Journal of Biomedical Informatics 117 (2021) 103767

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

TransforMED: End-to-End Transformers for Evidence-Based Medicine and Argument Mining in medical literature

Nikolaos Stylianou [*], Ioannis Vlahavas

School of Informatics, Aristotle University of Thessaloniki, Greece

ARTICLE INFO

Keywords:
Deep learning
Natural language processing
Evidence based medicine
Argument mining

ABSTRACT

Argument Mining (AM) refers to the task of automatically identifying arguments in a text and finding their relations. In medical literature this is done by identifying Claims and Premises and classifying their relations as either Support or Attack. Evidence-Based Medicine (EBM) refers to the task of identifying all related evidence in medical literature to allow medical practitioners to make informed choices and form accurate treatment plans. This is achieved through the automatic identification of Population, Intervention, Comparator and Outcome entities (PICO) in the literature to limit the collection to only the most relevant documents. In this work, we combine EBM with AM in medical literature to increase the performance of the individual models and create high quality argument graphs, annotated with PICO entities. To that end, we introduce a state-of-the-art EBM model, used to predict the PICO entities and two novel Argument Identification and Argument Relation classification models that utilize the PICO entities to enhance their performance. Our final system works in a pipeline and is able to identify all PICO entities in a medical publication, the arguments presented in them and their relations.

1. Introduction

Evidence-Based Medicine (EBM) enables medical practitioners to form treatment plans based on the complete available evidence. These evidence are the products of research studies, usually in the forms of Randomized Control Trials (RCTs) or Clinical Trials (CTs) that investigate the effects of a treatment on a specific group of patients and present their findings.

The increasingly growing number of medical publications is making it extremely difficult for healthcare staff and medical practitioners to stay updated with the latest research and guidelines [1]. During the 2020 COVID-19 pandemic, more than fifty thousand research articles about the novel coronavirus have been published on PubMed[1] alone. The PICO framework, named after its elements, Population, Intervention, Comparator, Outcome, enables refined searching in RCTs and CTs [2]. However, the process remains time-consuming and the information presented is still in a very raw and unstructured form [3].

Recent advancements in Medical Argument Mining (MAM) [4–7] along with the publication of high quality datasets and systems, enables further processing of medical documents and their presentation in a compact, practical form. The task of Argument Mining (AM), and by extension MAM, is a two-part problem of first identifying the arguments in text (Argument Identification) and second identifying their relations (Argument Relation Classification). These relations are used to build argument graphs with the relation types being the connections and the arguments being the edges.

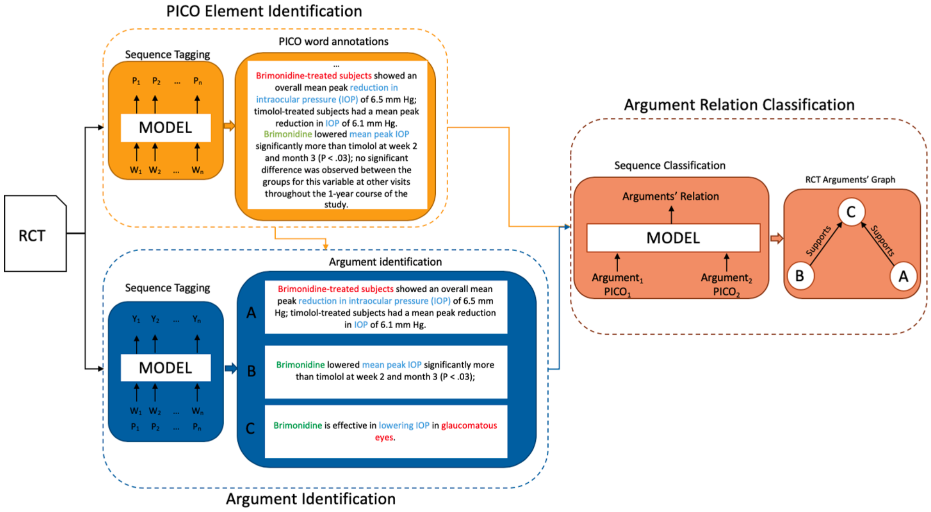
However, even though strides have been made, these systems rarely work in unison and are usually presented as individual components. What is more, the advancements in EBM can greatly benefit MAM systems by introducing latent semantic information in the raw data and increase the model's performance. Consequently, medical practitioners would only need to read the research directly related to each case while also being presented with latent arguments extracted from the text, annotated with PICO entities, along with their relations.

