^i, \overrightarrow{\boldsymbol{H}}^i) \quad (1)$$

$$\boldsymbol{A}_j = \text{Sigmoid}(\boldsymbol{W}^a[\overleftarrow{\boldsymbol{H}}^i, \hat{\boldsymbol{H}}^i]_j + \boldsymbol{b}^a) \quad (2)$$

$$\boldsymbol{U}^i = \boldsymbol{A} \odot \hat{\boldsymbol{H}}^i + (1 - \boldsymbol{A}) \odot \overleftarrow{\boldsymbol{H}}^i, \quad (3)$$

where MultHead is the multi-head attention operation (Vaswani et al., 2017) whose query is $\overleftarrow{\boldsymbol{H}}^i$ and key & value is $\overrightarrow{\boldsymbol{H}}^i$, $\boldsymbol{A}_j$ is the fusion gate for the j-th word embedding, $[\cdots]$ is the concatenation, $\odot$ is the element product operation, and $\boldsymbol{W}^a$ and $\boldsymbol{b}^a$ are trainable matrix and bias for fusion gating. There are two reasons why using $\overleftarrow{\boldsymbol{H}}^i$ as the key of the multi-head attention: 1) [CLS] exists in the $\overleftarrow{\boldsymbol{H}}^i$ while the replaced token [CTX] appears in $\overrightarrow{\boldsymbol{H}}^i$ when $i \neq 0$; 2) The position ids start from 0 when computing $\overleftarrow{\boldsymbol{H}}^i$. The fused [CLS] token embedding $\boldsymbol{u}^i$ is selected to represent the whole claim.

3.2.3 Context Path Information Gathering

After extracting sentence-level claim representations $\boldsymbol{u}^0, \boldsymbol{u}^1, \cdots, \boldsymbol{u}^l$, a transformer layer is used to gather longer-range context representations. The transformer layer includes a position embedding layer to provide sinusoid positional embeddings, a gated multi-head attention layer, a feed-forward network, and a layer normalization. The position embedding layer in DISCOC is different from that in the vanilla Transformer because it generates position ids in a reversed order, i.e. $l, l-1, \cdots, 0$. The reversed order is helpful to model the contexts of variable length because the claim to be classified has the same position embedding. We also choose a gate to maintain the scale instead of using a residual connection. The gated transformer can generate meaningful representations because each claim can attend any other claims and itself. On the other hand, it perfectly fits the pair-wise encoding that imitates the recurrent networks to reduce the noise in irrelevant contexts and enhance the nearest context's correlations. For example, in Figure 1, **S2** is predicted as a result of **S1** (with a probability

of 39.17%) and a restatement (with a probability of 19.81%), and **S1** is also a result of **thesis** (with a probability of 70.57%). Consequently, **S2** is high-relevant to the **thesis** as a potential result if "physical torture is acceptable", which can be captured by DISCOC. Finally, a 2-layer MLP with batch normalization is applied to $\boldsymbol{v}^l$ of the last claim to predict its impact.

4 Experiments

4.1 Baseline Models

Majority. The baseline simply returns *Impactful*.

SVM. Durmus et al. (2019) created linguistic features for a SVM classifier, such as named entity types, POS tags, special marks, tf-idf scores for n-grams, etc. We report the result from their paper.

HAN. HAN (Yang et al., 2016) computes document vectors in a hierarchical way of encoding and aggregation. We replace its BiGRU with BiLSTM for the sake of comparison. And we also extend it with pretrained encoders and transformer layers.

Flat-MLMs. Pretrained masked languages, e.g., RoBERTa, learn word representations and predict masked words by self-attention. We use these encoders to encode the flat context concatenation like [CTX] C^0 [SEP] [CTX] $\cdots$ [CTX] C^{l-1} [SEP] as Segment A and [CLS] C^l [SEP] as Segment B. After getting [CTX] and [CLS] representations, a gated transformer layer and a MLP predict impacts. As for XLNet, we follow its default setting so that [CTX] and [CLS] are located at the end of claims.

Interval-MLMs. Flat-MLMs regard the context path as a whole segment and ignore the real discourse structures except the adjacency, e.g., distances between two claims are missing. We borrow the idea from BERT-SUM (Liu and Lapata, 2019): segment embeddings of C^i are assigned depending on whether the distance to C^l is odd or even.

