Sequential Multi-Task Learning for Biomedical Argument Mining

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

Abstract—Biomedical argument mining (BAM) is a challenging task that involves automatically identifying the structure of argumentation in the biomedical text. BAM plays an important role in extracting exact information from a large-scale medical text data, and is of great significance to the development of Evidence-based decision making. However, the existing multi-task model addresses BAM through a conventional multi-task learning framework, ignoring the sequential dependency between the Argument Component Classification (ACC) task and the Relation Identification (RI) task. In addition, unlike the other kinds of text, it is necessary to combine multiple evidence to obtain the final claim in the biomedical text. The existing models focus solely on the information of argument component pair itself on the RI task, ignoring the semantic information of the other related argument component pairs. In order to solve the above issues, in this paper, we propose a Sequential Multi-Task (SeqMT) learning model for BAM. In specific, to model the sequential dependency between the ACC task and the RI task, the representation of the input and output in the ACC task is transferred into the RI task. Furthermore, we structure the argument component pairs of the same local neighborhood into a pair graph network. Then, we employ a Graph Convolutional Network to model collection-level context information for an argument component pair. The proposed method is evaluated on a benchmark dataset and the experimental results show that our method outperforms the state-of-the-art methods.

Index Terms—Computer Society, IEEE, IEEEtran, journal, LaTeX, paper, template.

1 Introduction

In the biomedical domain, there is an increasing interest in Evidence-based decision making, which can help clinicians make the best decision for the medical case under evaluation. Most work [1], [2], [3] focus on the reasoning stage in Evidence-based decision making, but the mining stage receive little attention. The target of the mining stage is to extract, from a large amount of medical data for different diseases and treatments, the useful information for clinicians and further to present this information in a structured way [4]. Biomedical Argument Mining (BAM) aims at identifying the argumentative structures in biomedical text, including Argument Components (AC) and relations among them. Because of its aptness to automatically identify important information in biomedical text and present them in a structured manner, which is consistent with the target of the mining stage, BAM is of great significance for supporting clinicians to make accurate medical decisions accurately.

BAM is the application of Argument Mining (AM) in the biomedical domain that consist in: (1) *argument component identification* (ACI), which involves separating the arguments components from non-argumentative text; (2) *argument component classification* (ACC), which involves identifying the argument components with different types (i.e., *majorclaim, claim, and evidence*); (3) *relation identification* (RI), which involves recognizing the argumentative relations (i.e., *support*, *attack* and *none*) within a single argument component pair (ACp) such as *evidence-claim* [5], [6]. As the Fig. 1 shows, compared to text in other genres (e.g., *student essay*), biomedical text generally contains numerical indicators (e.g, $P = 0.049$), statistical findings and domain-specific terminologies (e.g., *GJJ*), which bring difficulties for models to understand the complex semantic informationi within the biomedical text [7], [8]. In addition, a common situation in biomedical text is that the argumentative relation within an ACp should be determined via combinating of multiple consecutive evidence. For example, as is shown in Fig. 1, we construct two ACps (i.e., ACp1: *evidence1* $\rightarrow$ *claim*, ACp2: *evidence2* $\rightarrow$ *claim*). *evidence1* talks about initial costs where 'GJJ' is higher and *evidence2* talks about follow-up costs where 'GJJ' and 'stent placement' are the same. When considering ACp1 and ACp2 separately, it is difficult to recognize their relation types, due to the claim talks about total costs. Combining the information of ACp1 and ACp2, the relation types of ACp1 and ACp2 can be easily recognized as *support*.

Two previous studies have been conducted to address BAM. Mayer et al. [4] released an AbstRCTs dataset of 659 RCT abstracts about five diseases, including neoplasm, glaucoma, hypertension, hepatitis and diabetes. They employed a complete pipeline to address three subtasks of BAM, where boundaries and component types of AC are identified through token-level tagging, then relation types of ACp can be recognized with AC provided from previous step. Furthermore, to avoid error propagation caused by the pipeline, Galassi et al. [9] employed a conventional sentence-level multi-task learning framework to jointly address the ACC task, RI task, and link prediction task, which contains a shared encoder, four separate classifiers with the golden ACs provided in advance.

Although proved to be effective, the existing methods have two shortcomings: (1) The existing multi-task learning models in BAM handle different tasks with multiple

- *M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: see http://www.michaelshell.org/contact.html*
- *J. Doe and J. Doe are with Anonymous University.*

Manuscript received April 19, 2005; revised August 26, 2015.

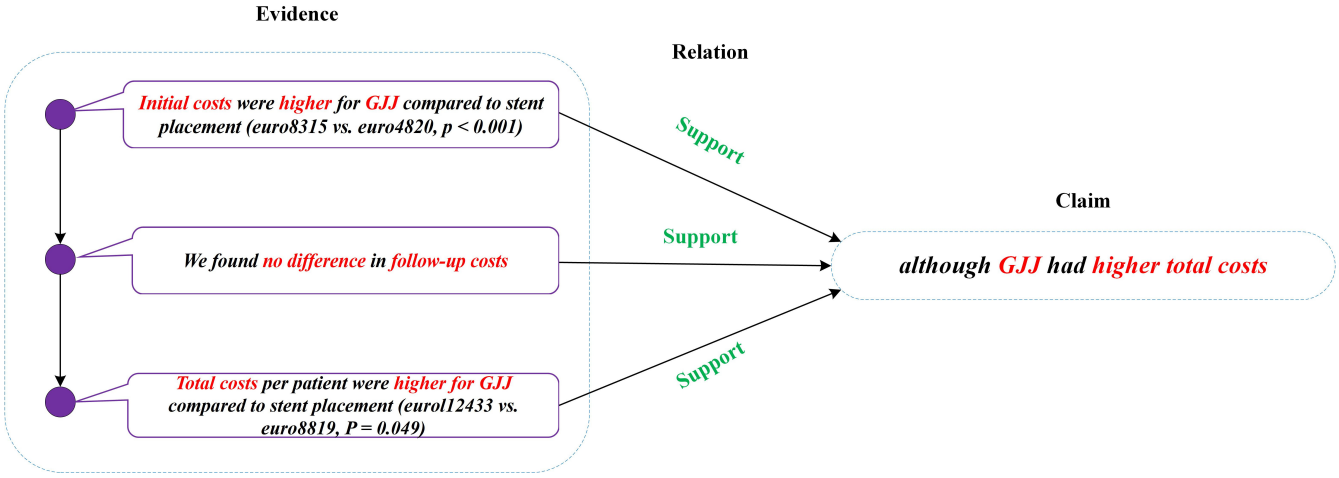


Fig. 1. An example of biomedical argument mining. Important words are marked in red

independent classifiers, however, they largely overlooked the implied sequential dependency between the ACC task and the RI task. Actually, most of the support or attack relation occurs between *evidence-claim* pair rather than *evidence-evidence* pair, which indicates that the type information of ACs learned in the ACC task could be beneficial for accurately recognizing the relation types of ACps. (2) Approaches to relation identification task have largely focused on the information of single ACp, loosing out potentially valuable context from the borader collection of ACps within a RCT abstract. Generally, the relation types of ACp need to be verified by combining the information from surrounding relevant ACps for biomedical text (e.g., *evidence1-claim* and *evidence2-claim* as shown in Fig. 1), which shows the significance of incorporating the collection-level contextual informatioon from related ACps within the same text.