In this paper we present TransforMED, a state-of-the-art EBM and MAM system that is comprised of three models working in unison. TransforMED uses the EBM model to extract PICO entities from medical publications that are then used in the MAM system. The MAM system is consisted of an Argument Identification model and an Argument Relation Classification model, working in a pipeline. The trained system masters performance increase over the previous state-of-the-art in all three models individually and in overall performance. Furthermore, we

[*] Corresponding author at: School of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece.
E-mail addresses: nstylia@csd.auth.gr (N. Stylianou), vlahavas@csd.auth.gr (I. Vlahavas).

[1] https://www.ncbi.nlm.nih.gov/pubmed/.

https://doi.org/10.1016/j.jbi.2021.103767
Received 2 December 2020; Received in revised form 3 March 2021; Accepted 28 March 2021
Available online 31 March 2021



...
Brimonidine-treated subjects showed an overall mean peak reduction in intraocular pressure (IOP) of 6.5 mm Hg; timolol-treated subjects had a mean peak reduction in IOP of 6.1 mm Hg. Brimonidine lowered mean peak IOP significantly more than timolol at week 2 and month 3 (P < .03); no significant difference was observed between the groups for this variable at other visits throughout the 1-year course of the study.

A: Brimonidine-treated subjects showed an overall mean peak reduction in intraocular pressure (IOP) of 6.5 mm Hg; timolol-treated subjects had a mean peak reduction in IOP of 6.1 mm Hg.

B: Brimonidine lowered mean peak IOP significantly more than timolol at week 2 and month 3 (P < .03);

C: Brimonidine is effective in lowering IOP in glaucomatous eyes.

Fig. 1. TransforMED system pipeline view. Example text appears in Red for Populations, Green for Interventions and Light Blue for Outcomes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

demonstrate the effects of EBM to MAM and provide higher-level output by combining the merits of EBM and MAM when working in concert.

2. Related work

Evidence-Based Medicine has been tackled for many years, with a plethora of approaches and methodologies [8–10]. In latest research, EBM is approached as a problem of identifying PICO elements in RCTs and CTs. The scope of such approaches has also become increasingly detailed, starting from identifying PICO elements to sentences [11,12] and reaching the ideal of identifying the specific spans that describe the PICO elements in the sentences through the use of newly created datasets [13] and state-of-the-art systems [14,15]. Specifically, [14] introduces a Recurrent Neural Network (RNN)-based approach, with the use of ELMo [16] and character-based word embeddings, utilizing a Conditinal Random Fields (CRF) layer to identify PICO elements in a sequence-to-sequence approach. In [15], PICO identification is approached as a Question-Answering task using BERT [17], a Transformer-based neural language model, to extract high-quality representations to be used by a simple RNN-based network.

Argument Mining has been of increasing research interest in the last decade [18–21] and has been attempted in a variety of domains [22–24], including healthcare [5,7], however few of these approaches handle both AM tasks [7]. In [22], the proposed method is based on feature-based linear programming to jointly model both tasks. The approach however is limited by the self-imposed constraint that each document, in this case essay, has only one major claim. An end-to-end TreeLSTM [24] based approach is used in [25] along with a decoupling of the two tasks to improve the performance and account for the propagated error. While their approach has merit, the TreeLSTM architecture does not scale well with long texts, imposing distance constraints to the results. In a similar approach, [7] uses Transformer-based language models, along with RNNs to address both tasks and remove previously introduced architectural constraints.