Context-MLMs. We also compare pretrained encoders with context masks. A context mask is to localize the attention scope from the previous to the next. That is, C^i can attends words in C^{i-1} and C^{i+1} except for itself if $1 \leq i < l$; C^0 can only attend C^0, C^1, and C^l can only attend C^{l-1}, C^l.

Memory-MLMs. XLNet utilizes memory to extend the capability of self-attention to learn super long historical text information. We also extend Flat-MLMs under this setting.

4.2 Model Configuration and Settings

We use pretrained base models [2] in DISCOC and baselines. We follow the same finetuning setting: classifiers are optimized by Adam (Kingma and Ba, 2015) with a scheduler and a maximum learning rate 2e-5. The learning rate scheduler consists of a linear warmup for the 6% steps and a linear decay for the remaining steps. As for BiLSTM and HAN, the maximum learning rate is 1e-3. The hidden state dimension of linear layers, the hidden units of LSTM layers, and projected dimensions for attention are 128. The number of the multi-head attention is set as 8. Dropout is applied after each layer and the probability is 0.1. We pick the best context path length l for each model by grid search from 0 to 5 on validation data with the batch size of 32 in 10 epochs. Each model runs five times.

4.3 Argument Impact Classification

Table 4 shows experimental results of different models. It is not surprising that neural models can easily beat traditional feature engineering methods in overall performance. But linguistic features still bring the highest precision. We also observe a significant 3.49% improvement with context vectors aggregating in HAN-BiLSTM compared with the simple BiLSTM. This indicates that it is necessary to model contexts with higher-level sentence features. Models with pretrained encoders benefit from representative embeddings, and HAN-RoBERTa achieves a gain of 5.49%. Flat context paths contain useful information to help detect the argument impact, but they also involve some noise from unrelated standpoints. Interval segment embeddings do not reduce noise but make BERT confused. It is counterintuitive that the segment embeddings depend on whether the distance is odd or even because BERT uses these for next sentence prediction. Since XLNet uses relative segment encodings instead of segment embeddings, Interval-XNet is better than Flat-XLNet in all three metrics. On the other hand, context masks bring another side effect for BERT, RoBERTa, and XLNet. Although these masks limit the attention scope at first sight, distant word information is able to flow to words with the increment of transformer layers. As a result, the uncertainty and attention bias increase after adding context masks. The memory storing context representations is also not helpful. The main reason is

[2] BERT-base-uncased, RoBERTa-base, and XLNet-base-cased are downloaded from huggingface.co

Model	Precision	Recall	F1
Majority	19.43	33.33	24.55
SVM (Durmus et al., 2019)	**65.67**	38.58	35.42
BiLSTM	46.94 ± 1.08**	46.64 ± 0.71**	46.51 ± 1.11**
HAN-BiLSTM	51.93 ± 1.37**	49.08 ± 1.52**	50.00 ± 1.49**
HAN-BERT	53.72 ± 0.80**	53.45 ± 0.51**	53.46 ± 0.47**
HAN-RoBERTa	55.71 ± 1.12**	55.95 ± 0.90**	55.49 ± 0.62**
HAN-XLNet	53.91 ± 0.96**	55.56 ± 1.59**	54.53 ± 1.22**
BERT (Durmus et al., 2019)	57.19 ± 0.92	55.77 ± 1.05**	55.98 ± 0.70**
Flat-BERT	57.34 ± 1.56	57.07 ± 0.74*	56.75 ± 0.82**
Flat-RoBERTa	58.11 ± 1.34	56.40 ± 0.61**	56.69 ± 0.63**
Flat-XLNet	55.86 ± 1.74*	56.20 ± 1.17**	55.57 ± 0.95**
Interval-BERT	55.56 ± 2.03*	55.52 ± 1.44**	55.34 ± 1.50**
Interval-RoBERTa	58.31 ± 0.89	56.46 ± 1.44*	56.61 ± 1.24*
Interval-XLNet	57.54 ± 0.50	56.78 ± 1.63*	56.52 ± 1.00**
Context-BERT	54.96 ± 0.93**	56.09 ± 0.83**	55.44 ± 0.83**
Context-RoBERTa	57.28 ± 0.97	55.29 ± 0.26**	55.83 ± 0.54**
Context-XLNet	54.56 ± 0.71**	56.28 ± 1.22**	55.10 ± 0.72**
Memory-BERT	54.33 ± 0.83**	57.57 ± 0.67*	55.22 ± 0.61**
Memory-RoBERTa	55.08 ± 0.89**	55.55 ± 1.59**	54.76 ± 1.38**
Memory-XLNet	55.44 ± 1.15**	55.45 ± 1.25**	54.91 ± 0.96**
DISCOC	57.90 ± 0.70	**59.41 ± 1.41**	**58.36 ± 0.52**