Given the above considerations, we propose a novel **Seq**uential **M**ulti-**T**ask learning model (SeqMT) for BAM. In specific, to model the sequential dependency between the ACC task and the RI task, the representation of the input and output in the ACC task is incorporated into the RI task through the information transfer module. Moreover, to capture collection-level contextual information, we extract a set of the local neighborhoods from a single RCT abstract, where each local neighborhood is defined as the collection of ACps with the second AC unchanged. Then a graph is constructed for each local neighborhood, where nodes represent ACp and edges represent different possible relationship between the corresponding two ACps in a graph. Finally, Graph Convolutional Network (GCN) [10] is employed to learn the representation of each ACp via information propagation from the related ACps in the same local neighborhood.

The contributions of this paper can be summarized as follows:

- A novel **Seq**uential **M**ulti-**T**ask learning model (SeqMT) is proposed, which is, to our best knowledge, the first attempt of exploring the effect of sequential dependency between the ACC task and the RI task for BAM.
- We demonstrate that considering collection-level contextual information from the related ACps within a single RCT abstract can improve the performance of biomedical argument mining.
- Our model is evaluated on a benchmark BAM dataset for the ACC task and the RI task. The experimental results show that our model outperforms the state-of-the-art methods.

The rest of this paper is organized as follows: Section 2 introduces the related works of argument mining and graph conventional network. Then we describe the details of our proposed model in Section 3. The experimental setting and results are presented in Section 4. In Section 5, We conclude our work and make some ideas for work in the future.

2 Related Work

2.1 Argument Mining

Recent researches for argument mining mainly focues on the subtasks of AM, such as argument component identification [11], [12], [13], [14] and argument component classification [15], [16]. With a strongly internal relatedness of these subtasks, a growing number of works employed multi-task or joint learning framework to address these subtasks. Peldszus et al. [17] utilized a evidence graph to combine different subtasks for joint training, with the assumption that the ACI task has been finished. Stab et al. [6], for the first time, employed a sequence tagging model to identify the boundaries of AC and applied a joint Integer Linear Programming (ILP) model to detect argumentation structures. Potash et al [18] utilized pointer network to identify the types of AC and detect whether the link exists between ACs with an assumption that all ACs have already been identified. Eger et al. [19] proposed an end-to-end sequence tagging model ,which firstly integrate the BIO label of ACI task, the types of ACs, and the relation labels between ACs into the same label space, and utilized Bi-directional LSTM (BiLSTM) to predict the final results of three subtasks. In [20], a structured learning framework based on factor graphs was employed to identify the elementary unit type and argumentative relation.

Biomedical argument mining is a newly emerged research field and developed rapidly in recent years due to the potential application for medical diagnosis. Mayer et al. [21] annotated ACs within RCT abstracts and utilized SubSet Tree Kernel (SSTK) to classify the types of AC by taking the Bag-Of-Words (BOW) of biomedical text as input. In a similar vein, Mayer et al. [4] created the first completely biomedical argument mining dataset to deal with three subtasks, where component classification task and boundary detection task were intergrated as one problem with BIO scheme, then the relation types of all pairs of ACs were classified as *support*, *attack* and *none*. Furthermore, The utilization of various contextualize word embedding was explored to address biomedical argument mining task, such as BERT [22], BioBERT [23], SciBERT [24], RoBERTa [25] et al., which are pre-trained on the large corpora in the different domains. Galassi et al. [9] employed multi-task framework with attentive residual network to address the ACC task, the RI task, and link prediction task of BAM, based on an assumption that ACs had been detected.

Different from existing previous research, our work handle biomedical argument mining with consideration of sequential dependency between the ACC task and the RI task. Moreover, we construct a set of graphs for an RCT abstract to model the collection-level contextual information propagation from the related ACps in the same neighborhood.

2.2 Graph Convolutional Network

Kipf et al. [10] fristly proposed the GCN for node classification, which showed state-of-the-art results on many benchmark graph datasets. With the flexible operation on the graph structure, GCN has been widely applied to various Natural Language Processing (NLP) tasks recently. Yao et al. [26] employed GCN to accomplish text classification, where the text graph was construted by word co-occurrence and document-word relations. Sun et al. [27] modeled the contextual information and dependency information between opinion words and aspect words by GCN. Chen et al. [28] proposed a pairGCN for emotion-cause pair extraction, which models the dependency information among related emotion-cause pairs.

In this paper, GCN is applied to aggregate the relevant information from related ACps in the same neighborhood to model the collection-level contextual information.

3 Proposed Model

In this section, we present an overview of the architecture of SeqMT. As shown in Fig 2, the model consists of four main modules: (I) Representation Module, which learns the word representation of AC and the relation representation of ACp; (II) Argument Component Classification (ACC) Module, which assigns the argument component with different types of the label through a set of BiLSTM networks; (III) Information Transfer (IT) Module, which generates the representations of ACps and further transferred to the relation identification module; (IV) Relation Identification (RI) Module, which employs GCN to model the collection-level contextual information from the related ACps in the same local neighborhood.

3.1 Task Definition

Given the RCT abstract, the objective of our work is to identify the component types of AC and relation types of ACp. Following Galassi et al. [9], we define an RCT abstract $D = \{c_1, c_2, ..., c_L\}$ as a sequence of L ACs. The ACs in D are formed into a set of ACps P by Cartesian product:

$$P = \left\{c^p_{1,1}, ..., c^p_{i,j}, ..., c^p_{L,L}\right\} \tag{1}$$

$$c^p_{i,j} = (c_i, c_j) \tag{2}$$

where c_i is i-th AC of D and c_j is the j-th AC of D, and there are totally $L * L$ ACps in P. The ACC task aims to identify c_i as *claim* or *evidence* and the RI task aims to identify the relation type of $c^p_{i,j}$ as *support*, *attack*, or *none*.

3.2 Representation Module

This section explains the details of how SeqMT extracts the word representations of AC and semantic representations of ACp through an encoder.

Firstly, we adopt the BioBERT [23] as encoder, which is pre-trained on large-scale biomedical corpora PubMed and performs well in biomedical NLP tasks. Given $c^p_{i,j} = (c_i, c_j)$, where $c_i = \{w_1, w_2, ..., w_{l_i}\}$ and $c_j = \{w_1, w_2, ..., w_{l_j}\}$, l_i and l_j are the length of c_i and c_j. In order to be encoded by BioBERT, we add [CLS] token at the beginning of word sequence, then use [SEP] token to distinguish different ACs:

$$\boldsymbol{W}_{i,j} = [\mathrm{CLS}]; c_i; [\mathrm{SEP}]; c_j; [\mathrm{SEP}] \tag{3}$$

where $\boldsymbol{W}_{i,j}$ is the input of the encoder. The input words in both c_i and c_j will learn the inforamtion of the relationship between them using multi-layer transformer encoders [29]. Subsequently, the relation representation of $c^p_{i,j}$ and word representations of $c^p_{i,j}$ can be computed by:

$$\boldsymbol{r}_{i,j}, \boldsymbol{I}_{word} = \mathrm{BioBERT}(\boldsymbol{W}_{i,j}) \tag{4}$$

where the hidden state of [CLS] token $\boldsymbol{r}_{i,j} \in \mathbb{R}^d$ represents the relation representation of $c^p_{i,j}$, $\boldsymbol{I}_{word} \in \mathbb{R}^{l_p \times d}$ represents word representation in $c^p_{i,j}$, where d is the dimension of representation space and $l_p = l_i + l_j$ is the length of $c^p_{i,j}$. $\boldsymbol{I}_{word}$ can also be described by Formula 5 specifically:

$$\boldsymbol{I}_{word} = \left\{e_1, e_2, ..., e_{l_p}\right\} \tag{5}$$

where e_t is the word representation of t-th words in $c^p_{i,j}$.