Recently, works focused only on specific subtasks of AM, due to the high performance achieved in argument identification. In [26,27] they target only argument linking, an intermediate step between Argument Identification and Argument Relation Classification. This is achieved with an attention based pointer network in [26], while [27] uses structured learning. Argument Relation Classification was also tackled as a single task in [27], through BERT. However, these tasks assume that the arguments are already identified and are unable to perform as standalone systems.

3. Material and methods

TransforMED highlights the effects of Evidence-Based Medicine to the performance of the Argument Identification and Argument Relation Classification models. The system, which consists of three end-to-end Transformer models, outperforms all other AM systems on both tasks, while maintains a better performance than dedicated EBM systems on the PICO element identification task. Additionally, TransforMED outputs information on a higher level than current AM systems as each document is enhanced with PICO annotations, combining the merits of EBM and AM.

3.1. System overview

TransforMED works in a pipeline, as illustrated on Fig. 1. For every new document, we first identify PICO annotation through our EBM model. Then the PICO annotated document is used to extract arguments. The extracted arguments, along with the respective PICO annotations are used for argument relation classification. As a result, for each document, TransforMED outputs the Population, Intervention/Comparator and Outcome of the study as well as the arguments identified and their relations which can be later constructed into a graph.

In practice, the output of TransforMED will be a PICO annotated version of the RCT that was given as input, along with the identified and classified arguments and their relations. Providing an example from literature, using the abstract of [28] as the input RCT to TransforMED (Fig. 1), in the following excerpt from that abstract we highlight the PICO annotations as predicted by TransforMED.

> "*Brimonidine-treated subjects showed an overall mean peak reduction in intraocular pressure (IOP) of 6.5 mmHg; timolol-treated subjects had a mean peak reduction in IOP of 6.1 mmHg. Brimonidine lowered mean peak IOP significantly more than timolol at week 2 and month 3 (P < .03); no significant difference was observed between the groups for this variable at other visits throughout the 1-year course of the study.*"

We used Red, Green and Light Blue colors for the spans of text that identify Population, Intervention and Outcome entities respectively.

What is more, in the same RCT, TransforMED also identified three arguments, classified them and categorized them based on their relations. The arguments identified are: (A) "*Brimonidine-treated subjects showed an overall mean peak reduction in intraocular pressure (IOP) of 6.5 mmHg; timolol-treated subjects had a mean peak reduction in IOP of 6.1*

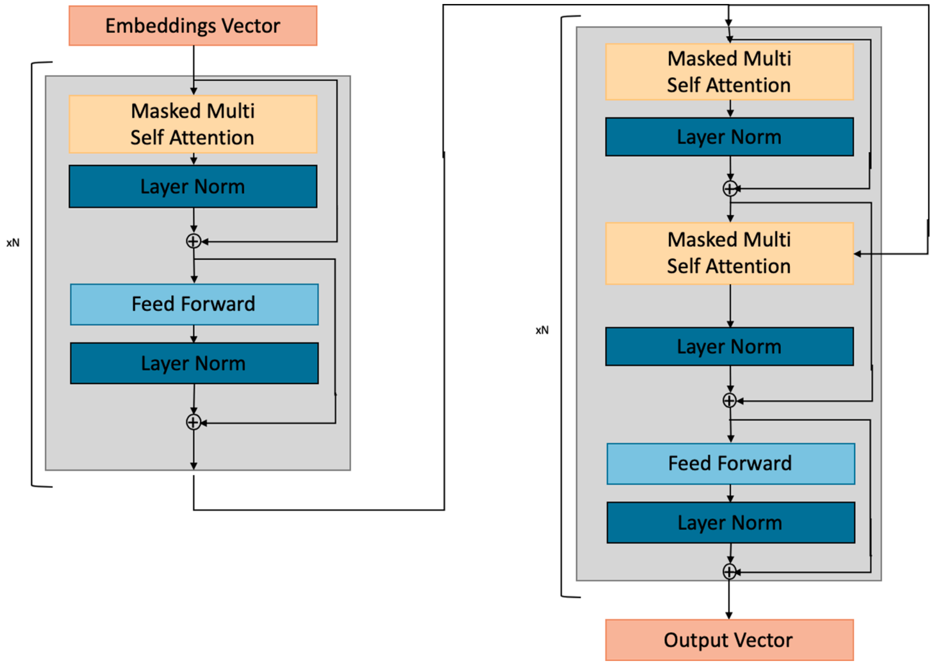


Fig. 2. TransforMED Encoder Decoder Layers.

