A Comprehensive and Large-Scale Dataset for Integrated Argument Mining Tasks

Anonymous ACL submission

Abstract

Traditionally, a debate usually requires a manual preparation process, including reading plenty of articles, selecting the claims, identifying the stances of the claims, seeking the evidences for the claims, etc. As the AI debate attracts more attention these years, it is worth exploring the methods to automate the tedious process involved in the debating system. In this work, we introduce a comprehensive and large dataset, which can be applied to a series of argument mining tasks, including claim extraction, stance classification, evidence extraction, etc. Our dataset is collected from over 1k articles related to 123 topics. Near 70k sentences in the dataset are fully annotated based on their argument properties (e.g., claims, stances, evidences, etc.). We further propose two new integrated argument mining tasks associated with the debate preparation process: (1) claim extraction with stance classification (CESC) and (2) claim-evidence pair extraction (CEPE). We adopt a pipeline approach and an end-to-end method for each integrated task separately. Promising experimental results are reported to show the values and challenges of our proposed tasks, and motivate future research on argument mining. [1]

1 Introduction

Debating has a long history and wide application scenarios in education field (Stab and Gurevych, 2014; Persing and Ng, 2016; Stab and Gurevych, 2017), political domain (Lippi and Torroni, 2016; Duthie et al., 2016; Menini et al., 2018), legal actions (Mochales and Moens, 2011; Grabmair et al., 2015; Teruel et al., 2018), etc. It usually involves tons of manual preparation steps, including reading the articles, selecting the claims, identifying the claim stances to the topics, looking for the evidences, etc. Since the machine has shown promising potential in processing large quantities of information in many other natural language processing tasks, it is also worthwhile to explore the methods for automating the manual process involved in debating.

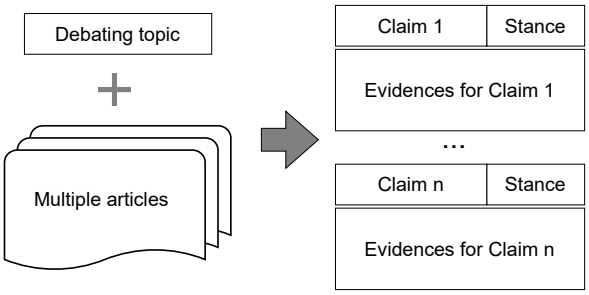

Figure 1: A flow chart showing the debating preparation process.

Argument mining (AM), as the core of a debating system (Bar-Haim et al., 2021), has received more attention in the past few years. Several AM tasks and datasets have been proposed to work towards automatic AI debate, such as: context dependent claim detection (CDCD) (Levy et al., 2014), claim stance classification (CSC) (Bar-Haim et al., 2017) , context dependent evidence detection (CDED) (Rinott et al., 2015), etc. All the above tasks are essential elements for AM and they are mutually reinforcing in the debating preparation process. In this work, we aim at automating the debating preparation process as shown in Figure 1. Specifically, providing with the debating topic and several related articles, we intend to extract the claims with its stance, and also the evidences supporting the claims.

However, none of the existing works can facilitate the study of all these tasks at the same time. Motivated by this, we introduce a comprehensive dataset to support the research of these tasks. We create our dataset by first collecting over 100 topics from online forums and then exploring over 1k articles related to these topics. All the sentences in those articles are fully-annotated following a set of carefully defined annotation guidelines. Given

[1] Our code and data are available at *URL*.

a specific topic, the annotators have to distinguish whether the given sentence is a claim to this topic and identify the relation between the selected claim and the topic (i.e., support or contest). Then given the claims, the annotators have to browse the contexts to find evidences supporting the claims. With all the labeled information, researchers can work towards these primary argument mining tasks simultaneously.

To better coordinate these individual tasks together, we propose two new integrated tasks: claim extraction with stance classification (CESC) and claim-evidence pair extraction (CEPE). Instead of treating the existing tasks (i.e., CDCD, CSC, CDED) as individual ones, the two proposed tasks can integrate the relevant primary tasks together, which are more practical and more effective in the debating preparation process. The CESC task can be divided into two subtasks: the claim detection task and the stance classification task. Intuitively, we conduct experiments on the CESC task with a pipeline approach to combine the two subtasks. As the two subtasks are mutually reinforcing each other, we also adopt an end-to-end classification model with multiple labels (i.e., support, contest, and no relation). The CEPE task is composed of the claim detection task and the evidence detection task. Similar to the annotation procedure, we apply a pipeline method to tackle this problem by first detecting the claims given the topics and then identifying the corresponding evidences of each claim. We also use a multi-task model to extract both claims and evidences as well as their pairing relation simultaneously. We conduct extensive experiments on our dataset to verify the effectiveness of our models and shed light on the challenges of our proposed tasks.

