AQE: Argument Quadruplet Extraction via a Quad-Tagging Augmented Generative Approach

Jia Guo[*†1,2] **Liying Cheng**[*1] **Wenxuan Zhang**[1] **Stanley Kok**[2] **Xin Li**[1] **Lidong Bing**[1]
[1]DAMO Academy, Alibaba Group [2]School of Computing, National University of Singapore
guojia@u.nus.edu, skok@comp.nus.edu.sg
{liying.cheng, saike.zwx, xinting.lx, l.bing}@alibaba-inc.com

Abstract

Argument mining involves multiple sub-tasks that automatically identify argumentative elements, such as claim detection, evidence extraction, stance classification, etc. However, each subtask alone is insufficient for a thorough understanding of the argumentative structure and reasoning process. To learn a complete view of an argument essay and capture the interdependence among argumentative components, we need to know *what* opinions people hold (i.e., claims), *why* those opinions are valid (i.e., supporting evidence), *which* source the evidence comes from (i.e., evidence type), and *how* those claims react to the debating topic (i.e., stance). In this work, we for the first time propose a challenging argument quadruplet extraction task (AQE), which can provide an all-in-one extraction of four argumentative components, i.e., claims, evidence, evidence types, and stances. To support this task, we construct a large-scale and challenging dataset. However, there is no existing method that can solve the argument quadruplet extraction. To fill this gap, we propose a novel quad-tagging augmented generative approach, which leverages a quadruplet tagging module to augment the training of the generative framework. The experimental results on our dataset demonstrate the empirical superiority of our proposed approach over several strong baselines. [1]

1 Introduction

The argument plays an important role in a wide range of human activities (Yuan et al., 2021), from casual discussions (Boltužić and Šnajder, 2015; Abbott et al., 2016; Dusmanu et al., 2017) to legal negotiations (Mochales and Moens, 2011; Poudyal, 2017; Niculae et al., 2017; Teruel et al., 2018), where multiple parties formulate reasons and draw conclusions. Computational argumentation, as a growing research field, aims to automatically identify and extract the argument components presented in natural language and to predict the relationships among them (Cabrio and Villata, 2018). Given the intricate nature of the reasoning process in argumentation, identifying the various components involved and their inter-dependencies allows us to gain a deep and comprehensive understanding of the argumentative structure, thus providing valuable information for downstream applications (Lawrence and Reed, 2019).

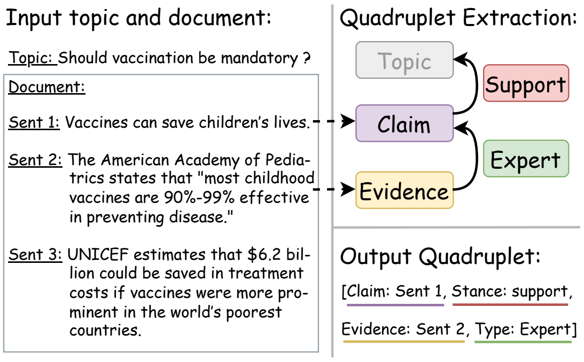

Input topic and document:
Topic: Should vaccination be mandatory ?
Document:
Sent 1: Vaccines can save children's lives.
Sent 2: The American Academy of Pediatrics states that "most childhood vaccines are 90%-99% effective in preventing disease."
Sent 3: UNICEF estimates that $6.2 billion could be saved in treatment costs if vaccines were more prominent in the world's poorest countries.
Quadruplet Extraction:
Topic
Support
Claim
Expert
Evidence
Output Quadruplet:
[Claim: Sent 1, Stance: support,
Evidence: Sent 2, Type: Expert]

Figure 1: A simplified example of argument quadruplet extraction (AQE) task from our dataset. Given the topic and a document containing multiple sentences, Sent 1 is a claim supporting the given topic, Sent 2 is a piece of expert evidence supporting the extracted claim.

Existing argument mining (AM) works focus on AM subtasks with one or a subset of the argument components, such as: claim extraction (Aharoni et al., 2014; Levy et al., 2014), evidence extraction (Rinott et al., 2015; Singh et al., 2019), evidence classification (Liga, 2019; Afrin et al., 2020), stance detection (Hasan and Ng, 2014; Bar-Haim et al., 2017; Hardalov et al., 2022), claim-evidence pair extraction (Cheng et al., 2022), argument pair extraction (Cheng et al., 2020). However, each of the tasks above could only provide a partial view of the whole argumentative structure, and few of them

[*] Equally Contributed.

[†] This work was done when Jia Guo was an intern at DAMO Academy, Alibaba Group.

[1] Our codes and datasets are available at https://github.com/guojiapub/QuadTAG.

Findings of the Association for Computational Linguistics: ACL 2023, pages 932–946
July 9-14, 2023 ©2023 Association for Computational Linguistics

have provided a detailed analysis of the complex interplay of various components. In this work, our goal is to get a thorough understanding of the overall argumentative structures. Hence, we propose a novel task named *Argument Quadruplet Extraction* (AQE). Specifically, provided with a controversial topic and document, our AQE task aims to answer: (1) *what* opinions the party holds towards the topic (i.e., claim), (2) *why* those opinions are tenable (i.e., evidence), (3) *which* source the evidence comes from (i.e., evidence type), and (4) *how* these opinions react to the debating topic (i.e., stance). A simplified example in Figure 1 illustrates the input and output of our AQE task.

To facilitate the study of this AQE task, a comprehensive dataset with all argumentative components (i.e., claim, evidence, stance, and evidence type) and their relations (i.e., claim-evidence pairing relations) is needed. Although a previous dataset (Cheng et al., 2022) has included multiple argument elements, the evidence-type information has been largely ignored. Without knowing the attributes and source of supporting evidence, it is difficult to determine the persuasiveness and adequacy of a claim for decision-making. Moreover, claims supported by a variety of evidence types tend to be more convincing than those relying solely on one type of evidence (Rinott et al., 2015).

Therefore, we carefully formulate five evidence types based on references from relevant works (Addawood and Bashir, 2016; Rinott et al., 2015): `Expert`, `Research`, `Case`, `Explanation`, `Others`. Our evidence types model the general way people recognize evidence and are widely applicable to various domains, such as online debates, policy reports, and academic writing. Both objective (i.e., `Research` and `Case`) and subjective (i.e., `Expert` and `Explanation`) categories of evidence are included. To ease the labeling labor, we additionally label the type information of each piece of evidence on top of the existing IAM dataset (Cheng et al., 2022). The resulting comprehensive dataset is able to support our AQE task which takes a step forward to fully understand the argumentative structures and is named as *Quadruplet Argument Mining* (QAM) dataset.