mmHg.", (B) "*Brimonidine lowered mean peak IOP significantly more than timolol at week 2 and month 3 (P* < *.03)*" and (C)"*Brimonidine is effective in lowering IOP in glaucomatous eyes.*", all classified as *Evidence*. Their relations are categorized as A *Supports* C, B *Supports* C, and A has *No Relation* to B.

3.2. Data and preprocessing

TransforMED uses two datasets for training and evaluation. The EBM model is trained using the EBM-NLP corpus [13], which consists of 5000 annotated abstracts of medical publications and provides hierarchical PICO annotations. The AM model is trained using the dataset published alongside [7], referred to as AMCT in this study, which contains 919 argument components across 159 abstracts. The data collection in AMCT consists of five different disease types (neoplasm, glaucoma, hepatitis, diabetes and hypertension) which are split in three groups of data (Neoplasm, Glaucoma and Mixed) based on the type of diseases described in the document that comprise them. Formally, Neoplasm is used for both training and evaluation based on the predefined splits, while Glaucoma and Mixed are only used as out-of-domain evaluation due to their different vocabulary and smaller size.

For our EBM model, we use the corpus directly, without any major preprocessing. We opted to use the higher level annotations instead of the hierarchical as in our use case we are only interested about the PICO entities themselves not the fine details of each entity. Additionally, the Intervention and Comparator entities are considered to be the same entity and treated as one [13]. As a result we identify *Population*, *Intervention* and *Outcome* using an IO (*Inside*, *Outside*) format for the labels to allow for sequence-to-sequence predictions.

The AM models uses the AMCT corpus in two formats for Argument Identification and for Argument Relation Classification respectively, both containing the same collection of RCT abstracts.

For Argument Identification, we convert the data in an IO format to facilitate a sequence-to-sequence prediction. In comparison to [7], in which they use IOB (*Inside*, *Outside*, *Before*) to represent the target labels, we remove the *Before* label as the arguments tend to be consisted of a large number of tokens, resulting in a heavy imbalance between *Before* and *Inside* labels without contributing any additional information. Furthermore, similarly to [7], we replaced the labels of all *Major Claim* arguments to just *Claim*. This transformation is due to the fact that both *Claim* and *Major Claim* labels are not handled differently by the pipeline and that the AMCT corpus identifies only 125 *Major Claim* labeled tokens in all 159 abstracts, while *Evidence* and *Claim* have 2808 and 1625 occurrences respectively.

For Argument Relation Classification, the identified *Claim* and *Evidence* arguments, i.e. the spans of text previously identified, are grouped in pairs. The AMCT corpus identifies four classes, *Support*, *Attack*, *Partial-Attack* and *No Relation*, indicating supporting, attacking on no relation accordingly. We merge *Partial-Attack* and *Attack* as there are no occurrences of the first in the training set. Even after merging *Attack* class only has 200 and 113 occurrences in the training and test sets respectively, accounting for only 10% of the total instances.

In all models we used WordPiece for the tokenization along with the "[CLS]" and "[SEP]" tokens to accommodate the BERT-required input representation when used. To achieve uniform inputs, padding is applied accordingly, using the "[PAD]" token, after the last occurrence of the "[SEP]" token when the sequence is smaller than the maximum sequence length. If the sequence is larger than maximum sequence length, it is trimmed to match it. For the EBM and Argument Identification models there is only one occurrence as they model one sequence at a time. The Argument Relation Classification model uses two arguments (sequences) at a time, as it models pairs of pre-identified arguments, and hence the padding is applied after the second occurrence of the "[SEP]" token.