To summarize, our contributions include:

- We introduce a fully-annotated argument mining dataset and provide thorough data analysis. This is the first dataset that supports comprehensive argument mining tasks. We will release the dataset to the community.
- We are the first to propose the CESC and CEPE tasks, which are practical task settings in the argument mining field and able to enlighten future research on this.
- We conduct preliminary experiments for all proposed tasks with the new dataset. We will also release the experiment code and detailed settings to the community.

2 Related Work

In recent years, there is a tremendous amount of research effort in the computational argumentation research field (Eger et al., 2017; Bar-Haim et al., 2021), such as argument components identification (Levy et al., 2014; Rinott et al., 2015; Lippi and Torroni, 2016; Daxenberger et al., 2017), argument relation prediction (Chakrabarty et al., 2019), argument pair extraction (Cheng et al., 2020, 2021), argument quality assessment (Habernal and Gurevych, 2016; Wachsmuth et al., 2017; Gretz et al., 2020; Toledo et al., 2019), listening comprehension (Mirkin et al., 2018), etc.

Meanwhile, researchers have been exploring new datasets and methods to automate the debating preparation process, such as project debater (Slonim et al., 2021), etc. Bilu et al. (2019) work on argument invention task in the debating field to automatically identify which of these arguments are relevant to the topic. Li et al. (2020) explore the role of argument structure in online debate persuasion. Levy et al. (2014) introduce a dataset with labeled claims and work on the task of context-dependent claim detection (CDCD). Bar-Haim et al. (2017) modify Aharoni et al. (2014)'s dataset by further labeling the claim stances and tackle the problem of stance classification of context-dependent claims. Rinott et al. (2015) propose a task of detecting context-dependent evidences that support a given claim (CDED) and also introduce a new dataset for this task.

Unlike previous works with a specific focus on only one argument mining task, we introduce a comprehensive dataset that is able to support different tasks related to the debating system. Such a dataset not only enlightens future research on argument mining but also shows strong potential for various practical applications. Another difference is that existing tasks (e.g., CDCD, CDED, CSC, etc.) could be considered as subtasks in the emerging wider field of argumentation mining (Levy et al., 2014). While in this paper, we propose two integrated tasks (i.e., CESC and CEPE) incorporating the existing subtasks in the debating system, which takes a step forward to automatic AI debate.

3 Our Dataset

We introduce a large and comprehensive dataset to facilitate the study of several essential AM tasks in the debating system. We describe the collection process, annotation details and data analysis here.

	Topic: Will artificial intelligence replace humans	Claims	Stances	Evidences
1	Job opportunities will grow with the advent of AI; however, some jobs might be lost because AI would replace them.	C_1	+1	
2	Any job that involves repetitive tasks is at risk of being replaced.	C_2	+1	
3	In 2017, Gartner predicted 500,000 jobs would be created because of AI, but also predicted that up to 900,000 jobs could be lost because of it.			E_1 | E_2
4	The number of industrial robots has increased significantly since the 2000s.			E_3
5	The low operating costs of robots make them competitive with human workers.			E_3
6	In the finance sector, computer algorithms can execute stock trades much faster than a human, needing only a fraction of a second.			E_3
7	As these technologies become cheaper and more accessible, they will be implemented more widely, and humans might be increasingly replaced by AI.	C_3	+1	
8	According to Harvard Business Review, most operations groups adopting RPA have promised their employees that automation would not result in layoffs.	C_4	-1	E_4
9	AI is incredibly smart, but it will never match human creativity.	C_5	-1	

Table 1: Sample topic and labeled claims with their stances and evidences. Note that different blocks refer to the sentences from different articles, and we only extract claim-evidence pairs from the same article. For clarity, we label the indices in ascending order, which may not reflect the real indices in the dataset.

3.1 Data Collection

First, we collect 123 debating topics with a wide variety from online forums. For each topic, we explore around 10 articles from English Wikipedia with promising contents. The most number of articles explored for one topic is 16, while the least number is 2. This is because it is difficult to find enough resources for unpopular topics such as "Should nuclear waste be buried in the ground". However, most topics (i.e., 91 topics) are relatively popular with more than 8 related articles collected for each of them. In total, there are 1,010 articles collected for all the topics. After we obtain all the relevant articles, we use NLTK package (Bird et al., 2009) to split the corpus into 69,666 sentences from these articles for further annotation.