Furthermore, to obtain the word representation in c_i, we first design a special mask vector MASK, which is presented as follows :

$$\mathrm{MASK} = [1, ..., 1, 0, ..., 0] \tag{6}$$

$$\boldsymbol{E}_{ij} = \boldsymbol{I}_{word} \cdot \mathrm{MASK} \tag{7}$$

where the length of MASK is equal to l_p, the number of "1" in MASK is equal to l_i, and "$\cdot$" means dot product. Then the average operation is applied to the L ACps with the same AC c_i to obtain the final word representation:

$$\boldsymbol{E}_i = \frac{\sum_{j=1}^{L} \boldsymbol{E}_{ij}}{L} = \{e_1, e_2, ..., e_{li}\} \tag{8}$$

where $\boldsymbol{E}_i \in \mathbb{R}^{l_i \times d}$ is final representation of each word in c_i.

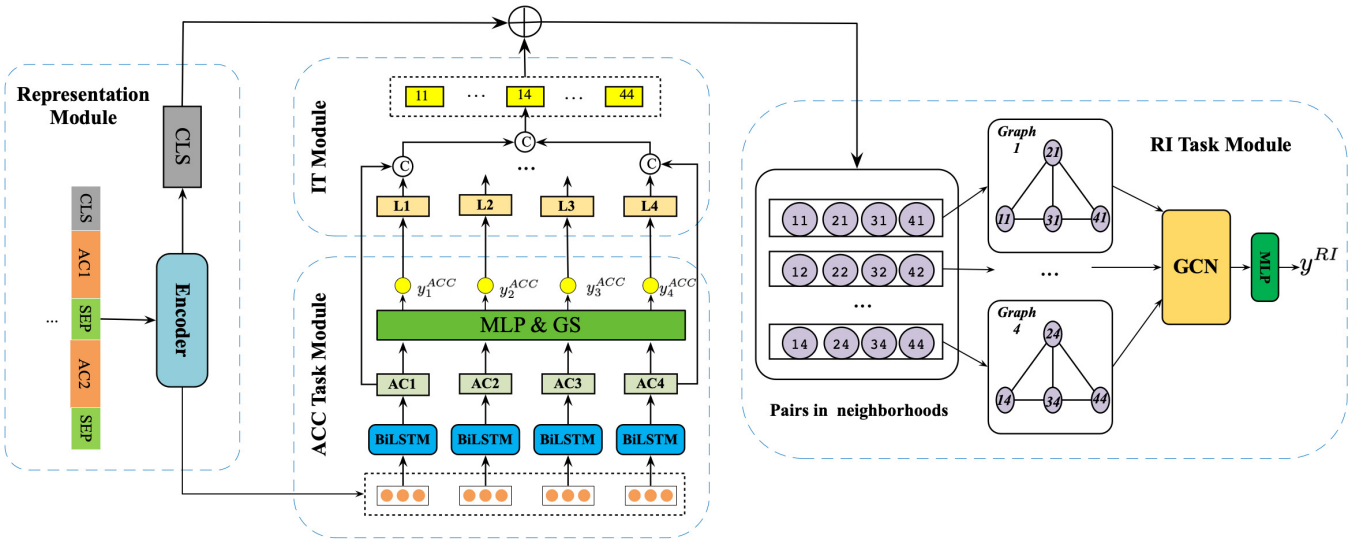


Fig. 2. The architecture of SeqMT

3.3 Argument Component Classification Module

This section describes how to get the semantic representation and type of AC.

As an effective text sequence encoder, BiLSTM can capture the forward and backward dependencies among words in the text sequence [30]. Therefore, we employ a set of word-level BiLSTM networks, each of which corresponds to one AC, to generate the semantic representation of each AC by taking the word representation $\boldsymbol{E}_i$ as input. The hidden state of the k-th word in the i-th AC is $h_{i,k}$ that can be obtained by:

$$h_{i,k} = [\overrightarrow{LSTM}(e_{i,k}); \overleftarrow{LSTM}(e_{i,k})] \tag{9}$$

where $h_{i,k} \in \mathbb{R}^{2d}$, which is the concatenation of the forward hidden state and backward hidden state. Subsequently, the semantic representation of AC can be obtained by summing all of the hidden states in c_i:

$$\boldsymbol{H}_i = \sum_{k=1}^{li} h_{i,k} \tag{10}$$

Where $\boldsymbol{H}_i \in \mathbb{R}^{2d}$ is the final semantic representation of the i-th AC, which is utilized as the features for the ACC task. Then, a multi-layer perceptron (MLP) with one hidden layer is used to generate the component type label prediction probability distribution of ACs:

$$\hat{y}_i = \mathrm{MLP}(\boldsymbol{H}_i) \tag{11}$$

where $\hat{y}_i \in \mathbb{R}^C$, C is the number of component types.

3.4 Information Transfer Module

Based on the illustration in Section 1, we observe that the support or attack relation commonly occurs between the evidence and the claim, which indicates the benefit of component type information of AC for learning relation type information of ACp accurately. Therefore, how to incorporate the label information of AC is a key component in our model. This section elaborates on the process of how to obtain the label information of AC with a sampling strategy and generate the pair-wise representation of ACp transferred to the RI task.

Sampling Strategy. To model the label information of an AC, we need to obtain the label of AC, which is a one-hot vector. During the training stage, we can exploit the gold labels of ACs. However, we can only use predicted labels of ACs during the testing stage, which leads to the train-test discrepancy. Meanwhile, using the argmax method to obtain the label value from probability distribution of the label will cause the gradient to be unable to return during the training stage. In order to solve the above problem, we apply Gumbel-Softmax (GS) [31] as a effective sampling strategy, which samples a label from $\hat{y}_i$ and output the one-hot vector of label [32]. In specific, GS utilizes a re-parameter trick to avoid the problem that the gradient cannot be returned. The sampling process of GS is as follows:

$$\tilde{y}_i = softmax((\hat{y}_i + g)/\tau) \tag{12}$$

where g samples from Gumbel$(0, 1)$ distribution, τ denotes the temperature, which controls the final result generation trend, where $\tilde{y}_i$ will be closer to a one-hot vector when $\tau \to 0$. We replace gold label with $\tilde{y}_i$ during the training stage.

Pair-wise Representation. To incorporate the label information into the semantic representation learning of ACp, we firstly encode $\tilde{y}_i$ to label embedding le_i for AC c_i:

$$\boldsymbol{le_i} = W_{le}\tilde{y}_i \tag{13}$$

where $W_{le} \in \mathbb{R}^{d \times C}$ is the learned parameter matrix. Furthermore, the new representation $\boldsymbol{R}_i^l$ for c_i with its label information can be obtained by concatenating the semantic representation of AC $\boldsymbol{H}_i$ and label embedding $\boldsymbol{le}_i$:

$$\boldsymbol{R}_i^l = W_l[\boldsymbol{le}_i; W_r\boldsymbol{H}_i] \tag{14}$$

where $W_l \in \mathbb{R}^{2d \times d}$ and $W_r \in \mathbb{R}^{2d \times d}$ are parameter matrix, $[;]$ is the concatenation operation. Moreover, we employ concatenation operation to obtain the pair-wise representation $\boldsymbol{R}_{i,j}^{acc}$ for $c_{i,j}^p$ specific to ACC task:

$$\boldsymbol{R}_{i,j}^{acc} = W_p[\boldsymbol{R}_i^l; \boldsymbol{R}_j^l] + b_p \tag{15}$$

where $W_p \in \mathbb{R}^{2d\times d}$ and $b_p \in \mathbb{R}^d$ are the learned parameter matrix.