Recently, the pre-trained generative models (e.g., Raffel et al., 2020) have shown effectiveness in information extraction (Zhang et al., 2022, 2021). However, most generative works operate at the word level and cannot learn the dependencies among sentences explicitly. To tackle the complex reasoning at the sentence level for the quadruplet extraction task, we for the first time propose a **Quad**-**T**agging **A**ugmented **G**enerative approach (QuadTAG), which leverages a novel quad-tagging approach as the augmentation module to enhance the generative framework by explicitly capturing the cross-sentence interactions for various components. The experimental results on our dataset demonstrate the effectiveness of our model over several strong baselines.

To summarize, our contributions include:

- We propose a novel AQE task to extract a more comprehensive argument term consisting of multiple components and relations from unstructured text.
- To support the investigation of the proposed task, we introduce a new dataset QAM by additionally annotating the evidence types to an existing dataset.
- We propose an integrated generative framework augmented by a quad-tagging module for the AQE task, which can well capture the interrelations among multiple argument components. We demonstrate the empirical effectiveness on the proposed challenging QAM dataset.

2 Related Work

2.1 Argument Mining Tasks

Argument Mining Subtasks As introduced earlier, there are four main elements for understanding the argument structures: *what* (i.e., claims), *why* (i.e., evidence), *which* (i.e., types) and *how* (i.e., stances). Existing works focused on either one element or a subset of the four elements. First, most earlier works only focused on *subtask extraction*. For instance, Levy et al. (2014) proposed a task of context-dependent claim detection (CDCD). In order to find the arguments supporting the extracted claims, Rinott et al. (2015) introduced the task of context-dependent evidence detection (CDED). Addawood and Bashir (2016) worked on evidence classification subtask. Hasan and Ng (2014) explored the task of stance classification. Second, Cheng et al. (2022) proposed a claim-evidence *pair extraction* (CEPE) task. Third, in terms of AM *triplet extraction* task, researchers (Persing and Ng, 2016; Eger et al., 2017; Ye and Teufel, 2021) aimed to extract claims, premises and their relations (i.e., stances) simultaneously. In this work, we take a step further by proposing the argument *quadruplet extraction* task, by incorporating the evidence type information.

Argumentation Analysis Argumentation analysis is critical to understand argumentative structures. Stab and Gurevych (2014) classified argumentative sentences into four classes: major claim, claim, premise, none. Park and Cardie (2014) proposed the task of classifying the propositions into 3 categories: unverifiable, verifiable non-experimental, and verifiable experimental. In this work, we focus on evidence classification, which has been shown in previous works that a claim can be supported using different types of evidence in different use cases (Rieke and Sillars, 1984; Seech, 1993; Rieke et al., 2005). In social media domain, Addawood and Bashir (2016) classified the evidence into six types, including: news, expert, blog, picture, other, and no evidence. For a similar data domain to our work (i.e., Wikipedia), Rinott et al. (2015) classified evidence into three categories: study, expert and anecdotal. Inspired by the above, we further define 5 types of evidence by considering the context of claims, which includes: case, expert, research, explanation, and others.

2.2 Argument Mining Models

There are mainly two general types of end-to-end models for multiple AM subtasks, one is discriminative models and the other is generative models. In terms of the discriminative models, Chernodub et al. (2019) built a BiLSTM-CNN-CRF neural sequence tagging model to identify argumentative units and to classify them as claims or premises. Cheng et al. (2021) adopted a multi-task model with table-filling approach (Miwa and Sasaki, 2014) for claim-evidence pair extraction task. In terms of generative Models, Potash et al. (2017) applied pointer network sequence-to-sequence attention modeling for a joint argument relation extraction task and argument classification task. Bao et al. (2022) employed a pre-trained BART (Lewis et al., 2020) sequence-to-sequence language model with a constrained pointer mechanism (CPM) for an AM triplet extraction task. In this work, we aim to design a novel model with good generalization ability that is able to capture the sentence-level pairing relation explicitly by combining both discriminative and generative models.

3 QAM Dataset

To facilitate the study of the proposed argument quadruplet extraction (AQE) task, we create a fully annotated dataset based on the IAM dataset (Cheng et al., 2022). We first describe the background of the original IAM dataset, followed by our data processing and human annotation details.

3.1 The Original IAM Datset and Data Processing

As described in Cheng et al. (2022), the IAM dataset is collected from English Wikipedia, which covers 123 debating topics. This dataset is designed to support three tasks in argument mining, including claim extraction, evidence extraction, and stance classification. Thus, it is fully labeled on the three argument components (i.e., claim, evidence, stance) and their internal relations. In total, there are 69,666 sentences from 1,010 articles. 4,890 claims with stances towards the given topics and 9,384 pieces of evidence together with the pairing relations of the extracted claims are labeled. We remove some invalid sentences (e.g., only symbols or numbers) from the dataset, and eliminate those documents without any claim-evidence pair. After the pre-processing, there are 34,369 sentences from 801 articles, with 3,407 claims and 8,319 pieces of evidence.

3.2 Data Annotation

With the filtered dataset, we aim to further identify the specific relations between the extracted claim and evidence sentences. This enables the extended dataset to support our AQE task and highlights the critical role of evidence types in the overall argumentative structure. The evidence type reflects how sufficiently the claims are supported. Without the evidence types, it is difficult to determine which claim is more compelling for decision-making. For example, arguments supported by evidence from research findings are more likely to be adopted in policy decisions than those that rely on subjective reasoning to support their opinions. In the debating domain, a comprehensive speech typically incorporates various types of evidence, such as citing authoritative opinions from well-known figures or specific real-life cases. This approach enhances persuasiveness compared to relying solely on one type of evidence. Therefore, it is a non-trivial task to understand the type information of each piece of evidence in the corpus.

Type	# Evidence	% Evidence	Classification F_1
Case	1,073	12.8%	74.26
Expert	1,538	18.3%	70.18
Research	1,298	15.4%	77.71
Explanation	4,234	50.4%	89.78
Others	264	3.1%	27.91

Table 1: Statistics and analysis of evidence types.