3.2 Data Annotation

The annotation process is mainly separated into two stages: (1) detecting the claims given the topics, (2) detecting the evidences given the claims. A context-dependent claim (CDC), claim in short, is a general and concise statement that directly supports or contests the given topic (Levy et al., 2014). The annotators are asked to extract the claims by following this definition. Meanwhile, the annotators have to identify the stance of the extracted claim towards the given topic. In the second stage, the annotators have to read through the context surrounding the claims, and extract the evidences following the definition that a context-dependent evidence (CDE) is a text segment that directly supports a claim in the context of the topic. Since only the surrounding sentences are content-relevant in most cases, we only search 10 to 15 sentences before and after the claim sentence to label the evidences. Note that the claim itself could be the evidence as well.

Professional data annotators are hired from a data annotation company and are fully paid for their work. Each sentence is labeled by 2 professional annotators working independently in the first round. 69,666 sentences are labeled in total and the Cohen's kappa is 0.44 between the two annotators, which is a reasonable and relatively high agreement considering the annotation complexity. Whenever there is any inconsistency, the third professional annotator will judge the annotation result in the confirmation phase to resolve the disagreement.

Table 1 shows a sample topic "Will artificial intelligence replace humans" and its labeled claims with their stances and evidences. The claims are labeled as "C_index" and the evidences are labeled as "E_index". For stances, "+1" represents the current claim supporting the topic, while "-1" represents the claim contesting the topic. A claim and an evidence form a claim-evidence pair (CEP) if the indices match with each other under a specific topic. An evidence can support multiple claims, such as Sent 3, as an evidence, it supports two claims, i.e. Sent 1 and Sent 2. Similarly, a claim can have different evidences, such as Sent 7, as a claim, it has three paired evidence sentences (i.e., Sent 4 - 6). As mentioned, one sentence can be considered as both the claim and the evidence. For instance, in Sent 8, there is a clear and concise statement "automation would not result in layoffs"

	Topics	Articles	Articles with claims	Claims	Support	Contest	Claims with evidences	Evidences	CEPs
Levy et al. (2014)	32	326	-	976	-	-	-	-	-
Rinott et al. (2015)	39	274	274	1,734	-	-	1,040	3,057	5,029*
Aharoni et al. (2014)	33 (12)	586	321 (104)	1,392	-	-	(350)	(1,291)	1,476*
Bar-Haim et al. (2017)	55	-	-	2,394	1,324	1,070	-	-	-
Ours	123	1,010	814	4,890	2,613	2,277	3,302	9,384	10,635

Table 2: Overall statistics comparison of the existing datasets and our dataset. Note that: (1) in Aharoni et al. (2014)'s dataset, the numbers in the parenthesis refer to the evidences labeling data; (2) the numbers with * are calculated by us since they are not shown in the original papers.

	Ours
Avg. length of all sentences	21.05
Avg. length of claims	23.44
Avg. length of evidences	25.09
Avg. % vocab shared in each CEP	20.14%
Avg. % vocab shared in each sent pair	8.73%

Table 3: Dataset statistics on argument lengths and vocabulary sharing.

contesting the given topic directly, which is considered as a claim. There is also a text segment at the beginning of the sentence showing the testimony from an organization (i.e., "Harvard Business Review") directly supporting this claim stated in the latter part of the sentence. Therefore, this sentence is labeled as an evidence as well. Last but not least, there are some claims without evidences found in the context in our dataset, such as Sent 9.

3.3 Dataset Analysis

We present the dataset statistics comparison with existing datasets in Table 2, and list the key differences below. First, as mentioned earlier, the existing datasets have their own focus on particular tasks, and none of them can support all the essential argument mining tasks related to the debate preparation process. Levy et al. (2014) only label data for claims, Rinott et al. (2015) only focus on detecting the evidences given the claims, Aharoni et al. (2014) only label a partial dataset for evidences, and Bar-Haim et al. (2017) only tackle the claim stance classification problem. In contrast, our dataset is fully annotated for all the key elements related to argument mining tasks, including claims, stances, evidences, and relations among them. Although combining Aharoni et al. (2014) and Bar-Haim et al. (2017)'s datasets can obtain a comprehensive dataset with 12 topics supporting all the subtasks, in terms of the dataset size, our dataset is significantly larger than it and the existing datasets. We explore 123 topics in total, which is more than twice of Bar-Haim et al. (2017)'s dataset. Accordingly, we obtain much more claims and evidences by human annotation on all sentences in the corpus, as compared to the previous datasets, which could add potential value to the argument mining community.