3.5 Relation Identification Module

How to incorporate the related contextual information is important to recognize the relation type of ACp within biomedical text as we mentioned in Section 1. Motivated by the remarkable results obtained by graph-based architectures in NLP tasks, we first define the local neighborhood as a graph with the subset of ACps $N_j = \{c^p_{1,j}, c^p_{2,j}, ..., c^p_{L,j}\}$ to model the collection-level contextual information from related ACps, where nodes denote the a set of ACps with the same AC c_j. If an RCT abstract has L ACs, there are L local neighborhoods in total. Then, the GCN is employed to enhance the representation of ACp by information propagation from related ACps within the same local neighborhood. The construction process of a pair graph is as follows:

Nodes. Given a set of ACps N_j belong to the same local neighborhood, each ACp is considered as a node. $\boldsymbol{V}_j \in \mathbb{R}^{L\times d}$ represents the j-th local neighborhood, the initial representation of i-th node $c^p_{i,j}$ in j-th local neighborhood is computed by summing up the pair-wise representation and the relation representation of ACp:

$$\boldsymbol{v}_{i,j} = \boldsymbol{R}^{acc}_{i,j} + \boldsymbol{r}_{i,j} \tag{16}$$

Edges. To capture the information of consecutive AC that support or attack the same AC normally, we take the current ACp as the center ACp and only conder other ACps within windows size of 2 (2.18 ACs have relations with the same AC on average, as shown in Tab. 1). For $c^p_{i,j}$, the following ACps should be paid more attention, and have edges with $c^p_{i,j}$:

$$c^p_{[i-2,i+2],j} = \left\{c^p_{i-2,j}, c^p_{i-1,j}, c^p_{i+1,j}, c^p_{i+2,j}\right\} \tag{17}$$

Then, we denote $G_j = (V_j, E_j)$ as the j-th graph of local neighborhood, where E_j represents the edges, the weight of edge between two nodes is set to 1.

Furthermore, we can easily construct the adjacency matrix A_j and its degree matrix D after the construction of graph, where $D_{ij} = \sum_j A_{ij}$. Then a GCN with one layer is employed to extract collection-level contextual information:

$$\boldsymbol{Q}^c_j = D^{-\frac{1}{2}} A_j D^{-\frac{1}{2}} \boldsymbol{V}_j W_{gcn} \tag{18}$$

where $\boldsymbol{V}_j = [\boldsymbol{v}_{1,j}, \boldsymbol{v}_{2,j}, ..., \boldsymbol{v}_{L,j}] \in \mathbb{R}^{L\times d}$ is a feature matrix of G_j, $W_{gcn} \in \mathbb{R}^{d\times d}$ is the learned weight matrix. $\boldsymbol{Q}^c_j = [\boldsymbol{q}_{1,j}, \boldsymbol{q}_{2,j}, ..., \boldsymbol{q}_{L,j}]$ is the representation by incorporating the collection-level contextual information from related ACps via information propagation in GCN. Then, a residual operation is employed to update the representation of $c^p_{i,j}$ further:

$$\boldsymbol{r}^{final}_{i,j} = \boldsymbol{q}^c_{i,j} + \boldsymbol{v}_{i,j} \tag{19}$$

Finally, A MLP with one hidden layer is used to generate the label probability distribution of RI task.

$$\bar{y}_{i,j} = \mathrm{MLP}(relu(\boldsymbol{r}^{final}_{i,j})) \tag{20}$$

where $relu$ is the activation function.

Algorithm 1 SeqMT Algorithm

Input: A set of ACs, a set of ACps P(constructed based on the set of ACs)
Initialize: Randomly initialize the model parameters.
Output: argumentative types y_i of AC and relation types $y_{i,j}$ of ACp.

while not converge **do**
  for ACp $c^p_{i,j}$ in P **do**
    Learn the relation representation $\boldsymbol{r}_{i,j}$ and word representation $\boldsymbol{I}_{word}$ using Eq.(4).
  end for
  Obtain the word representations $\boldsymbol{E}_i$ of each word in c_i by Eq.(8).
  Learn the semantic representation $\boldsymbol{H}_i$ of AC c_i by Eq.(9),(10).
  Make label prediction distribution of ACC task $\hat{y}_i$ using Eq.(11).
  Obtain the label information of AC c_i by Eq.(12).
  Compute the pair-wise representation $\boldsymbol{R}^{acc}_{i,j}$ for $c^c_{i,j}$ by Eq.(13),(14),(15).
  for ACp $c^p_{i,j}$ in P **do**
    Obtain the initial node representation by Eq.(16).
    Determine the node set N_j, the edge set E_j, construct the pair graph G_j.
    Calculate adjacency matrix A_j and degree matrix D_j
    Incorporate the collection-level contextual information using GCN.
    Update the representation of $c^p_{i,j}$ by Eq.(19).
    Make label prediction distribution of RI task $y_{i,j}$ by Eq.(20).
  end for
  Calculate the combined loss $\mathcal{L}$ by Eq.(21),(22),(23).
  Update parameters.
end while

3.6 Training Objective

The ACC and RI task are trained jointly, with a multi-task framework. Therefore, a joint loss function is considered as the training objective.

We use cross entropy loss function to compute the loss of ACC task $\mathcal{L}_{acc}$ and RI task $\mathcal{L}_{ri}$. $\mathcal{L}_{acc}$ and $\mathcal{L}_{ri}$ can be computed as follow:

$$\mathcal{L}_{acc} = \frac{1}{N}\sum_i -y.log(p) - (1-y).log(1-p) \tag{21}$$

$$\mathcal{L}_{ri} = \frac{1}{N}\sum_i \sum_{i=k}^{C} -y_k log(p_k) \tag{22}$$

where $C = 3$ is the number of categories of RI task. Finally, the combined loss $\mathcal{L}$ function is defined as follows:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{acc} + \lambda_2 \cdot \mathcal{L}_{ri} \tag{23}$$

where λ_1 and λ_2 are hyper-parameters. Algorithm 1 describes the overall of training process of our model SeqMT.

TABLE 1
Statistics of the average number of ACs that have relation with the same AC. R_{num} means the number of ACPs, whose relation types are not *none*. T_{num} means the number of different tail ACs in ACPs, whose relation types are not *none*

	T_{num}	R_{num}	Average
neo-train	642	1427	2.22
neo-dev	92	210	2.28
neo-test	209	424	2.03
gla-test	166	367	2.21
mix-test	158	329	2.08
total	1267	2757	2.18

TABLE 2
Statistics of AbstRCTs dataset

Dataset	Evi	Cla	MajCl	Sup	Att
neo-train	1537	666	64	1213	214
neo-dev	218	99	9	186	24
neo-test	438	228	20	364	60
gla-test	404	183	7	334	33
mix-test	388	182	30	305	24

4 EXPERIMENTS

In this section, we describe the dataset, baselines, evaluation metric and experiment setting details before analyzing the experimental results.