Table 4: The averages and standard deviations of different models on the argument impact classification. The marker * refers to p-value < 0.05 and the marker ** refers to p-value < 0.001 in t-test compared with DISCOC.

that the last claim's update signal can not be used to update previous context representations. That is, Memory-models degenerate to models with frozen path features or even worth. DISCOC that we proposed can capture useful contexts and fuse in a comprehensive manner. Finally, DISCOC outperforms the second best model Flat-BERT over 1.61% and its backbone Flat-RoBERTa over 1.67%, the previous best model BERT by 2.38%.

4.4 Ablation Study

Influence of the Context Path Length

Different claims have different contexts. We only report the best performance with a fixed maximum context path length in Table 4. Figure 5 shows F1 scores of models with different hyper-parameters. DISCOC always benefits from longer discourse contexts while other models get stuck in performance fluctuation. Most models can handle one context claim, which is consistent with our idea of pair-wise encoding. DISCOC has consistent performance gains; instead, other models cannot learn long-distance structures better. Each token in Flat-RoBERTa and Interval-RoBERTa can attend all other tokens, and the two are the most competitive baselines. However, Context-RoBERTa and Memory-RoBERTa limit the attention scope to the tokens of one previous claim, making models unable to make use of long-distance context information.

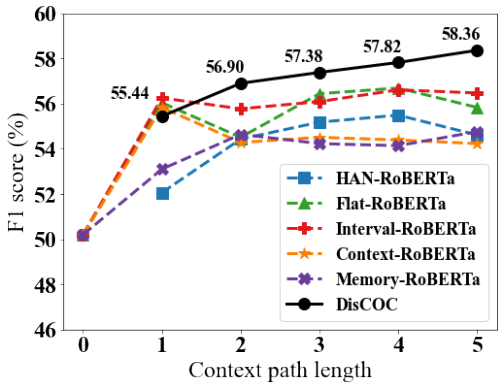

Figure 5: F1 scores of different models on varying the maximum path length.

Model	Precision	Recall	F1
DISCOC	57.90	59.41	**58.36**
DISCOC (E-BERT)	57.84	**59.46**	58.04
DISCOC (w/o StanceE)	**58.68**	58.12	57.74
DISCOC (w/o DiscoE)	57.81	58.42	57.29
DISCOC (F-BiLSTM)	58.58	57.87	57.72
DISCOC (F-Conv)	58.20	58.53	57.82
DISCOC (w/o GTrans)	56.04	54.71	54.78

Table 5: Ablation Studies of DISCOC.

RoBERTa vs. BERT

As shown in Table 4, there is little difference between the performance of RoBERTa variants and that of BERT variants. We conduct the experiment for DISCOC (E-BERT) with BERT as the encoder reported in Table 5. Its performance has achieved a significant boost over 1.29% despite the small gap between itself and DISCOC.

Impactful	Medium Impact	Not Impactful
Reason-Contrast	Conjunction-Reason	Restatement-Reason
Restatement	Conjunction-Contrast	Contrast-Restatement
Reason	Contrast-Conjunction	Chosen Alternative
Restatement-Conjunction	Conjunction-Restatement	Restatement-Restatement
Restatement-Contrast	Contrast-Contrast	Reason-Restatement
Contrast-Instantiation	Contrast-Reason	Chosen Alternative-Reason
Conjunction-Instantiation	Conjunction-Conjunction-Restatement	Contrast
Restatement-Restatement	Conjunction-Restatement-Conjunction	Chosen Alternative-Conjunction
Reason-Conjunction	Conjunction-Reason-Conjunction	Result-Reason
Restatement-Result	Conjunction-Conjunction	Chosen Alternative-Restatement

Table 6: Discourse path patterns that corresponding to the largest top 10 coefficients of the binary LR.