Table 3 shows more statistics of our dataset. In terms of the sentence lengths in our dataset, the average number of words in a sentence is around 21. The average length of sentences containing claims is generally longer, and evidences are even slightly longer than claims. However, since the length differences are subtle, it shows the challenges to distinguish the claims and evidences using the length differences among the sentences. We also calculate the average percentage of vocabulary shared between each claim-evidence sentence pair, which is 20.14%; while the same percentage between any two sentences from our corpus is only 8.73%. This shows that extracting CEP is a reasonable task setting as it has a higher percentage of vocabulary sharing than other sentence pairs, but it is also challenging as the absolute percentage is still low.

4 Tasks

In the debating system, our ultimate goal is to automate the whole debate preparation process as shown in Figure 1. With the introduced annotated dataset, we can tackle all core subtasks involved in the process at the same time. In this section, we first review the existing subtasks, and then propose two integrated argument mining tasks.

4.1 Existing Tasks

Task 1: Claim Extraction Similar to the CDCD task proposed by Levy et al. (2014), this task is defined as: given a specific debating topic and related articles, automatically extract the claims from the articles. Claim extraction is a primary argument mining task as the claim is a key argument component.

Task 2: Stance Classification As introduced by Bar-Haim et al. (2017), this task is defined as: given a topic and a set of claims extracted for it, determine for each claim whether it supports or contests the topic. As shown in Table 2, the number of claims from two stances is approximately balanced (i.e., 53.4% are support and 46.6% are contest), which also indicates the high quality of our dataset and annotations.

Task 3: Evidence Extraction In Rinott et al. (2015)'s work, this task is defined as: given a concrete topic, a relevant claim, and potentially relevant documents, the model is required to automatically pinpoint the evidences within these documents. In this paper, we only explore the evidence candidate sentences from the surrounding sentences of the claims, as long-distance sentences may not be content-relevant in most cases.

4.2 Integrated Tasks

In order to further automate the debating preparation process, exploring integrated tasks rather than individual subtasks is non-trivial. In this work, we introduce two integrated argument mining tasks as below to better study the subtasks together.

Task 4: Claim Extraction with Stance Classification (CESC) Since claims stand at a clear position towards a given topic, the sentences with clear stances should have a higher possibility to be the claims. Hence, identifying the stances of the claims is supposed to benefit the claim extraction task. By combining Task 1 and Task 2, we define the first integrated task as: given a specific topic and relevant articles, extract the claims from the articles and also identify the stance of the claims towards the topic.

Task 5: Claim-Evidence Pair Extraction (CEPE) Since evidences are clearly supporting the corresponding claims in an article, claims and evidences are mutually reinforcing each other in the context. Therefore, we hypothesize the claim extraction task and the evidence extraction task may benefit each other. By combining Task 1 and Task 3, we define the second integrated task as: given a specific topic and relevant articles, extract the claim-evidence pairs (CEPs) from the articles.

5 Approaches

To tackle the two integrated tasks, we first adopt a pipeline approach to pipe the corresponding subtasks together by using sentence-pair classification on each subtask. We also propose two end-to-end models for the two integrated tasks.

5.1 Sentence-pair Classification

We formulate Task 1, Task 2, and Task 3 as sentence-pair classification tasks. We train a sentence-pair classifier based on pre-trained models such as BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019). The sentence pairs are concatenated and fed into the pre-trained model to get the hidden state of the "[CLS]" token. Then, a linear classifier will predict the relation between the two sentences. Specifically, for Task 1, the topic and the article sentence are concatenated and fed into the model. If they belong to the same pair, the article sentence is considered as a claim, and vice versa. For Task 2, the model predicts the stance between a topic and a claim. Task 3 is similar to Task 1, where the model predicts if the given claim and the article sentence form a pair, i.e., if the sentence is an evidence of the claim. All these three tasks can be considered as binary classification tasks, and cross-entropy loss is used as the loss function.

Negative Sampling For Task 1 and Task 3, the binary labels are unbalanced as the number of claims/evidences is far smaller than the total number of sentences. To overcome this difficulty, we adopt negative sampling techniques (Mikolov et al., 2013). During the training of these two tasks, for each claim/evidence sentence, we randomly select a certain amount of non-claim/non-evidence sentences as negative samples. These negative samples together with all claims/evidences form a new training dataset for each task.