3.3. Transformers

Our models are built using the following Encoder and Decoder architectures (Fig. 2), each with its own set of parameters and configuration. These architectures are originally described in [29]. Each Transformer layer takes a sequence of vectors as input and outputs a sequence of vectors. Both the Encoder and the Decoder are consisted of N number of layers and h number of heads. The encoder can be used independently of the decoder and does not require an decoder layer to follow, while the decoder can only be used after an encoder layer.

The encoder is composed of two sublayers, a multi-headed self-attention layer and a feed-forward layer. These layers are applied sequentially, while both are preceded by a normalization layer [30] as

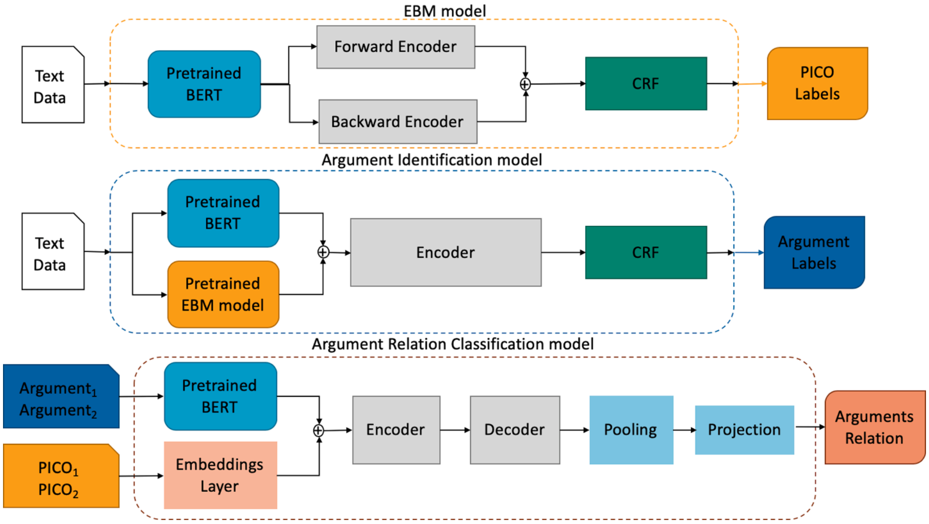


Fig. 3. TransforMED models' architectures.

well as a residual connection [31]:

$$Y = (\mathtt{norm}(\mathtt{self}-\mathtt{attention}(X))+X) \tag{1}$$

$$Z = (\mathtt{norm}(\mathtt{feed}-\mathtt{forward}(Y))+Y) \tag{2}$$

The decoder is composed of three sublayers, two multi-headed self-attention layers and a feed-forward layer. The layers are also applied sequentially in an identical manner to the encoder, using normalization and residual connection layers.

$$T = (\mathtt{norm}(\mathtt{self}-\mathtt{attention}(T))+T) \tag{3}$$

$$P = (\mathtt{norm}(\mathtt{self}-\mathtt{attention}(T,Z))+T) \tag{4}$$

$$H = (\mathtt{norm}(\mathtt{feed}-\mathtt{forward}(P))+P) \tag{5}$$

The first self-attention layer applies changes to each element of the input sequence by looking at the other elements, while the second self-attention mechanism computes the attention over the encoder output. The feed-forward layer processes each element independently via a two-layer Multilayer Perceptron (MLP). These layers are described in detail in [29].

3.4. Model architectures

TransforMED is an end-to-end transformer system, that does not use any RNN layers. As a result, all three models used in the pipeline are based on the previously described Transformer Encoder and Decoder architectures. The pre-trained BERT model creates semantically richer representations from the input sequence, modelling the sequence at a token level, as it is trained on a significantly larger collection of documents than our training corpora. The following Transformer Encoder layers extract latent features from these representations using the whole sequence, which are then utilized accordingly by each model. The Transformer Decoder layers are used only in our Argument Relation Classification model as we are dealing with two sequences and further modelling is required to define the relation between the two given sequences.