Are Stances and Discourse Senses Helpful?

We also remove either the stance token embedding or the discourse sense embeddings from DISCOC. The results in Table 5 suggest that both sides of structures are essential for modelling the correlation between the parent claim and the child claim. By comparison, discourse sense embeddings are more vital.

Are Gated Transformers Necessary?

We add a gated transformer layer to gather sentence-level vectors. Such gathering is necessary for the proposed framework because each claim can only attend limited contexts. BiLSTM and convolutions can also be used for this purpose, so we replace the gated transformer layer with a BiLSTM or a convolutional layer. Moreover, we also remove it to make predictions by $\boldsymbol{u}^l$ directly. The results in Table 5 show that the gated transformer is the irreplaceable part of DISCOC because it retains the contextualized representations and remains their scales. Simple removing it hurts recall enormously.

4.5 What Makes Claims Impactful?

High-coefficient Discourse Relation Patterns

We use Logistic Regression to mine several interesting discourse relation patterns. Detailed settings are described in Appendix A, and results including the most high-coefficient patterns are listed in Table 6. We observe that some discourse relation path patterns are distinguishing for classifying individual impact labels. *Instantiation* is a typical relation that only occurs in the top patterns of *Impactful*. Also, *Restatement* is relatively frequent for *Impactful* (5 of top 10), but it is the relation between the grandparent and the parent. Providing additional resources (*Restatement-Result*) or objecting others' repetitions (*Restatement-Contrast*) can increase the persuasive power. For the *Medium Impact* class, its top 10 significant patterns are the longest on average. That indicates some views are usually considered ordinary in complex structures. *Conjunction* is the dominant relation (8 of top 10) so that we are suggested to avoid to go along with others. The case of *Not Impactful* is a little clearer, in the sense that it has a unique relation *Chosen Alternative* as one of the most significant patterns. *Restatement* also appears frequently, showing neither generalization, nor specification, nor paraphrasing of others' views can help make claims stand out.

Discourse Patterns	DISCOC	DISCOC (w/o DiscoE)
Reason-Contrast	**65.56**	43.33
Restatement	56.63	**57.59**
Reason	**58.91**	54.96
Conjunction-Reason	**78.97**	72.14
Conjunction-Contrast	**80.64**	66.17
Contrast-Conjunction	**55.15**	42.38
Restatement-Reason	**38.00**	37.35
Contrast-Restatement	66.10	**76.24**
Chosen Alternative	**73.33**	42.86
All	**59.04**	58.06

Table 7: F1 score differences between two best models on top 9 discourse relation patterns and all patterns.

Case Study

In Appendix A, we define $Pr(r^1, \cdots, r^l)$ as the joint probability to generate the discourse relation path $(r^1, \cdots, r^l)$ given the context $(C^0, C^1, \cdots, C^{l-1})$ and the claim C^l. For example, the $Pr(\textit{Reason}, \textit{Contrast})$ is 56.59% which corresponds to an *Impactful* claim "There is no evidence for this" with its parent claim "Our bodies know how to recognise and process current foods; changing them through genetic modification will create health issues". Furthermore, we find 5 of top 5 and 8 of top 10 are voted as *Impactful* claims after sorting based on $Pr(\textit{Reason}, \textit{Contrast})$. For a complex pattern *Restatement-Restatement* appearing in both top patterns of the *Impactful* and the *Not Impactful*, 3 cases with the maximum probabil-

ities are *Not Impactful* while the following 7 cases are *Impactful*. It is interesting that the thesis of the top 3 claims is the same discussion about an American politician. There are 25 *Impactful* claims and 22 *Not Impactful* claims in this topic, 24 of which are restatements of their parent claims. As for *Restatement-Reason*, the most top pattern of the *Not Impactful*, we find 7 of the top 10 claims relevant to politics, 2 of them about globalization, and one food-related. Therefore, there is no perfect answer in these quite controversial topics, and that is why *Restatement* and *Reason* appear frequently.