5.2 Multi-Label Model for CESC

Apart from the pipeline approach, we propose a multi-label model for the CESC task. Instead of handling the two subtasks separately, we concatenate the topic and article sentences to feed into a pre-trained model and define 3 output labels specifically for this task: support, contest, and no-relation. Support and contest refer to those claims with their corresponding stances to the topic, while no-relation stands for non-claims. Since the sentence pairs with no-relation labels are much more than those with support/contest, we also apply the negative sampling here for a more balanced training process.

	train	dev	test
# sents as claim candidates	55,544	7,057	7,065
# claims	3,871	492	527
# support claims	2,098	259	256
# contest claims	1,773	233	271
# claims with evidences	2,616	347	375
% claims with evidences	67.6%	70.3%	71.2%
# sents as evidences candidates	57,398	7,487	8,172
# evidences	7,278	909	1,108
Avg. # evidences per claim	2.78	2.62	2.95

Table 4: Dataset statistics split on train/dev/test sets.

5.3 Multi-Task Model for CEPE

Inspired from Cheng et al. (2021)'s work, we adopt a multi-task model (i.e., an attention-guided multi-cross encoding based model) for the CEPE task. Provided with a sequence of article sentences and the topic, we first concatenate the topic and individual sentences as the claim candidates, and use the sequence of article sentences as the evidence candidates. We reformulate the claim extraction and evidence extraction subtasks as sequence labeling problems. Then, the sequence of claim candidates and the sequence of evidence candidates go through the pre-trained models to obtain their sentence embeddings respectively. To predict whether two sentences form a claim-evidence pair, we adopt a table-filling approach by pairing each sentence in the claim candidates with each sentence in the evidence candidates to form a table. All three features (i.e., claim candidates, evidence candidates, table) update each other through the attention-guided multi-cross encoding layer as described in Cheng et al. (2021)'s work. Lastly, the two sequence features are used to predict their sequence labels, the table features are used for pair prediction between each claim and evidence. Compared to the pipeline approach, this multi-task model has stronger subtask coordination capability, as the shared information between the two subtasks is learned explicitly through the multi-cross encoder.

6 Experiments

6.1 Experimental Settings

We split our dataset randomly by a ratio of 8:1:1 for training, development, and testing. The dataset statistics are shown in Table 4. In the training set, since the number of claims (3,871) and the number of non-claims (51,673) are not balanced with a ratio of 1:13.3, we conduct experiments by selecting different numbers of negative samples and evaluate the effectiveness of the negative sampling strategy.

Models	Macro F_1	Micro F_1	Claim F_1
BERT-base-cased	72.08	**92.51**	48.08
RoBERTa-base	**72.36**	91.09	**50.35**

Table 5: Claim extraction performance.

It turns out that using 5 random negative samples for each claim performs the best. For each claim with evidences, 10 to 15 sentences before and after the claims are chosen to be the evidence candidates. The negative sampling strategy is also applied for the evidences candidates in the training set, where the ratio of positive samples (i.e., 7,278 evidences) to negative samples (i.e., 50,120 non-evidences) is 1:6.9. It turns out that using 1 negative sample for each evidence is the best.

We implement the sentence-pair classification model and the multi-label model for CESC with the aid of SimpleTransformers (Rajapakse, 2019). The multi-task model for CEPE is based on the implementation of the multi-task framework by Cheng et al. (2021). All models are run with V100 GPU. We train our models for 10 epochs. We experiment with two pre-trained models: BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019). Batch size is set as 128 for claim extraction and stance classification, and 16 for evidence extraction. We use 1 encoding layer for the multi-task model, and other parameters are the same as the previous work. [2]

For the claim and evidence extraction subtasks, besides Macro F_1 and Micro F_1, we also report the claim-class F_1 and the evidence-class F_1, respectively. For CESC, we additionally report the claim-class F_1 of different stances (i.e., support and contest). For the claim stance classification subtask, we report overall accuracy and F_1 for each class, as this task can be simply considered as a binary classification problem with balanced labels. For CEPE, we report precision, recall, and F_1.

6.2 Main Results on Existing Tasks

Claim Extraction Performance Table 5 shows the performance on Task 1. The classification model with pre-trained RoBERTa-base performs slightly better than with BERT-base-cased. Recall that we adopt the negative sampling strategy for these two models by randomly selecting 5 negative samples during the training phase. We also com-

[2] More details about hyper-parameter settings (i.e., batch sizes in the sentence-pair classification model, number of layers in the multi-task model), runtime and performance on the development set could be found in Appendix A.