4.1 Dataset

Since BAM is a new task, there is only one complete dataset so far. Mayer et al. [4] created an AbstRCTs dataset, which extends their previous dataset [21]. The annotated data come from the abstracts of papers regarding RCT for different diseases, including neoplasm, glaucoma, hypertension, hepatitis and diabetes. The number of abstracts about neoplasm is 500 and 350 of these abstracts are selected as train set, 100 of these abstracts are selected as test set, 50 of these abstracts are selected as validate set. There are other test sets which are glaucoma and mixed, where glaucoma test set consists of 100 RCT abstracts of glaucoma and mixed test set consists of 20 RCT abstracts for each disease. Table 2 presents statistics of each subsets. Due to the small number of *majorclaim*, *majorclaim* is treated as a category of *claim*. Meanwhile, we can find that distribution of relation types is imbalanced, the number of *support* is far beyond *attack*, which leads to difficulty of learning the relation of *attack*.

In our work, we continue to insist on the original division settings of AbsRCTs and test our model on different test sets, which are neo-test, gla-test and mix-test.

4.2 Baselines

The SeqMT could be directly compared with [9], [33], but it is difficult to compare our model SeqMT directly with [4], which is because that SeqMT perform ACC task and RI task versus complete pipeline approach. Specifically, there are the following two problems to be solved: (1) we treat the ACC task as a sentence-level task, which makes our results unable to compare with the approaches that perform it token-wise. Therefore, we follow the strategy of [9], where each classified AC is split into tokens that share the same label with the AC. We can obtain the results of token-wise classification through tokenization method, which is acceptably comparable. (2) For RI task, we consider the golden ACs that are obtained in advance in SeqMT, which make the our results of RI task incomparable with [4]. In [4], errors of ACI task and ACC task will introduced to RI task, which means that non-argumentative components are treated as ACs and some ACs are lost. Galassi et al [9] make an analysis on this problem in detail, and concluded that this problem cannot be solved. Therefore, we choose the results of RI task in [4] to make a qualitative comparison.

After the above analysis, we select the following baselines including single-task approach, pipeline approach and multi-task approach to compare our model SeqMT, to evaluate the effectiveness of the our proposed model.

4.2.1 Single-task

• **BiLSTM** employs BiLSTM network with different transformer encoders (BERT, BioBERT, SciBERT) to acquire the semantic representation of ACs for ACC task.

4.2.2 Pipeline

ACC Task For reason we have explained, token-level evaluation is qualitative

• **GRU+CRF** [4] employs Gated Recurrent Unit (GRU) and Conditional Random Field (CRF) to address ACI task and ACC task jointly.

RI Task

• **Tree-LSTM** [34] employs Tree-LSTM to identify the type of relation between two ACs.

• **SentClf** [4] employs different transformers (BERT, BioBERT, SciBERT, Roberta) to encode ACps for RI task.

• **MultiChoice** [4] turns the RI task into a selection problem.

4.2.3 Multi-task

• **ResArg** [33] uses residual networks to address component type classification, relation classification and link prediction at the same time.

• **ResAttArg** [9] employed multi-tasking learning network to jointly address type classification, relation classification and link prediction, which improves previous work by introducing attention mechanism.

The word embeddings are acquired by using pre-trained GloVe embeddings [35] in the above model. Galassi et al. [9] employed two strategies to get the final results after repeating the training procedure 10 times using different seeds, that are (1) "average": considering the average of the results of 10 different networks; (2) "ensemble": assigning each element as the class voted by the majority of 10 networks.

4.3 Evaluation metrics and Implementation

Evaluation Metrics The evaluation metrics are F1 scores related to the macro average, the micro average, the *evidence* class and the *claim* class for ACC task. For RI task, the evaluation metric is macro F1.

Implementation The model is implemented with torch1.4 and trained on NVIDIA v100 GPU. For BioBERT encoder, the PyTorch implementation of huggingface version 2.3 is

used. The representation dimensions for word, AC, ACp and label are the same, which are set as 768. For BiLSTM, we set the hidden vector dimension $\delta = 768$. For GCN, the number of layer for GCN is 1 and the hidden size of GCN is 768. while training, adam optimizer is employed to update all parameters. The batch size is 1 that means that every input is a RCT abstract, the learning rate is 5×10^{-5}, and the temperature of GS τ is 0.05. The model is trained for 3 epochs. In the global loss function, λ_1 and λ_2 are set to 0.1 and 1.

4.4 Main Results

ACC Task The token-level and component-level experimental results of ACC task are presented in Table 3.

• **Token-level Evaluation** For reason we have explained, token-level evaluation is qualitative. We can observe from Table 3 that SeqMT achieves the best results on three test sets for what concerns the micro F1 score and macro F1 score. SeqMT outperforms all other models on the three test sets for what concerns E-F1 score. However, for what concerns C-F1, SeqMT is defeated by the BERT-based approach on the neoplasm test set, and by BioBERT-based approach and SciBERT-based approach on the glaucoma and mixed test sets.

Although the result of token-level analysis is only for reference, it can also illustrate the high performance of SeqMT in token-level classification.

• **Component-level Evaluation** It can be observed from Table 3 that: (1) our model achieves the best performance on three test sets, except *evidence* F1 on the neoplasm test set. Consistent with baseline, our model achieved better scores for *evidence* F1 than *claim* F1 on all test sets. In addition, our model performs better on the glaucoma test set than the other two test sets. The excellent results show that SeqMT can distinguish the types of ACs well. (2) For the approaches based on BERT or its variants, BiLSTM+BioBERT achieves the best performance on the neoplasm and glaucoma test set, and approximate results on the mixed test set compared with BiLSTM+SciBERT. BiLSTM+BERT obtains the worst performance on the three tset sets. The above analysis shows that it is effective to employ a pre-trained language model that incorporates biomedical knowledge in the face of biomedical problems. (3) For the approaches based on Golve embeddings, ResAttArg with ensemble achieves the best results on the three test sets. However, the performance of ResAttArg with ensemble is far worse than models that use BERT or its variants as encoder, which indicate the superiority of language model after pre-training on a large corpus.

RI Task Table 4 shows the experimental results of RI task. It can be observed from Table 4 that: (1) On the neoplasm test set, SeqMT achieves the best result, which beats the best baseline ResAttArg(Ensemble) by 0.32 percentage points. (2) On the glaucoma test set, SeqMT achieves a great improvement, which beats the best baseline ResAttArg(Ensemble) by 4.87 percentage points. (3) On the mixed test set, SeqMT achieves the best performance, which beats the best baseline SciBERT+SC by 3.71 percentage points.

In order to evaluate the performance of SeqMT comprehensively, we report some additional details about F1 scores in different classes on SeqMT and ResAttArg with ensemble in Fig 3. For *none* class and *support* class, SeqMT achieves the better performance than ResAttArg with ensemble on the three test sets. For *attack* class, SeqMT achieves the better performance on the glaucoma and mixed test sets, but on the neoplasm test set, the result of ResAttArg with ensemble is 11.54 percentage points higher than that of SeqMT. One possible reason is that ResAttArg with ensemble employs a special strategy, which introduces opposite relation types (e.g. A *attackedBy* B) during training, to alleviates the problem of category imbalance and helps *attack* class to be learned well [9]. In conclusion, we can find that the excellent results of SeqMT on the RI task comes from the balanced performance of three different classes.

The outstanding results in Table 4 and Fig 3 demonstrate the effectiveness of SeqMT in solving RI task of BAM.

4.5 Ablation Study

To explore the contributions of various components of our model, we perform the ablation study as follows: **SeqMT(-gcn)** without modeling the dependency among the related ACps in the same local neighborhood, **SeqMT(-le)** without transferring label information of ACs to RI task module, **SeqMT(-te)** without transferring features and labels information of ACs to RI task module, which means that SeqMT degenerates into a conventiona multi-task learning model, **SeqMT(-$\mathcal{L}_{acc}$)** without loss function of ACC task, which means that SeqMT only addresses RI task.