Evidence-Based Medicine. The EBM model takes as input a sequence of tokens $W = (w_1,\ldots,w_l)$ which are passed through a BERT model to produce a sequence of embeddings $X = (x_1,\ldots,x_n)$, where l is the number of tokens in the original sentence and n is the maximum number of tokens in a sequence, after padding. The resulted vector $X \in \mathbb{R}^{n\times d_{BERT}}$, where d_{BERT} is the BERT embedding size, is considered a forward vector ($\overrightarrow{X}$) and is also reversed to form a backward vector $\overleftarrow{X}$. We opt to use both a forward and a backword encoder, as the PICO entities are highly volatile and contextual information of what follows is required to make accurate predictions. To that end, we use our forward and backward encoders similarly to how a Bidirectional Long Short-Term Memory (BiLSTM) network operates, i.e. the two vectors are passed through individual Transformer Encoders with N_{ebm} number of layers, each with h_{ebm} attention heads, concatenated and then passed through a CRF layer that handles predictions. CRF has been proven effective when predicting sequence tags with strong dependencies[32].

Argument Identification. The ArgId model takes as input a sequence of tokens $W = (w_1,\ldots,w_l)$ which are passed through a BERT model to produced a sequence of embeddings $X = (x_1,\ldots,x_n)$ and a sequence of PICO annotations $L = (l_1,\ \ldots,\ l_l)$. The PICO annotations are passed through a embedding layer to produce a sequence of embeddings $P = (p_1,\ldots,p_n)$, where l is the number of tokens in the original sentence and n is the maximum number of tokens in a sequence, after padding. The resulted vectors $X \in \mathbb{R}^{n\times d_{BERT}}$ and $P \in \mathbb{R}^{n\times d_{PICO}}$, where d_{BERT} and d_{PICO} are the BERT and PICO embedding sizes respectively, are concatenated to form $XP \in \mathbb{R}^{n\times (d_{BERT}+d_{PICO})}$. In comparison to our EBM model, we only pass vector XP through a Transformer Encoder with N_{ArgId} number of layers, each with h_{ArgId} attention heads. As Argument Identification is a semantically simpler task, where arguments are identified as long strings of text and their entity type is not dependent to other predictions made in the same sequence, we do not need to take into account the following context. Finally, a CRF layer handles predictions, following the same design principle as our EBM model.

Argument Relation Classification. The ArgRC model takes as input two sequences of tokens $W_1 = (w_{1_1},\ldots,w_{1_k})$ and $W_2 = (w_{2_1},\ldots,w_{2_m})$, where k and m are the number of tokens in each argument. Along with the arguments, the model is provided with the respective PICO annotations $L_1 = (l_{1_1},\ldots,l_{1_k})$ and $L_2 = (l_{2_1},\ \ldots,\ l_{2_m})$. Both argument sequences concatenated, before padding, with the "SEP" special token inserted between the two sequences, before being passed through a BERT model to produce a sequence of embeddings $X = (X_{1_1},\ldots,X_{1_k},X_{sep},X_{2_1},\ldots,X_{2_m},\ldots,X_n)$, of size n due to padding. For PICO annotations, we apply the same concatenation and padding, making $L = (L_{1_1},\ldots,L_{1_k},L_X,L_{2_1},\ldots,L_{2_m},\ldots,L_n)$ before passing it through an embedding layer that results in the sequence vectors P. The resulted vectors $X \in \mathbb{R}^{n\times d_{BERT}}$ and $P \in \mathbb{R}^{n\times d_{PICO}}$, where d_{BERT} and d_{PICO} are the BERT and PICO embedding sizes respectively, are concatenated to form $XP \in \mathbb{R}^{n\times (d_{BERT}+d_{PICO})}$ which is passed through a Transformer Encoder layer, with N_{ArgRC} number of layers, each with h_{ArgRc} attention heads. The encoded output is forward through a Transformer Decoder, with the same number of layers and attention

Table 1
EBM Models performance comparison.