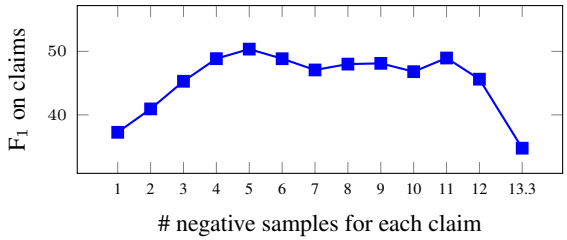

Figure 2: Effect of negative sampling for claim extraction with RoBERTa-base model.

Models	Acc.	Support F_1	Contest F_1
BERT-base-cased	73.43	73.08	73.78
RoBERTa-base	**81.21**	**81.21**	**81.21**

Table 6: Stance classification performance.

pare the performance of using different numbers of negative samples for each claim as shown in Figure 2. Generally speaking, the model performs better as the number of negative samples increases from 1 to 5, and starts to drop afterward. As the ratio is more balanced, i.e., from no sampling (1:13.3) to 5 negative samples, the F_1 score increases as expected. As the number of negative samples decreases further to 1, the ratio is even more balanced. However, it sacrifices the number of training data, which leads to worse performance.

Stance Classification Performance Table 6 shows the performance on Task 2. In both models, the F_1 scores on each stance are very close to each other, which is as expected because the two stances are balanced as shown in Table 4. Although the pretrained RoBERTa model outperforms the BERT model, there is still ample room for improvement as the accuracy of the RoBERTa model (81.21) is not relatively high for a binary classification task. One possible reason is that some claim sentences are too long to intuitively show the stances. For example, for the topic "Should vaccination be mandatory", a claim sentence "Young children are often at increased risk for illness and death related to infectious diseases, and vaccine delays may leave them vulnerable at ages with a high risk of contracting several vaccine-preventable diseases." is classified as "+1" according to the human evaluation, but is predicted as "-1" from the RoBERTa model.

Evidence Extraction Performance Table 7 shows the performance on Task 3. Again, the RoBERTa model performs better than the BERT model. For this task, we experiment with two settings: (1) given the topic and the claim (T+C), (2) only given the claim (C), to identify the evidences

Models	Macro F_1	Micro F_1	Evi. F_1
BERT-base-cased (T+C)	58.17	72.75	38.15
RoBERTa-base (T+C)	62.43	78.13	**40.89**
BERT-base-cased (C)	58.01	72.65	37.92
RoBERTa-base (C)	**63.37**	**80.29**	40.16

Table 7: Evidence extraction performance.

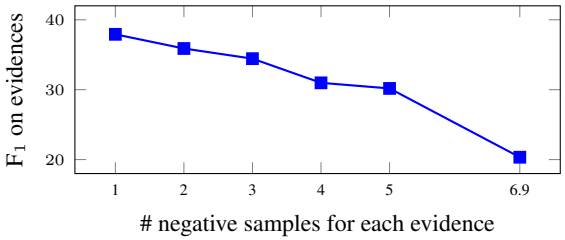

Figure 3: Effect of negative sampling for evidence extraction with BERT-base-cased (C).

Models	Macro F_1	Micro F_1	Support F_1	Contest F_1
Pipeline	55.95	88.56	33.39	40.10
Multi-label	**60.25**	**91.22**	**38.34**	**47.31**

Table 8: CESC task performance.

from the candidate sentences. For the (T+C) setting, we simply concatenate the topic and the claim as a sentence, and pair up with the evidence candidates to predict whether it is an evidence of the given claim under the specific topic. Comparing the results of these two settings, adding the topic sentences as inputs does not significantly improve the performance further, which suggests that claims have a closer relation with evidences, while the topic is not a decisive factor to evidence extraction. Here, 1 negative sample for each evidence sentence is randomly selected. The comparison of different numbers of negative samples is shown in Table **??**. Unlike the trend shown in the claim extraction task, the model achieves the best performance when the ratio is exactly balanced at 1:1.

6.3 Main Results on Integrated Tasks

For these two integrated tasks, we first use a pipeline method to pipe the best performing model on each corresponding subtask together, and then compare the overall performance with the proposed end-to-end models.