We define 5 different evidence types based on previous work (Rinott et al., 2015) as follows:

- `Case`: specific real-life cases, events, examples, etc.
- `Expert`: authoritative opinions of a professional, authority figure, scholar, official organization, etc.
- `Research`: results or conclusions from scientific research, statistical report, survey, etc.
- `Explanation`: detailed elaboration or explanation of the claim itself, reasons or impacts of the claim.
- `Others`: none of the above.

To conduct the data annotation work, 4 professional data annotators are hired from a data company to annotate the type of each piece of evidence by following the annotation guidelines[2]. The annotators are fully compensated for their work. Each evidence sentence is independently labeled by 2 different annotators, and a third professional annotator will resolve the disagreement between the two annotators. There are 8,392 evidence sentences annotated in total and the inter annotator agreement (IAA) is measured using Cohen's Kappa with a value of 0.864.

3.3 Data Analysis

To examine the characteristics of our defined categories for evidence types, we conduct an exploratory analysis and train a simple RoBERTa-based sentence classifier for the claim and evidence sentences. The overall classification F_1 score is 81.79. The distribution and classification performance in F_1 scores of each evidence type are shown in Table 1. The classification performance on evidence sentences with `Explanation` types achieves a higher F_1 score due to sufficient data available for this type. When comparing types of `Case`, `Expert` and `Research`, the objective types `Case` and `Research` outperform the subjective type `Expert`, despite having a relatively lower portion of quantities.

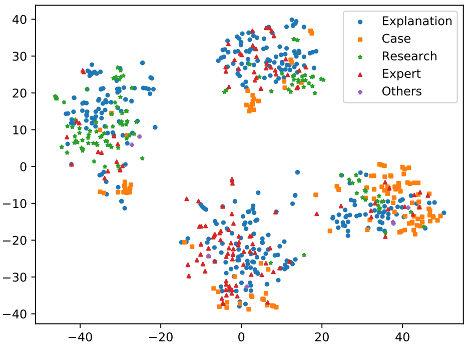


Figure 2: The t-SNE visualization for different evidence types across four topics.

To further analyze the properties of each evidence type, we use t-SNE algorithm (van der Maaten and Hinton, 2008) to visualize the evidence sentences in two-dimensional space. Specifically, we randomly select four topics that have a relatively higher amount of evidence sentences: "*Should we support family education?*", "*Should alcohol be forbidden?*", "*Should intellectual property rights be abolished?*" and "*Should we fight for the Olympics?*". It can be observed from Figure 2 that the distributions of evidence types vary significantly across different topics. Furthermore, evidence sentences of types `Case` and `Research` demonstrate distinct characteristics and exhibit clear clustering within the same topic. Conversely, evidence sentences of types `Explanation` and `Expert` show some overlap and are comparatively more challenging to differentiate. This confirms that the evidence types pose distinct challenges, thereby indicating the highly demanding nature of performing our proposed AQE task.

4 Task Formulation

More formally, given a document $\mathcal{D} = [s^1, s^2, \dots, s^n]$ with n sentences and its topic sentence s^0, our task aims to extract the set of argumentative quadruplets $\mathcal{Q} = \{q_k | q_k = (s^c_k, s^e_k, a_k, t_k)\}_{k=1}^{|\mathcal{Q}|}$ from the document $\mathcal{D}$, where $s^c_k, s^e_k \in \mathcal{D}$ $(c, e \in \{1, \dots, n\})$ respectively denote the claim sentence and evidence sentence. $a_k \in A$ represents the stance of the current claim sentence s^c_k to the topic sentence s^0, $A = \{\texttt{Support}, \texttt{Against}\}$ is the set for stance labels. $t_k \in T$ denotes the evidence type for the quadruplet q_k. $T = \{\texttt{Expert}, \texttt{Research}, \texttt{Case}, \texttt{Explanation}, \texttt{Others}\}$ is the set of all evidence categories.

[2] More detailed annotation guidelines and examples are shown in Appendix A.

5 Model

Distinct from existing subtasks of argument mining, our argument quadruple extraction (AQE) task brings unique challenges to current methods. It requires not only good compatibility to accommodate each argument component well but also building up the shared modeling capacities that are conducive to each subtask. The emergence of pre-trained generative model presents us with a good choice as a backbone framework to unify multiple targets into a general text-to-text learning paradigm. However, simply linearizing the argument quadruplets into a natural language sentence still can not fully exploit the underlying semantic dependencies among related components. To facilitate the task of argument quadruplet extraction, we propose an integrated framework augmented by a novel quad-tagging approach.

5.1 Generative Encoder

Reformulated Input Given a document $\mathcal{D} = [s^1, s^2, \ldots, s^n]$ with n sentences and its topic sentence s^0, sentence $s^i = [w^i_1, w^i_2, \ldots, w^i_m]$ contains m words. The output of AQE task requires identifying a sentence pair with the associated stance label and evidence type. However, when adapting to the text generation framework, it is inefficient to generate the original sentence of the input document during decoding especially when multiple quadruplets share the same claim or evidence sentence. To identify the sentence of interest in an efficient way and reduce the searching space of outputs, we assign each sentence with a unique symbolic ID denoted as "#i", ($i \in [1, n]$), and insert it at the beginning of each sentence. With this symbol, we can easily recognize each sentence by its unique ID.

For our proposed quad-tagging approach, we need to obtain the hidden representation of each sentence. Inspired by the recent success of the special marker technique in information extraction (Zhou et al., 2021), we insert two special tokens, i.e., <SS> and <SE>, at the start position and end position of the original sentence respectively, along with the symbolic ID. The contextual embedding of token <SS> computed by the pre-trained encoder model will be used as the sentence representation.

Sentence Encoding The reformulated input text for our proposed generative framework is defined as $\mathcal{I}(s^i) = [\texttt{<SS>}, \#i, w^i_1, w^i_2, \ldots, w^i_m, \texttt{<SE>}]$. We concatenate the reconstructed topic sentence and all sentences in the document as long text and feed it into the T5 encoder model. The hidden representations of each input token are calculated as follows:

$$\mathbf{H}_{enc} = \text{T5\_Encoder}([\mathcal{I}(s^0), \ldots, \mathcal{I}(s^n)]), \quad (1)$$

where $\mathbf{H}_{enc} \in \mathbb{R}^{L \times d}$ denotes the hidden representations of encoder states with length L after encoding. Specifically, we use $\mathbf{h}^i_s$ to represent the contextual token embedding of <SS> for i-th sentence, which will be used as i-th sentence embedding in our proposed framework.