Effect of Pair-level Contextual Information We first explore how the model performance on different test sets is influenced by GCN component. The experimental results of **SeqMT(-gcn)** on three test sets of RI task are presented in Table 5. It can be observed from the first row in Table 5 that: After removing the GCN component of SeqMT, the results on the three test sets have dropped significantly, especially on the glaucoma test set, which dropped by 7.71 percentage points. The above results illustrate the importance of modeling the dependency among the related ACps in the same local neighborhood for RI task of BAM.

Effect of Sequential Information Transfer We explore the effectiveness of sequential information transfer between ACC task and RI task with the removal of information transfer module. The experimental results of **SeqMT(-le)** and **SeqMT(-te)** on three test sets of RI task are presented in Table 5. It can be observed from the second and third rows in Table 5 that: (1) On the neoplasm and mixed test sets, model shows a slight drop in performance without transferring the label information of ACs. However, the results of model have dropped significantly on the glaucoma test set, which means that label information of ACs is extremely important for glaucoma test set. (2) As the results on the neoplasm and glaucoma test sets show, the performance of **SeqMT(-te)** that does not transfer label and feature information of ACs, is worse than that of **SeqMT(-le)**. However, the opposite situation occurs on the mixed test set. One possible reason is that larger proportion of *support* and *attack* relation occur between evidence and claim in the mixed test set, which make label information of ACs more important. The above results illustrate that modeling the sequential relationship between ACC task and RI task is effective for BAM.

TABLE 3
Results of ACC on AbstRCTs(%). f_1 means micro F1 and $F1$ means macro F1. The binary F1 for claims are reported as C-F1 and for evidence as E-F1. Best results are marked in bold. RA means ResArg, RAA means ResAttArg, BL means BiLSTM.

		Neoplasm				**Glaucoma**				**Mixed**			
Level	Model	f_1	F_1	E-F1	C-F1	f_1	F_1	E-F1	C-F1	f_1	F_1	E-F1	C-F1
Token	BERT+GRU+CRF	89	85	78	90	89	86	76	89	90	88	81	91
	BioBERT+GRU+CRF	90	84	90	87	92	91	91	93	92	91	92	91
	SciBERT+GRU+CRF	90	87	92	88	91	89	91	93	91	88	93	90
	RA(avg)	90.66	88.20	93.58	82.81	91.84	87.49	94.86	80.11	91.25	87.79	94.29	81.29
	RA(Ensemble)	90.75	88.10	93.72	82.47	92.50	88.48	95.28	81.68	91.61	88.21	94.54	91.61
	RAA(avg)	90.80	88.60	93.59	83.61	92.02	88.02	94.93	81.11	91.58	88.72	94.40	83.04
	RAA(Ensemble)	92.12	90.04	94.56	85.72	92.92	89.35	95.52	83.19	92.79	90.26	95.23	85.30
	SeqMT	**94.38**	**92.96**	96.08	89.84	**93.76**	**94.02**	96.98	91.06	**94.51**	**94.26**	97.65	90.87
Component	RA(avg)	87.42	86.18	90.31	82.04	88.08	85.53	91.59	79.48	88.20	86.74	91.13	82.35
	RA(Ensemble)	87.76	86.38	90.71	82.05	89.39	87.13	92.53	81.74	89.00	87.59	91.77	83.42
	RAA(avg)	87.32	86.19	90.11	82.27	88.50	86.26	91.79	80.72	88.65	87.51	91.27	83.74
	RAA(Ensemble)	88.92	87.87	91.44	84.30	89.73	87.71	92.69	86.54	90.67	89.70	92.86	82.72
	BL + BERT	90.23	89.14	92.58	85.71	89.39	86.99	92.57	81.41	88.83	87.48	91.59	83.37
	BL + SciBERT	90.08	89.06	92.41	85.71	89.16	88.00	91.74	84.26	91.24	89.43	93.80	85.05
	BL + BioBERT	92.08	90.67	94.30	87.05	92.16	91.38	93.98	88.78	90.67	89.74	92.82	86.66
	SeqMT	**92.56**	**91.89**	94.22	**89.57**	**93.43**	**92.35**	**95.22**	**89.48**	**92.83**	**92.21**	**94.49**	**89.73**

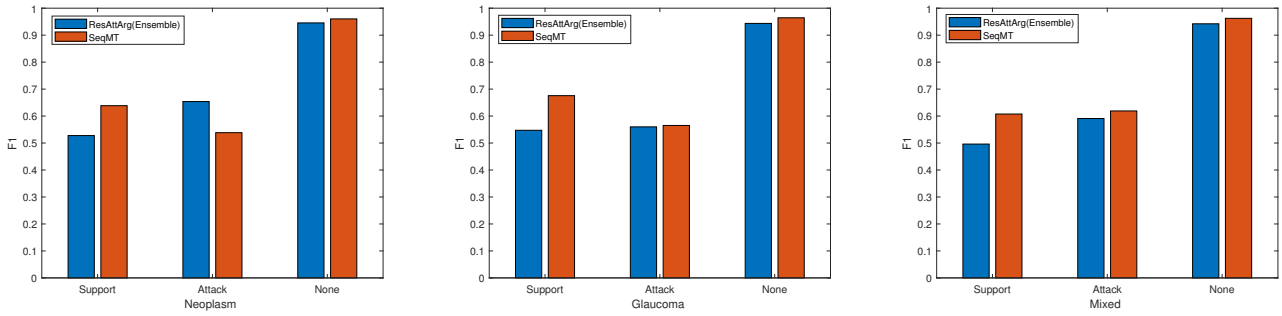


Fig. 3. $F1$ score of the respective classes of different models

Effect of Multi-task Framework We explore the impact of the multi-task framework on the performance of the model by retaining only the loss weight of the RI task The experimental results of **SeqMT(-$\mathcal{L}_{acc}$)** on three test sets of RI task are presented in Table 5. It can be observed from the fourth row in Table 5 that the multi-task learning framework helps the model to improve performance by sharing relevant information between tasks.