CESC Task Performance Table 8 shows the results of two approaches for the CESC task. For both two methods, we randomly select 5 negative samples for each positive sample (i.e., claim) during training. The pipeline model trains two subtasks independently and pipes them together to predict whether a sentence is a claim and its stance. Al-

Models	Precision	Recall	F_1
Pipeline	16.58	22.11	18.95
Traversal	24.06	**38.74**	29.69
Multi-task	**43.54**	30.57	**35.92**

Table 9: CEPE task performance.

though it achieves the best performance on each subtask, the overall performance is poorer than the multi-label model. It shows that identifying the stances of the claims can benefit the claim extraction subtask, and such a multi-label model is beneficial to the integrated CESC task.

CEPE Task Performance Table 9 shows the overall performance comparison among different approaches. Apart from the pipeline and the multi-task models as mentioned, we add another baseline model named "traversal". In this model, all possible pairs of "topic + claim candidate" and "evidence candidate" are concatenated and fed into the sentence-pair classification model. Both the traversal model and the multi-task model outperform the pipeline model in terms of the overall F_1 score, which implies the importance of handling these two subtasks together. The better performance of the multi-task model over the traversal model demonstrates the strong subtask coordination capability of the multi-task architecture.

6.4 Case Study

We present a few examples in Table 10 to compare the prediction results from the pipeline approach and the multi-task method for the CEPE task. Given the topic "should we ban human cloning", both models successfully identify the claim sentence. The first two sentences are not labeled as evidences supporting this claim based on the human annotation. The multi-task model labels these two sentences correctly, while the pipeline model predicts them as evidences by mistake. We notice that phrases of giving examples (e.g., "countries like") and numbers (e.g., "40 million", "year 2060") are very common elements in an evidence, which are the typical evidence types like demonstration with examples and digital evidences. We further explore the label predictions of these two sentences toward other claims and observe the pipeline approach classifies them as evidences as well. Without understanding the true meaning of the sentences, the pipeline approach only learns the common words and the structure. For the third evidence candidate, both models correctly predict this sentence and the extracted claim as a claim-evidence pair. However, the pipeline model fails to identify the last evidence candidate sentence as an evidence supporting the extracted claim. This is plausibly because the claim and the last evidence candidate sentence share few vocabularies. Although "genetic modification" is different from "cloning humans", they still share some similarities in terms of semantic comprehension in the context, thus the second sentence can also support the claim. Compared to the pipeline approach simply using the sentence-pair classification on the current sentences step by step, the multi-task model can learn a better sentence representation by utilizing the context information and coordinating two subtasks explicitly through the attention-guided multi-cross encoding layer, which finally leads to better performance. See Appendix B for more examples.

Topic: Should we ban human cloning	**Gold**	**PL**	**MT**
Claim: Cloning humans could reduce the impact of diseases in ways that vaccinations cannot.	C	C	C
This method could help countries like Japan who are struggling with low birth rates.		E	
The Japanese culture could see a reduction of up to 40 million people by the year 2060 without the introduction of cloning measures.		E	
Human cloning could help us to begin curing genetic diseases such as cystic fibrosis or thalassemia.	E	E	E
Genetic modification could also help us deal with complicated maladies such as heart disease or schizophrenia.	E		E

Table 10: Examples of model predictions for CEPE task. PL stands for the pipeline, and MT stands for the multi-task. We select four sentences from the evidence candidates to demonstrate the prediction results here.

7 Conclusions

In this paper, we introduce a comprehensive and large dataset for argument mining to facilitate the study of multiple tasks involved in the debating system. Apart from the existing primary argument mining tasks for debating, we propose two integrated tasks to work towards the debate automation, namely CESC and CEPE. We experiment with a pipeline method and an end-to-end approach for both integrated tasks. Experimental results and analysis are presented as baselines for future research, and demonstrate the value of our proposed tasks and dataset. In the future, we will continue studying the relations among the argument mining subtasks and also explore more useful research tasks in the debating system.

References

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 argumentation mining*.

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

Roy Bar-Haim, Liat Ein Dor, Matan Orbach, Elad Venezian, and Noam Slonim. 2021. Advances in debating technologies: Building ai that can debate humans. In *Proceedings of ACL-IJCNLP*.

Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. 2019. Argument invention from first principles. In *Proceedings of ACL*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathleen McKeown, and Alyssa Hwang. 2019. Ampersand: Argument mining for persuasive online discussions. In *Proceedings of EMNLP-IJCNLP*.

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-IJCLP*.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of EMNLP*.

Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. Mining ethos in political debate. In *Proceedings of International Conference on Computational Models of Argument*.

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

Matthias Grabmair, Kevin D Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R Walker. 2015. Introducing luima: an experiment in legal conceptual retrieval of vaccine injury decisions using a uima type system and tools. In *Proceedings of international conference on artificial intelligence and law*.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of AAAI*.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of ACL*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

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

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

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *Proceedings of AAAI*.