5.2 Structural Generation for Argument Quadruplet Extraction

The straightforward way of transforming a learning task to text generation is to reformulate the expected outputs in the form of natural language sentences. However, our AQE task faces new challenges when directly adapting to text-to-text generation. As our AQE task requires identifying sentences of claim and evidence from the input document, directly generating the original text of the target sentences is space-consuming since the text can be easily retrieved from the given input document. Besides, a claim sentence is usually supported by multiple evidence sentences, repetitively generating the same claim sentence for different quadruplets will inevitably cause redundant output and a waste of computation memory.

To conduct the structural generation for our AQE task in a coherent and concise way, we first define three generative templates, i.e., $\mathcal{T}_s, \mathcal{T}_{st}, \mathcal{T}_{et}$, for the generation outputs of target sentences (s^c, s^e), stance a and evidence type t in a quadruplet, respectively. Concretely, $\mathcal{T}_s(s^i) =$ "#i" represents the original sentence using its symbolic sentence ID. $\mathcal{T}_{st}(a)$ transforms the stance label $a \in \{\texttt{Support}, \texttt{Against}\}$ to two natural language phrases, i.e., $\mathcal{T}_{st}(\texttt{Support}) =$ "*supports the topic*" and $\mathcal{T}_{st}(\texttt{Against}) =$ "*is against the topic*"[3]. We keep the original text of evidence type in the generation output, $\mathcal{T}_{et}(t) = t$, $(t \in T = \{\texttt{Expert}, \texttt{Research}, \texttt{Case}, \texttt{Explanation}, \texttt{Others}\})$. For a quadruplet $q_k = (s^c_k, s^e_k, a_k, t_k)$, we denote the expected form of its generated output as below:

$$\mathcal{T}(q_k) = \text{``}\mathcal{T}_s(s^c_k)\ \mathcal{T}_{st}(a_k)\ :\ \mathcal{T}_s(s^e_k)\ \mathcal{T}_{et}(t_k)\text{''}. \quad (2)$$

[3] We also attempt another template for stance label, i.e., $\mathcal{T}'_{st}(\texttt{Support}) =$ "*positive*" and $\mathcal{T}'_{st}(\texttt{Against}) =$ "*negative*", please see Section 6.6 for detailed analysis.

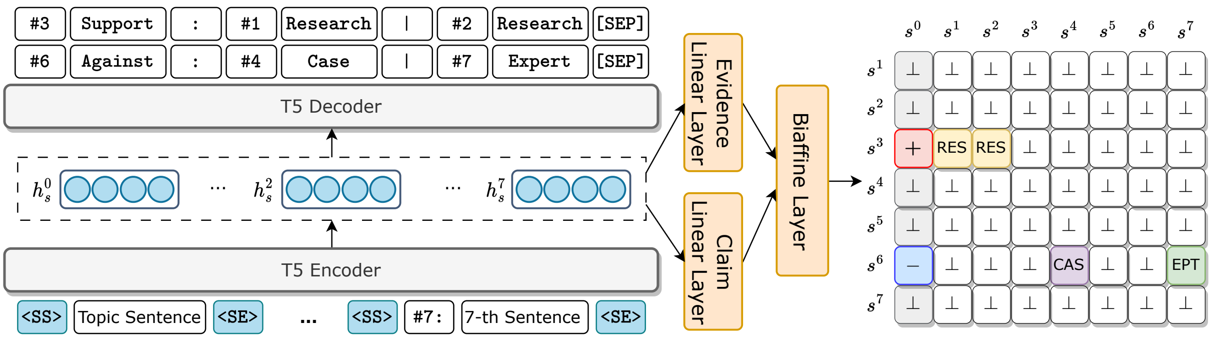

#3 Support : #1 Research | #2 Research [SEP]
#6 Against : #4 Case | #7 Expert [SEP]
T5 Decoder
h_s^0 ... h_s^2 ... h_s^7
T5 Encoder
<SS> Topic Sentence <SE> ... <SS> #7: 7-th Sentence <SE>
Evidence Linear Layer
Claim Linear Layer
Biaffine Layer
s^0 s^1 s^2 s^3 s^4 s^5 s^6 s^7
s^1 s^2 s^3 s^4 s^5 s^6 s^7
⊥
+
RES RES
−
CAS
EPT

Figure 3: The overview of our proposed QuadTAG model.

For claims supported by multiple evidence sentences, we use the symbol "|" to concatenate different evidence and evidence types, i.e., the part of "$\mathcal{T}_s(s^e_k)$ $\mathcal{T}_{et}(t_k)$". For a document with multiple claim sentences, we use a special token [SEP] to separate them. We provide a concrete example in the upper-left part of Figure 3.

5.3 The Quad-Tagging Augmented Module

To facilitate the information sharing and modeling capacities for different subtasks, we propose a novel quadruplet tagging approach built in the generative backbone to explicitly enhance the interactions among sentences. For a document with n sentences, we construct a table with the size of $n \times (n+1)$. Each entry has a tagging label y_{ij} $(i \in [1, n], j \in [1, n+1])$. As shown in Figure 3, the entries in the leftmost column of the table handle the stance detection task, i.e., $y_{i0} \in \{\perp\} \cup A$ and $\perp$ is a null label. The entries in the rest table of $n \times n$ will perform the joint tagging for the (claim, evidence, evidence type) task, i.e., $y_{ij} \in \{\perp\} \cup T$, $(j \neq 0)$. For instance, the sentence s^3 in Figure 3 is a claim sentence and supports the topic. It is supported by two evidence sentences, i.e., s^1 and s^2, both of which belong to the Research type. For a non-claim sentence, such as s^2 in the second row, all entries in the row will be tagged with a null label "$\perp$".