Empirical Results

On the other hand, we check the performance of testing examples to verify the effectiveness of these discourse relation patterns. We choose the best model of DISCOC, whose F1 score is 59.04% as well as the best model of DISCOC (w/o DiscoE) whose F1 score is 58.06%. We select testing examples with specific discourse patterns, and performance differences are shown in Table 7. DISCOC benefits from 7 of the top 9 patterns and the performance margins are even more significant than the improvement of the overall results. Without giving discourse relation patterns, the model still has trouble capturing such implicit context influences. Empirical results support our idea that implicit discourse relations could affect the persuasiveness.

5 Related Work

There is an increasing interest in computational argumentation to evaluate the qualitative impact of arguments based on corpus extracted from Web Argumentation sources such as CMV sub-forum of Reddit (Tan et al., 2016). Studies explored the importance and effectiveness of various factors on determining the persuasiveness and convincingness of arguments, such as surface texture, social interaction and argumentation related features (Wei et al., 2016), characteristics of the source and audience (Durmus and Cardie, 2019; Shmueli-Scheuer et al., 2019; Durmus and Cardie, 2018), sequence ordering of arguments (Hidey and McKeown, 2018), and argument structure features (Li et al., 2020). The style feature is also proved to be significant in evaluating the persuasiveness of news editorial argumentation (Baff et al., 2020). Habernal and Gurevych (2016) conducted experiments in an entirely empirical manner, constructing a corpus for argument quality label classification and proposing several neural network models.

In addition to the features mentioned above, the role of pragmatic and discourse contexts has shown to be crucial by not yet fully explored. Zeng et al. (2020) examined how the contexts and the dynamic progress of argumentative conversations influence the comparative persuasiveness of an argumentation process. Durmus et al. (2019) created a new dataset based on argument claims and impact votes from a debate platform *kialo.com*, and experiments showed that incorporating contexts is useful to classify the argument impact.

Understanding discourse relations is one of the fundamental tasks of natural language understanding, and it is beneficial for various downstream tasks such as sentiment analysis (Nejat et al., 2017; Bhatia et al., 2015), machine translation (Li et al., 2014) and text generation (Bosselut et al., 2018). Discourse information is also considered indicative for various tasks of computational argumentation. Eckle-Kohler et al. (2015) analyzed the role of discourse markers for discriminating claims and premises in argumentative discourse and found that particular semantic group of discourse markers are highly predictive features. Hidey and McKeown (2018) concatenated sentence vectors with discourse relation embeddings as sentence features for persuasiveness prediction and showed that discourse embeddings helped improve performance.

6 Conclusion

In this paper, we explicitly investigate how discourse structures influence the impact and the persuasiveness of an argument claim. We present DISCOC to produce discourse-dependent contextualized representations. Experiments and ablation studies show that our model improves its backbone RoBERTa around 1.67%. Instead, HAN and other attention mechanisms bring side effects. We discover distinct discourse relation path patterns and analyze representatives. In the future, we plan to explore discourse structures in other NLU tasks.

Acknowledgements

This paper was supported by the NSFC Grant (No. U20B2053) from China, the Early Career Scheme (ECS, No. 26206717), the General Research Fund (GRF, No. 16211520), and the Research Impact Fund (RIF, No. R6020-19 and No. R6021-20) from the Research Grants Council (RGC) of Hong Kong, with special thanks to the Huawei Noah's Ark Lab for their gift fund.

References

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the persuasive effect of style in news editorial argumentation. In *ACL*, pages 3154–3160.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *EMNLP*, pages 2212–2218.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *NAACL-HLT*, pages 173–184.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *CoRR*, abs/1904.10509.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*, pages 2978–2988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *NAACL-HLT*, pages 1035–1045.

Esin Durmus and Claire Cardie. 2019. Modeling the factors of user success in online debate. In *WWW*, pages 2701–2707.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. The role of pragmatic and discourse context in determining argument impact. In *EMNLP-IJCNLP*, pages 5667–5677.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *EMNLP*, pages 2236–2242.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *EMNLP*, pages 1214–1223.

Christopher Hidey and Kathleen R. McKeown. 2018. Persuasive influence detection: The role of argument sequencing. In *AAAI*, pages 5173–5180.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *EMNLP*, pages 1299–1308.