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. *arXiv e-prints*.

Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of AAAI*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Shachar Mirkin, Guy Moshkowich, Matan Orbach, Lili Kotlerman, Yoav Kantor, Tamar Lavee, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2018. Listening comprehension over argumentative content. In *Proceedings of EMNLP*.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*.

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

T. C. Rajapakse. 2019. Simple transformers. https://github.com/ThilinaRajapakse/simpletransformers.

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*.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature*.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of EMNLP*.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *CL*.

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*.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment-new datasets and methods. In *Proceedings of EMNLP-IJCNLP*.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of EACL*.

A More Experimental Details

A.1 Hyper-parameters

We manually tune the hyper-parameters in our models. Table 11 shows the results on claim extraction task with different batch sizes from 8 to 128. Here, we use the pre-trained RoBERTa-base model. 5 negative samples are randomly chosen for each claim during training. When the batch size is 128, the model achieves the best performance.

Models	Batch size	Macro F_1	Micro F_1	Claim F_1
RoBERTa-base	8	68.44	91.24	41.65
RoBERTa-base	16	70.92	**92.17**	45.94
RoBERTa-base	32	71.59	90.97	48.82
RoBERTa-base	64	72.11	91.89	48.85
RoBERTa-base	128	**72.36**	91.09	**50.35**

Table 11: Claim extraction performance with different batch sizes.

Table 12 shows the results of using different batch sizes ranging from 8 to 128 for the stance classification task. Each model (i.e., BERT and RoBERTa) achieves the best performance when the batch size is 128.

Models	Batch size	Accuracy
BERT-base-cased	8	69.45
BERT-base-cased	16	68.50
BERT-base-cased	32	76.28
BERT-base-cased	64	65.09
BERT-base-cased	128	73.43
RoBERTa-base	8	70.97
RoBERTa-base	16	75.71
RoBERTa-base	32	78.37
RoBERTa-base	64	79.32
RoBERTa-base	128	**81.21**

Table 12: Results of different batch sizes for stance classification task.

Table 13 shows the effect of using different numbers of layers in the multi-task model. More model details regarding each layer could be found in (Cheng et al., 2021)'s work. The multi-task model achieves the best F_1 score when the number of layers is 1.

Layers	Precision	Recall	F_1
1	43.54	**30.57**	**35.92**
2	44.79	26.60	33.38
3	**45.65**	22.64	30.27

Table 13: Effect of different numbers of layers used in the multi-task model.

A.2 Runtime and Validation Performance

In Table 14, we present the running time and the results on the development set of the multi-task model on the CEPE task. As the number of layers increases, it requires a longer training time.

Layers	RT (min)	Dev P.	Dev R.	Dev F_1
1	15	33.31	20.66	25.50
2	22	39.81	18.68	25.43
3	29	40.59	18.36	25.28

Table 14: Runtime (RT) per epoch (minutes), the precision (P.), recall (R.) and F_1 on the development set of the multi-task model with differnt numbers of layers.

B More Case Study

Table 15 shows more example predictions generated by the pipeline approach and the multi-task model for the CEPE task. In these examples, the multi-task model identifies most of the claim-evidence pairs while the pipeline method fails to do so. For the second topic which is shown earlier in Section 3.2, the pipeline model fails to detect the claim sentence nor the evidence sentence.

Topic: Should we fight for the Olympics	**Gold**	**PL**	**MT**
Claim: These often impose costs for years to come.	C	C	C
Sydney's Olympic stadium costs the city $30 million a year to maintain.	E		E
Beijing's famous "Bird's Nest" stadium cost $460 million to build and requires $10 million a year to maintain, and sits mostly unused.	E		E
Topic: Will artificial intelligence replace humans			
Claim: Any job that involves repetitive tasks is at risk of being replaced.	C		C
In 2017, Gartner predicted 500,000 jobs would be created because of AI, but also predicted that up to 900,000 jobs could be lost because of it.	E		E
Topic: Should we implement the network real-name system			
Claim: Real-name policy blurs the boundaries between personal information and personal privacy.	C	C	C
Due to the vague boundaries between privacy and personal information, today people are willing to distinguish this boundary between online behavior and offline ID.	E		
For example, as an Internet user, my words and deeds on the Internet, personal information published, such as political positions, belong to my personal information.	E		E
But once it matches my true identity, it is personal privacy.	E		E

Table 15: More example predictions of the CEPE task.