To obtain the tagging label y_{ij}, we adopt a biaffine transformation layer to compute the plausibility score, which has been proven effective in related tasks (Dozat and Manning, 2017). The probability of tagging label is computed as follows:

$$\begin{aligned} &\mathbf{x}_i, \mathbf{x}_j = \mathrm{Linear}_c(\mathbf{h}_s^i), \mathrm{Linear}_e(\mathbf{h}_s^j), \\ &P(y_{ij}) = \mathrm{Softmax}(\mathbf{x}_i^T \mathbf{U} \mathbf{x}_j + \mathbf{W}_i \mathbf{x}_i + \mathbf{x}_j^T \mathbf{W}_j), \end{aligned} \tag{3}$$

where $\mathbf{h}_s^i$ and $\mathbf{h}_s^j$ represent the hidden representations of i-th and j-th sentence obtained from Equation 1, respectively. Linear_e and Linear_c are linear transformation layers for claim and evidence, respectively. $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^{m \times 1}$ are the linearly transformed representations of the encoder outputs of claim and evidence. $\mathbf{U} \in \mathbb{R}^{m \times r \times m}$, $\mathbf{W}_i \in \mathbb{R}^{r \times m}$, $\mathbf{W}_j \in \mathbb{R}^{m \times r}$ are tunable weight parameters, r is the number of all possible tags in the table and $r = |A| + |T| + 1$.

To optimize the training process, we balance the label distribution of entries with null labels by negative sampling. Specifically, $\mathcal{N}$ denotes a subset of entries randomly sampled from all entries with null labels, and $\mathcal{P}$ represents all entries with non-null labels. We conduct extensive experiments to determine the optimal ratio of negative samples, i.e., $\eta = |\mathcal{N}|/|\mathcal{P}|$, please see Appendix B for more analysis. We adopt the cross-entropy loss function to train the quad-tagging augmented module:

$$\mathcal{L}_a = -\sum_{(i,j) \in \{\mathcal{N} \cup \mathcal{P}\}} \sum_{k=1}^{r} y_{ij}^k \log P_\phi(\hat{y}_{ij}^k). \tag{4}$$

5.4 Training

We finetune the pre-trained T5 model (Raffel et al., 2020) on our QAM dataset with the autoregressive loss function shown below:

$$\mathcal{L}_g = -\sum_{t=1}^{T} \log P_\theta(y_t \mid \mathbf{H}_{enc}, y_{<t}), \tag{5}$$

where y_t represents the decoder output at the t-th step, and $y_{<t}$ represents the previous outputs before the t-th step.

The final loss function for training our proposed model is defined as follows:

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_a. \tag{6}$$

For inference, we parse the predicted quadruplets Q' from the generated text sequence y' by matching them with the corresponding component slots defined in the template.

6 Experiments

6.1 Experimental Settings

The dataset is split randomly on the document level by a ratio of 8:1:1 for training, development and testing. The dataset statistics are shown in Table 2. We experiment with the pre-trained RoBERTa-base model (Liu et al., 2019) and T5-base model (Raffel et al., 2020) for our pipeline approaches and generative methods, respectively. The max length for the output text is 512. We finetune the T5-base model on our dataset for 10 epochs with a learning rate of 1e-4 and batch size of 1. We search over $\{1, 3, 5, 10\}$ for the number of negative examples used for the tagging loss and {1e-5, 3e-5, 1e-4, 3e-4} for the learning rate. The experimental results shown in Table 3 are average scores and standard deviations over three runs with different random seeds. We adopt precision, recall, and F_1 metrics for evaluation on the development and test set. For a predicted argument quadruplet $q_k' = (s_k^{c\prime}, s_k^{e\prime}, a_k', t_k')$ to be considered correct, it has to match the ground-truth quadruplet $q_k = (s_k^c, s_k^e, a_k, t_k)$ in terms of each element and their internal relations. We run all experiments on a NVIDIA Quadro RTX 8000 GPU with 48GB GPU memory.

6.2 Baselines

Since there is no existing model for the argumentative quadruplet extraction task, we introduce three competitive baselines based on recent strong pre-trained language models: the pipeline approach, the pre-trained generative model, and the tagging approach. (1) The **Pipeline Approach** tackles the integrated quadruplet extraction task by decomposing it into four subtasks handled by individual pre-trained language models. The pipeline approach facilitates the information flow between tasks by utilizing the output obtained from the preceding task as the input for the subsequent task. The decomposed subtasks for the pipeline approach are claim extraction (C), stance classification (S), evidence extraction (E), and evidence type classification (T). We introduce three variants of the pipeline approach with different orders of subtasks: C-E-T-S, C-E-S-T, and C-S-E-T. The orders are determined by the basic assumption and interdependencies among the components. Specifically, the claim forms the premise for constructing an argumentative quadruple, and the remaining three components all rely on the shared claim sentence. Moreover, the evidence type relies on both the claim and evidence sentence. For the processing details of the pipeline approach, please refer to Appendix C. (2) The **Generative Baseline** serves as a base generative model implemented on the T5-base pre-trained model (Raffel et al., 2020). It shares the same hyperparameter and template settings as our QuadTAG method. (3) The **Tagging Baseline** is the newly introduced tagging approach for our AQE task described in Section 5.3. This approach explicitly enhances the cross-sentence interactions among various components and serves as a strong discriminative baseline model. The Tagging Baseline method is trained with the encoder of the pre-trained T5-base model as the encoding backbone.

Statistics	Train	Dev	Test
# topics	96	52	53
# documents	639	80	82
# paragraphs	2,569	326	342
# claims	2,674	358	375
# pieces of evidence	6,563	808	948
# quadruplets	7,502	938	1,098

Table 2: Data statistics for the QAM dataset.