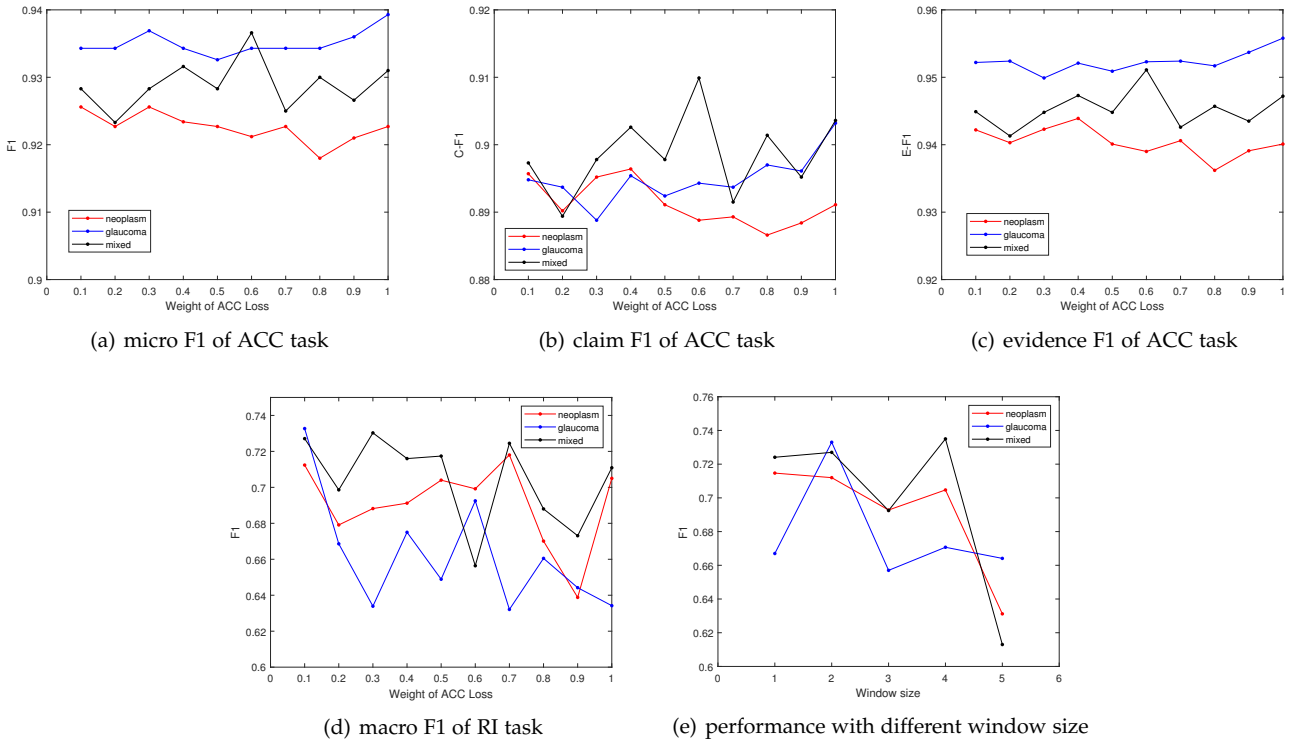
Overall, our model SeqMT with all four components gives the best performance on three test sets.

4.6 Hyper-parameter Analysis

In this section, we explore the impact of different hyper-parameters on model performance by conducting some comparative experiments. The experimental results are plotted in Fig 4.

Effect of Different Weight of Loss We keep the weight of loss of RI task and change the weight of loss of ACC task constantly to explore the most suitable weight of loss combination.

From the plot, it can be observed that: (1) For micro F1, claim F1 and evidence F1 of ACC task, the performance on the glaucoma and neoplasm test sets fluctuates slightly when $\mathcal{L}_{acc}$ increases at first, and then the performance improves when $\mathcal{L}_{acc}$ is close to 1. However, the performance of mixed test set fluctuates drastically when $\mathcal{L}_{acc}$ increases. One possible reason is that the internal connection between ACC task and RI task on the neoplasm test set is more obvious, which makes model difficult to deal with ACC task without considering the impact of RI task. (2) For macro F1 of RI task, the performance on different test sets varies greatly when $\mathcal{L}_{acc}$ increases. On the neoplasm test set, the performance in terms of macro F1 fluctuates slightly when $\mathcal{L}_{acc}$ increases from 0.1 to 0.7, and fluctuates drastically when $\mathcal{L}_{acc}$ increases from 0.8 to 1. On the glaucoma test set, the overall performance shows a downward trend when



(a) micro F1 of ACC task (b) claim F1 of ACC task (c) evidence F1 of ACC task

(d) macro F1 of RI task (e) performance with different window size

Fig. 4. The performance of SeqMT with different settings of hyper-parameter.

TABLE 4
Results of RI on AbstRCTs(%). Best results are marked in bold. MC means MultiChoice and SC means SentClf.

Model	Neoplasm	Glaucoma	Mixed
Tree-LSTM	37	44	39
BERT-MC	58	56	55
BioBERT-MC	61	58	57
SciBERT-MC	63	59	60
BERT-SC	62	53	66
BioBERT-SC	64	58	61
SciBERT-SC	68	62	69
RoBERTa-SC	67	66	67
ResArg(avg)	59.15	57.23	60.31
ResArg(Ensemble)	63.16	61.86	68.35
ResAttArg(avg)	66.49	62.68	63.47
ResAttArg(Ensemble)	70.92	68.40	67.66
SeqMT	**71.24**	**73.27**	**72.71**

TABLE 5
Ablation study results of RI task(%).

Model	Neoplasm	Glaucoma	Mixed
SeqMT(-gcn)	67.62	65.56	69.20
SeqMT(-te)	69.15	62.90	71.04
SeqMT(-le)	71.20	64.84	70.90
SeqMT(-$\mathcal{L}_{acc}$)	68.58	64.83	70.30
SeqMT	**71.24**	**73.27**	**72.71**

$\mathcal{L}_{acc}$ increases. On the mixed test set, the performance fluctuates drastically when $\mathcal{L}_{acc}$ increases. Based on the above analysis, we can find that the overall performance of the ACC and RI task is the best when $\mathcal{L}_{acc}$ is 0.1 and $\mathcal{L}_{ri}$ is 1.

Effect of Different Window Size We also investigate the influence of different window size on performance of RI task. The experimental results are presented in Figure 4. It can be observed that (1) The F1 value on the neoplasm test set shows a downward trend, when window size increases from 1 to 3. Then, a slight improvement occurs When window size changes from 3 to 4. The worst result is acquired when window size is 5. (2) For glaucoma test set, the F1 value fluctuates drastically when window size increases from 1 to 5. The best result is acquired when window size is 2. (3) when window size is 1, 2 and 4, the F1 value in mixed test is high, and terrible results are obtained, when window size is 3 and 5. (4) when window size is 2, the F1 value achieves good results in all three test sets. The results decline significantly when window size increases from 4 to 5, especially on the neoplasm and mixed test sets, indicating that irrelevant pair-level information has been captured, which has bad influence on performance of RI task. Based on the above analysis, we can conclude that the most suitable window size is 2.

TABLE 6
Case study of RI task. Label is the ground-truth.

Argument component pair	Label	SciBERT+SC	SeqMT
[Acute toxicity (WHO) of RCT was low, with less than 15% of patients experiencing grade 3 or higher toxicity]$_1$ [Neoadjuvant RCT is well tolerated and bears no higher risk for postoperative morbidity]$_4$.	*Support*	*Support*	*Support*
[The principal toxicity was diarrhea, with 12% in the postoperative RCT-arm and 11% in the pre-operative RCT-arm having grade 3-, and 1% in either arm having grade 4-diarrhea]$_2$. [Neoadjuvant RCT is well tolerated and bears no higher risk for postoperative morbidity]$_4$.	*Support*	*Support*	*Support*
[Erythema, nausea and leukopenia were the next common toxicities, with less than 3% of patients in either arm suffering grade 3 or greater leukopenia or nausea.]$_3$. [Neoadjuvant RCT is well tolerated and bears no higher risk for postoperative morbidity]$_4$.	*Support*	*None*	*Support*

4.7 Case Study

The Table 6 illustrates an example from the mixed test set to analyze the importance of modeling pair-level context information specifically. We select one claim and three consecutive evidences which support this claim from the same abstract. Then, We manually construct three ACps, which are listed in the table (three evidences are marked with subscripts 1, 2, 3 and the claim is marked with subscripts 4). It can be observed from the table that (1) Both SciBERT+SC and SeqMT predict the relation types of ACp1 and ACp2 are *support*, due to *evidence*1, *evidence*2 and *claim* all mention the related content of 'RCT' and low toxicity risk. (2) The relation type of AC3 is identified as *none* by SciBERT+SC, but SeqMT can correctly predict its relation type as *support*. In fact, it is difficult to identify the relation type of ACp3, just based on text information of itself, due to the experimental drug 'RCT' is not mentioned in $evidence$3. However, combining the contextual information of ACp3 from ACp1 and ACp2, we can know that AC3 describes the fact that RCT cause some slight toxicities, which means that the relation type of ACp3 is the same as ACp1 and ACp2. The above analysis show that modeling the pair-level contextual information from related ACps plays a important role in RI task and SeqMT can model this dependency relationship well.