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. In *EMNLP*, pages 8905–8912.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Assessing the discourse factors that influence the quality of machine translation. In *ACL*, pages 283–288.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *ACL*, pages 343–351.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the importance of word and sentence representation learning in implicit discourse relation classification. In *IJCAI*, pages 3830–3836.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP-IJCNLP*, pages 3728–3738.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Bita Nejat, Giuseppe Carenini, and Raymond Ng. 2017. Exploring joint neural model for sentence level discourse parsing and sentiment analysis. In *SIGDIAL*, pages 289–298.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The penn discourse treebank 2.0. In *LREC*.

Michal Shmueli-Scheuer, Jonathan Herzig, David Konopnicki, and Tommy Sandbank. 2019. Detecting persuasive arguments based on author-reader personality traits and their interaction. In *ACM-UMAP*, pages 211–215.

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 *WWW*, pages 613–624.

Siddharth Varia, Christopher Hidey, and Tuhin Chakrabarty. 2019. Discourse relation prediction: Revisiting word pairs with convolutional networks. In *SIGdial*, pages 442–452.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *ACL*, pages 195–200.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *CoNLL*, pages 1–16.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5754–5764.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*, pages 1480–1489.

Jichuan Zeng, Jing Li, Yulan He, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2020. What changed your mind: The roles of dynamic topics and discourse in argumentation process. In *WWW*, pages 1502–1513.

A Discourse Relation Path Patterns

To explicitly explore important high-order discourse relation patterns, we model the process of yielding a concrete discourse relation path $p_{\text{disco}}(C^l) = (r^1, \cdots, r^l)$ as a generative process. For a given context path $(C^0, C^1, \cdots, C^{l-1})$ and the claim C^l, we define the **pattern set** as all possible patterns connected to C^l. Mathematically, it is denoted as $\mathcal{P} = \sum_{i=1}^{l} \bigtimes_{j=i}^{l} \mathcal{R}$, where $\bigtimes$ is the Cartesian product.

We assume that every $r^i \in p_{\text{disco}}(C^l)$ is independent and identically distributed (i.i.d). Under this assumption, the joint probability of a given path of discourse relations $(r^1, \cdots, r^l)$ is

$$Pr(r^1, \cdots, r^l) = \Pi_{i=1}^{l} \boldsymbol{d}^i[r^i], \tag{4}$$

where $\boldsymbol{d}^i$ is the discourse relation distribution between C^{i-1} and C^i, $\boldsymbol{d}^i[r^i]$ is the probability of a specific relation sense r^i. Observing the consistently increased performance of BiLSTM on discourse relations in Figure 3 when l starts from 1 to 3 and no noticeable enhancement with longer contexts, we analyze path-generated distributions for up to three previous claims. We compute the joint probabilities $Pr(r^l), Pr(r^{l-1}, r^l), Pr(r^{l-2}, r^{l-1}, r^l)$ respectively and then concatenate these probabilities to get path pattern features $\boldsymbol{x} \in \mathbb{R}^{(|\mathcal{R}|+|\mathcal{R}|^2+|\mathcal{R}|^3)}$ where each dimension of $\boldsymbol{x}$ corresponds to the probability of a pattern belonging to $\mathcal{P}$. Next, the feature vector $\boldsymbol{x}$ is fed into a logistic regression (LR) model to train a one-vs-rest binary classifier for each of the three impact labels.

We report the largest top 10 coefficients of converged LR models in Table 6. Some relation path patterns are shown distinguishing for classifying individual impact labels. Coefficients vary differently among different LRs except for *Restatement-Restatement*, which occurs in both *Impactful* and *Not Impactful*. In general, *Instantiation* is a typical relation that only occurs in the top patterns of *Impactful*. Also, *Restatement* is relatively frequent for *Impactful* (5 of top 10), but it is the relation between the grandparent and the parent. Providing additional resources (*Restatement-Result*) or objecting others' repetitions (*Restatement-Contrast*) can increase the persuasive power. For the *Medium Impact* class, its top 10 significant patterns are the longest on average. That indicates some views are usually considered ordinary in complex structures. *Conjunction* is the dominant relation (8 of top 10) so that we are suggested to avoid to go along with others. The case of *Not Impactful* is a little clearer, in the sense that it has a unique relation *Chosen Alternative* as one of the most significant patterns. *Restatement* also appears frequently, showing that neither generalization, nor specification, nor paraphrasing of others' views can help make claims stand out. These interesting correlations between discourse relation path patterns and argument quality could be further analysis from the linguistic perspective in future works.