6.3 Main Results

Table 3 shows the overall performance of our proposed QuadTAG model on the AQE task compared to the aforementioned strong baselines. As shown in Table 3, our QuadTAG model outperforms all baselines by a large margin on the F_1 score for both the development and test dataset. The pipeline approaches address four subtasks sequentially by separate models. We observe that both the pipeline approach (C-E-S-T) and the pipeline approach (C-S-E-T) perform worse than the pipeline approach (C-E-T-S). This is because these two approaches additionally consider the dependencies between stance and evidence type, which renders them more susceptible to the issue of error propagation. Compared to the pipeline approaches, the end-to-end models (e.g., the generative baseline and our QuadTAG model) perform much better on three metrics. This shows that the modeling abilities developed for each subtask can be effectively transferred and leveraged for other tasks, which also implies the necessity and rationale behind the proposed AQE task in terms of empirical benefits. The tagging baseline described in Section 5.3 addresses the AQE task by treating it as a classification task. How-

Model	Dev			Test		
	Precision	Recall	F_1	Precision	Recall	F_1
Pipeline Approach (C-E-T-S)	12.02±0.95	16.13±6.95	13.33±2.42	14.02±1.40	15.77±5.34	14.40±1.29
Pipeline Approach (C-E-S-T)	11.61±0.49	11.73±1.67	11.63±0.98	13.47±1.35	11.57±1.27	12.44±1.23
Pipeline Approach (C-S-E-T)	9.51±1.51	16.11±6.67	11.40±0.40	10.74±1.58	16.05±6.86	12.50±2.27
Generative Baseline (T5-base)	17.14±2.68	16.60±2.58	16.87±2.63	21.16±3.55	18.16±2.49	19.54±2.94
Tagging Baseline (T5-base)	13.98±0.89	**18.87**±1.04	16.06±0.88	16.30±3.11	18.09±2.69	17.14±2.92
QuadTAG (Ours)	**20.55**±1.62	18.82±1.66	**19.64**±1.65	**24.47**±3.01	**19.01**±1.53	**21.39**±2.11

Table 3: Experimental results of our QuadTAG model and baselines for the AQE task.

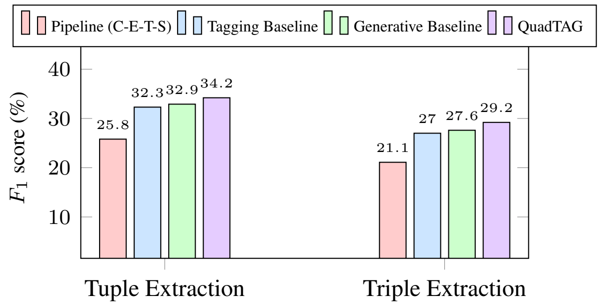


Figure 4: Performance comparison on the tuple extraction and triple extraction tasks.

ever, it still falls short of both the generative baseline and our QuadTAG model, which demonstrates the generalizability of generative models for such an integrated task with multiple diverse targets involved. Our QuadTAG model exhibits substantial improvements of 16.4% and 9.5% in terms of the F_1 score on the development and test datasets respectively when compared to the generative baseline. The experimental results demonstrate the effectiveness of our proposed augmented module, indicating that the generative model can be further enhanced by explicitly capturing the cross-sentence interactions and the semantic inter-dependencies among distinct components. Both the tagging and generative baseline in Table 3 serve as two ablations of our QuadTAG model.

6.4 Evaluation on Tuple and Triple Extraction

To further explore the differences in model capabilities, we present a performance comparison in Figure 4 focusing on the extraction of a subset of argument components. Specifically, we evaluate the performance of our model and baselines in terms of extracting the (claim, evidence) tuple and the (claim, evidence, evidence type) triple. All models are trained on the argument quadruplet dataset and evaluated on the corresponding task. We observe that both generative models (e.g., our QuadTAG model and the generative baseline) outperform the discriminative models (e.g., the pipeline approach and the tagging baseline) for the tuple extraction and triple extraction, which further confirm the superiority of the generative framework on the complex sentence-level extraction tasks. Moreover, we observe that the tagging baseline performs comparably to the generative baseline in both tasks. This finding suggests that our proposed tagging module effectively captures the cross-sentence interactions between the claim and evidence sentences, thereby enhancing the prediction of evidence types. By harnessing the strengths of both the generative model and tagging module, our model achieves superior performance and surpasses all other models.

6.5 Performance Breakdown on Subtasks

We provide the performance breakdown of our model in Table 4. We evaluate our QuadTAG model on multiple subtasks at different granularities, ranging from component extraction to triple extraction. The claim component forms the basis of a quadruplet. Given that the remaining three components rely on the claim and cannot be considered alone, comparing the model performance on different joint extractions can offer valuable insights into the error distribution within the challenging AQE task. We observe that in comparison to the claim extraction, introducing the joint extraction with evidence and stance resulted in a relative decline of 37.8% (33.08 vs. 53.20) and 26.5% (39.12 vs. 53.20), respectively. Incorporating the extraction of evidence type, the model performance for triple extraction of (claim, evidence, evidence type) decreases by 14.9% (28.16 vs. 33.08) compared to the tuple extraction of (claim, evidence). Furthermore, the overall performance of quadruplet extraction (i.e., 21.39 on F_1) is even lower than that of any of the aforementioned subtasks. The above performance degradation illustrates the challenges posed by each component and also highlights the difficulty in accurately capturing the complete quadru-

Task	Test		
	Precision	Recall	F_1
(Claim)	58.94	48.48	53.20
(Claim, Evidence)	37.79	29.42	33.08
(Claim, Stance)	43.36	35.64	39.12
(Claim, Evidence, Evidence Type)			
- *Trained on full quadruplets*	32.16	25.05	28.16
- *Trained on quadruplets with dummy stance*	31.59	24.57	27.63
(Claim, Evidence, Stance)			
- *Trained on full quadruplets*	28.02	21.74	24.48
- *Trained on quadruplets with dummy type*	26.37	19.94	22.71

Table 4: Model performance breakdown for different subtasks.

Template	Test		
	Precision	Recall	F_1
Prompt-based template **Template:** *Claim Index*: `#[c]`, *Stance*: `[a]`, *Evidence Index*: `#[e]`, *Evidence Type*: `[t]` **Example:** *Claim Index*: `#3`, *Stance*: `positive`, *Evidence Index*: `#1`, *Evidence Type*: `Research [SEP]` *Claim Index*: `#3`, *Stance* : `positive`, *Evidence Index*: `#2`, *Evidence Type*: `Research`	13.34	11.30	12.24
Order-differentiated template **Template:** `#[c]`, `#[e]`, `[t]`, `[a]` **Example:** `#3`, `#1`, `Research`, *supports the topic* `[SEP]` `#3`, `#2`, `Research`, *supports the topic*	16.11	14.29	15.15
Template: `#[e]`, `#[c]`, `[a]`, `[t]` **Example:** `#1`, `#3`, *supports the topic*, `Research [SEP]` `#2`, `#3`, *supports the topic*, `Research`	17.65	15.35	16.42
Template with other paraphrase **Template:** `#[c] [a] : #[e] [t]` **Example:** `#3 positive : #1 Research | #2 Research`	20.45	16.79	18.44

Table 5: Experimental results of our model with different templates.

plet structure. To examine the benefit gained from integrating multiple argumentative components, we manually assign a dummy value to the argument component (e.g., we set all evidence types in the QAM dataset as `Others`), and compare the model performance with the original QuadTAG model trained on the full quadruplet dataset. From Table 4, we found that both models trained with dummy values are much worse than the original model. This further emphasizes the tight interdependence of the four components. Our quadruplet extraction can benefit subtasks by introducing other associated components and facilitating the propagation of information among them.