4.8 Error Analysis

We randomly selected 50 incorrect instances of RI task from the three test sets, and categorize the main errors. The first type of error is caused by the ability of numerical reasoning. This is because some evidences contain comparisons experiment results and the corresponding claim is the conclusion after comparisons. For example, the evidence is "*The mean TWIST was 27.05 months with CAP, 31.5 months with ChOP and 32.95 months with fludarabine*" , where claim describes that "*patients with advanced CLL have a moderate benefit in terms of Q-TWIST when treated with fludarabine over ChOP*". The model cannot draw the conclusion of the median by comparing three numbers. The second type of error is due to inability to understand the terminology in the biomedical domain. For example, the evidence is that "*Many severely anemic and transfusion-dependent patients with advanced MM, NHL, and CLL and a low performance status benefited from epoetin therapy, with elimination of severe anemia and transfusion need, and improvement in QOL*", which poses a challenge to the model's understanding of professional vocabulary and acronyms. The third type of error is thanks to the lack of background knowledge. For example, the evidence is that "*Cox regression analysis showed an estimated hazards ratio of 1.309 (P =.052) favoring epoetin alfa*", the claim is that "*Epoetin alfa safely and effectively ameliorates anemia and significantly improves QOL in cancer patients receiving nonplatinum chemotherapy*". The model cannot understand that *hazards ratio of 1.309 (p = .052)* means safe and effective without enough background knowledge.

5 Conclusions

In this paper, we propose a novel Sequential Multi-task Learning model named SeqMT, to jointly address argument component classification and and relation identification of BAM. The proposed SeqMT has abilities to model the sequential relationship between the ACC task and RI task by introducing information transfer module, where the label of ACs is sampled by the gumbel-softmax method and passed to RI task together with its features. In addition, SeqMT can also model pair-level contextual information for a given ACp by aggregating information from related ACps with GCN. We evaluated SeqMT on a public dataset, which contains 659 abstracts of randomized clinical trial papers about five diseases that are neoplasm, glaucoma, hypertension, hepatitis and diabetes. The experimental results show that SeqMT defeats the most advanced baseline on three test sets, which shows the effectiveness of SeqMT. Therefore, we believe that our model SeqMT can promote the development of intelligent healthcare by providing a means to automatically extract important information from large-scale medical data and present this information in a structured way

In the future, we will focus on the imbalanced problem on the BAM dataset and the problem of lack of interpretability, which is important in biomedical issues especially. For the imbalanced problem, we consider designing a special loss function, similar to focal loss, to assign larger weight to rare categories. For the problem of lack of interpretability, we consider using a text generation method with an attention mechanism to generate text spans, which are utilized to explain results of the predictive model.

ACKNOWLEDGMENTS

The authors would like to thank...

REFERENCES

[1] R. Craven, F. Toni, C. Cadar, A. Hadad, and M. Williams, "Efficient argumentation for medical decision-making," in *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

[2] A. Hunter and M. Williams, "Aggregating evidence about the positive and negative effects of treatments," *Artificial intelligence in medicine*, vol. 56, no. 3, pp. 173–190, 2012.

[3] M. Al Qassas, D. Fogli, M. Giacomin, and G. Guida, "Analysis of clinical discussions based on argumentation schemes," *Procedia Computer Science*, vol. 64, pp. 282–289, 2015.

[4] T. Mayer, E. Cabrio, and S. Villata, "Transformer-based argument mining for healthcare applications," in *ECAI 2020, 24th European Conference on Artificial Intelligence*, 2020.

[5] I. Persing and V. Ng, "End-to-end argumentation mining in student essays," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1384–1394.

[6] C. Stab and I. Gurevych, "Parsing argumentation structures in persuasive essays," *Computational Linguistics*, vol. 43, no. 3, pp. 619–659, 2017.

[7] M. Zhu, B. Celikkaya, P. Bhatia, and C. K. Reddy, "Latte: Latent type modeling for biomedical entity linking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9757–9764.

[8] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi, "Fact or fiction: Verifying scientific claims," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7534–7550.

[9] A. Galassi, M. Lippi, and P. Torroni, "Multi-task attentive residual networks for argument mining," *arXiv preprint arXiv:2102.12227*, 2021.

[10] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[11] M.-F. Moens, E. Boiy, R. M. Palau, and C. Reed, "Automatic detection of arguments in legal texts," in *Proceedings of the 11th international conference on Artificial intelligence and law*, 2007, pp. 225–230.

[12] E. Florou, S. Konstantopoulos, A. Koukourikos, and P. Karampiperis, "Argument extraction for supporting public policy formulation," in *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 2013, pp. 49–54.

[13] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim, "Context dependent claim detection," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 1489–1500.

[14] M. S. Rasooli and J. R. Tetreault, "Yara parser: A fast and accurate dependency parser," *Computing Research Repository*, vol. arXiv:1503.06733, 2015, version 2. [Online]. Available: http://arxiv.org/abs/1503.06733

[15] M. Lippi and P. Torroni, "Context-independent claim detection for argument mining," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[16] H. Wang, Z. Huang, Y. Dou, and Y. Hong, "Argumentation mining on essays at multi scales," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5480–5493.

[17] A. Peldszus and M. Stede, "Joint prediction in mst-style discourse parsing for argumentation mining," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 938–948.

[18] P. Potash, A. Romanov, and A. Rumshisky, "Here's my point: Joint pointer architecture for argument mining," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1364–1373.

[19] S. Eger, J. Daxenberger, and I. Gurevych, "Neural end-to-end learning for computational argumentation mining," *arXiv preprint arXiv:1704.06104*, 2017.

[20] V. Niculae, J. Park, and C. Cardie, "Argument mining with structured svms and rnns," *arXiv preprint arXiv:1704.06869*, 2017.

[21] T. Mayer, E. Cabrio, M. Lippi, P. Torroni, and S. Villata, "Argument mining on clinical trials." in *COMMA*, 2018, pp. 137–148.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[23] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[24] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3606–3611.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[26] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7370–7377.

[27] K. Sun, R. Zhang, S. Mensah, Y. Mao, and X. Liu, "Aspect-level sentiment analysis via convolution over dependency tree," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5683–5692.

[28] Y. Chen, W. Hou, S. Li, C. Wu, and X. Zhang, "End-to-end emotion-cause pair extraction with graph convolutional network," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 198–207.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[30] G. Liu and J. Guo, "Bidirectional lstm with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, 2019.

[31] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[32] W. Chen, J. Tian, L. Xiao, H. He, and Y. Jin, "Multi-task learning for logically dependent tasks from the perspective of causal inference," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2213–2225.

[33] A. Galassi, M. Lippi, and P. Torroni, "Argumentative link prediction using residual networks and multi-objective learning," in *Proceedings of the 5th Workshop on Argument Mining*, 2018, pp. 1–10.

[34] M. Miwa and M. Bansal, "End-to-end relation extraction using lstms on sequences and tree structures," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1105–1116.

[35] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

PLACE PHOTO HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.