6.6 Generative Template Design

To investigate the effects of different template designs, we evaluate the performance of our model using various templates. As shown in Table 5, the prompt-based template provides some prompting words for each component, such as "*Claim Index*" and "*Stance*". However, it achieves poorer results than other templates, which may be due to the verbose output of the prompts, causing confusion with the original target. The order-differentiated template aims to sequentially generate four components for a quadruplet. We can observe that the empirical performance varies with different generating orders. Additionally, we offer a template with alternative textual paraphrases for the stance label, which shows the comparatively lower performance than ours. We will leave the investigation into the effects of template design for future research.

7 Conclusions

In this work, we propose a novel argument quadruplet extraction (AQE) task. To facilitate this task, we annotate a large-scale quadruplet argument mining dataset (QAM) and propose a novel quad-tagging augmented generative model (QuadTAG). Extensive experimental results and analysis validate the effectiveness of our proposed model.

Acknowledgements

This work is supported by Alibaba Group through Alibaba Research Intern Program.

Limitations

For this work, we have several limitations: first, as described in Section 6.6, we found that the choice of different templates and the order of generating content will both lead to performance variation. It is worthwhile to conduct a detailed investigation on this interesting problem, however, due to the limit of pages, we only experimented with limited alternative templates. Second, our proposed AQE task shares some similarities with some tasks in other domains, which means that it is possible to adapt our proposed framework to other tasks, such as relation extraction and sentiment analysis. We will leave this for future research and demonstrate its effectiveness in other domains. Last, subject to both the economic and time cost of dataset annotation, we only expand one existing dataset for our proposed AQE task. We will explore more possibilities for dataset construction for future work.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Proceedings of LREC*.

Aseel Addawood and Masooda Bashir. 2016. "what is your evidence?" a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining, ArgMining@ACL*.

Tazin Afrin, Elaine Lin Wang, Diane Litman, Lindsay Clare Matsumura, and Richard Correnti. 2020. Annotation and classification of evidence and reasoning revisions in argumentative writing. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL*.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argument Mining, ArgMining@ACL*.

Jianzhu Bao, Yuhang He, Yang Sun, Bin Liang, Jiachen Du, Bing Qin, Min Yang, and Ruifeng Xu. 2022. A generative model for end-to-end argument mining with reconstructed positional encoding and constrained pointer mechanism. In *Proceedings of EMNLP*.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of EACL*.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL*.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of IJCAI*.

Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. 2022. Iam: A comprehensive and large-scale dataset for integrated argument mining tasks. In *Proceedings of ACL*.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. Ape: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of EMNLP*.

Liying Cheng, Tianyu Wu, Lidong Bing, and Luo Si. 2021. Argument pair extraction via attention-guided multi-layer multi-cross encoding. In *Proceedings of ACL*.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Targer: Neural argument mining at your fingertips. In *Proceedings of ACL*.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of ICLR*.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on twitter: Arguments, facts and sources. In *Proceedings of EMNLP*.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of ACL*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of NAACL*.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of EMNLP*.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Comput. Linguistics*.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*.

Davide Liga. 2019. Argumentative evidences classification and argument scheme detection using tree kernels. In *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL*.

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.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of EMNLP*.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artif. Intell. Law*.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured svms and rnns. In *Proceedings of ACL*.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argument Mining, ArgMining@ACL*.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of NAACL*.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Here's my point: Joint pointer architecture for argument mining. In *Proceedings of EMNLP*.

Prakash Poudyal. 2017. A machine learning approach to argument mining in legal documents. In *AI Approaches to the Complexity of Legal Systems*. Springer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*

Richard D Rieke and Malcolm Osgood Sillars. 1984. *Argumentation and the decision making process*. Addison-Wesley Longman.

Richard D Rieke, Malcolm Osgood Sillars, and Tarla Rai Peterson. 2005. *Argumentation and critical decision making*. Pearson/Allyn & Bacon.

Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of EMNLP*.

Zachary Seech. 1993. Writing philosophy papers.

Keshav Singh, Paul Reisert, Naoya Inoue, Pride Kavumba, and Kentaro Inui. 2019. Improving evidence detection by leveraging warrants. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING*.

Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. 2018. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *Proceedings of LREC*.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*.

Yuxiao Ye and Simone Teufel. 2021. End-to-end argument mining as biaffine dependency parsing. In *Proceedings of EACL*.

Jian Yuan, Liying Cheng, Ruidan He, Yinzi Li, Lidong Bing, Zhongyu Wei, Qin Liu, Chenhui Shen, Shuonan Zhang, Changlong Sun, Luo Si, Changjiang Jiang, and Xuanjing Huang. 2021. Overview of argumentative text understanding for ai debater challenge. In *Proceedings of NLPCC*.

Shuai Zhang, Yongliang Shen, Zeqi Tan, Yiquan Wu, and Weiming Lu. 2022. De-bias for generative extraction in unified ner task. In *Proceedings of ACL*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of ACL*.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of AAAI*.

A Detailed Annotation Guidelines and Dataset Examples

In this section, we present our detailed annotation guidelines for human annotators. Given the topic and document information, the annotators are required to assign an evidence-type label to an evidence sentence, relying on a comprehensive comprehension of the document context and how the evidence supports its claim. As mentioned in Section 3.2, we pre-define five evidence types: `Case`, `Expert`, `Research`, `Explanation` and `Others`. We present the specific definition of each type below:

Case An evidence sentence of case type supports a claim by describing or referencing real-life cases, events, and examples to strengthen the claim. All of the following rules must be met: first, it must be an event, phenomenon, or occurrence that has taken place or existed in the real world. Second, the evidence must include at least one clearly defined and specific element related to the event, such as the individuals involved, the location, the time, and other relevant details.

The difference between this type and the explanation type is that the evidence of this type is supported by real and concrete examples, while the evidence of the explanation type remains focused on high-level analysis, reasoning, or illustration.

An argument quadruplet with case evidence is shown in the first block of Table 6. Since the sentence clearly quotes the specific event (i.e., “The 1984 Summer Olympics”) and the event place (i.e., “Los Angeles”), it is considered as a real-life case to support the given claim.

Expert Expert evidence supports its claim by citing the views, comments, or suggestions of a professional, authority figure, scholar, well-known organization, official institution, representative professional group, etc. Evidence belonging to this type can be clearly identified that the opinion or assertion in the sentence comes from a specific expert or organization, and it is essential to explicitly state the name of the expert or organization in the sentence.

Besides, we have to take note of the following: first, the difference between this type and the research type is that the evidence sentences of this

Topic	Claim & Evidence	Evidence Type	Stance
Should we fight for the Olympics?	**Claim:** The Olympics increase valuable tourism, which can boost local economies. **Evidence:** The 1984 Summer Olympics in Los Angeles netted the city a $215 million operating surplus and $289 million in broadcasting fees.	Case	Support
Should animal testing be banned?	**Claim:** Some cosmetics and health care products must be tested on animals to ensure their safety. **Evidence:** The US Food and Drug Administration endorses the use of animal tests on cosmetics to assure the safety of a product or ingredient.	Expert	Contest
Should we ban unsustainable logging?	**Claim:** Deforestation is occurring all over the world and has been coupled with an increase in the occurrence of disease outbreaks. **Evidence:** A 2017 study in the American Economic Review found that deforestation substantially increased the incidence of malaria in Nigeria.	Research	Support
Should we eliminate traditional universities?	**Claim:** Traditional universities are a rite of passage to independent life. **Evidence:** This means they have to start learning or practically using lots of skills of independent adults, such as financial management, cooking, being crime-aware, networking, and solving communication problems on their own.	Explanation	Contest

Table 6: Quadruplet examples for our AQE task. Each line represents a different quadruplet with varying evidence types and stances. We highlight the signal words in the evidence sentence of different evidence types in blue.

type come from the viewpoints, opinions, judgments, etc. of authoritative persons or institutions, which are subjective arguments, while the evidence sentences of research type are objective arguments. Second, if there is an overlap with the research type, it needs to be judged according to the subject of the sentence. Third, subjective opinions, positions, judgments, and estimations from media, newspapers, publications, writings, etc., can also be labeled as the expert type.

An argument quadruplet with expert evidence is shown in the second block of Table 6. "The US Food and Drug Administration" is an authoritative federal agency, and thus is labeled as expert type.

Research Evidence of the research type strengthens a claim by referencing perspectives or findings obtained from scientific research, surveys, investigation, statistical reports, or other reputable sources, including academic journals, scientific journals, etc. At least one of the following rules must be met: (1) The evidence sentence explicitly suggests that it pertains to a study, statistical report, or survey. Alternatively, the sentence conveys information derived from research, statistics, or surveys, typically related to research conclusions, findings, statistical data, etc. Usually, the evidence sentence of this type contains some keywords, such as "The research shows", "A survey found", "According to the report", etc. (2) The evidence sentence presents a substantial amount of factual statistics or numbers derived from concrete studies, surveys, and statistics, to enhance the persuasiveness of its claim rather than relying on rough estimations.

An argument quadruplet with research evidence is shown in the third block of Table 6. This piece of evidence clearly states "A 2017 study ... found that ...", which quotes a finding of a specific study to support its claim, thus is labeled as research type.

Explanation This type of evidence enhances its claim by offering supplementary details, explanations, and elaborations on the claim sentence, as well as other relevant aspects such as the causes, consequences, and impacts associated with the claim.

An argument quadruplet with evidence of explanation type is shown in the last block of Table 6. This evidence supports its claim by expanding upon the original assertion with more details.

Others We categorize evidence sentences that do not fit into any of the aforementioned categories as "Others". However, we discourage our annotators from assigning this label, as it contributes limited information about the attribute of evidence.

With the pre-defined categories, we also ask our annotators to take note of the following:

- When encountering a sentence that is difficult to decide, it is crucial to thoroughly analyze the relationship between the evidence and the claim, along with the document context, in order to determine the appropriate type.
- It is essential to comprehensively consider the semantic relationship between the preceding and following evidence sentences.
- Multiple consecutive evidence sentences can

belong to different types depending on their content as well as their relationship with the claim and overall context.

Apart from providing the above annotation guidelines, we work closely with professional data annotators whenever they have questions or they are unsure about the labels to make sure the data annotation quality.

B The Effect of Negative Sampling Ratio

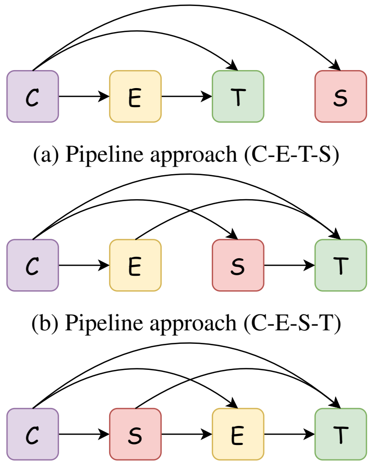
For determining the best negative ratio of the negative sampling method, we search over the range of {1,3,5,10}. As shown in Table 7, the model achieved the best performance when the negative ratio is 5.

# Negative Ratio	Precision	Recall	F_1
1	25.11	18.44	21.27
3	24.62	18.66	21.23
5	27.91	20.77	23.81
10	21.97	17.74	19.63

Table 7: Experimental results with different negative sampling ratios.

C Pipeline Processing Order

We provide the processing details in Figure 5 for pipeline approaches that handle four subtasks sequentially, including claim extraction (C), stance classification (S), evidence extraction (E), and evidence type classification (T). The arrow directions represent the input of each task.



(a) Pipeline approach (C-E-T-S)

(b) Pipeline approach (C-E-S-T)

(c) Pipeline approach (C-S-E-T)

Figure 5: The processing details of pipeline approaches.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☒ A2. Did you discuss any potential risks of your work?
No potential risk

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstract and section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All datasets are free for non-commercial usage.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
It's been discussed in the original dataset paper.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix and Section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Table 2

C ☑ Did you run computational experiments?

6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
6

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
6 and Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
6

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
3

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
3

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3