Models	Population			Intervention			Outcome			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERT	77.48	59.64	67.40	52.81	62.03	57.05	72.15	**84.72**	**77.93**	69.25	66.13	67.65
EBM+	**81.63**	80.03	80.81	54.97	71.22	62.05	80.14	58.81	67.84	70.93	70.02	70.47
QA-PICO	*87.00*	*88.00*	*87.00*	*74.00*	*78.00*	*74.00*	*68.00*	*69.00*	*67.00*	–	–	75.00
TransforMED	76.97	**91.24**	**83.49**	69.04	**79.24**	**73.79**	**84.38**	60.18	70.25	**79.32**	**80.33**	**79.82**

heads. The decoder output is passed through a pooling layer and a projection layer that handles classification predictions.

In Fig. 3 we illustrate the three models' architectures as previously described. The Encoder and Decoder layers are presented in an abstract form, as design information have been previously described (Section 3.3 and Fig. 2).

3.5. Experimental setup

Each model is trained individually, and has a unique set of parameters and weights. The trained models work in unison to produce the final outcome and are evaluated using the output of the previously trained models in the pipeline as input and compared to the gold labels. To that end, the ArgId and the ArgRC models both use PICO entity annotations from the EBM model, while the ArgRC model uses the identified arguments from the ArgId model.

In all models we experiment with different parameters for batch sizes {8, 16, 32}, learning rates {2e-4, 2e-5, 2e-6}, epochs {3, 5, 10, 15} and for PICO embeddings dimensions d_{PICO} {5, 20, 50, 100}. We also experimented with different number of layers {2, 4, 6, 8} and heads {2, 4, 6, 8}, as well as combinations between the two. Max sequence length for all tasks was set to $n = 128$ tokens. We used SciBERT [33] for all our BERT embeddings with $d_{BERT} = 768$. For the final models presented bellow, we used batch size 16, trained for 3 epochs and with a learning rate of 2e-5. For ArgId and ArgRC, $d_{PICO} = 100$. In all models we have the same number of layers and heads, resulting in $N_{ebm} = h_{ebm} = 4$, $N_{ArgId} = h_{ArgId} = 4$ and $N_{ArgRC} = h_{ArgRC} = 4$. Following the architecture design in [29] we use a 0.1 dropout inside the Encoder and Decoder layers.

The experiments were run on a computer with a single Titan V 12 GB graphics card, 32 GB of memory and an Intel i7-8700 processor. The code to reproduce the experiments is available at https://github.com/nstylia/transformed/.

4. Results

We compare the performance of our TransforMED system in two parts. First we compare our EBM model to EBM-NLP [13], BERT [17], EBM+ [14] and QA-PICO[15], which are the previous state-of-the-art EBM models, in terms or Precision, Recall and F1-score.

Secondly, we compare both AM tasks with AMCT [7], which uses as a basis a fine-tuned SciBERT model (AMCT-Sci) or a fine-tuned BioBERT model (AMCT-Bio) with a Bidirectional Gated Recurrent Unit (BiGRU) network and a CRF layer, and evaluate them in terms of micro and macro averaged F1-score. We also compare the Argument Relation Classification model to TreeAM [25] and ResidualAM [27] and Roberta [34] in terms of F1-score. Most noteably, while the performance of the Argument Identification and Argument Relation Classification models are shown independently, the Argument Relation Classification models models are evaluated using the predicted arguments from the best Argument Identification model, TransforMED, to provide a fair comparison.

Table 1 compares the performance of our EBM model, trained on the EBM-NLP dataset.We use bold to identify the best overall values for models predicting all PICO elements simultaneously, and underline for the best values per metric per entity, when predicting a single entity.To that end, it is important to note that the reported scores from QA-PICO in terms of Population, Intervention and Outcome are of models predicting each PICO label individually (only P or IC or O label), with different configurations and parameters each. The reported overall F1 score is of a single model handling all element predictions, which however has not reported Precision and Recall scores or individual element scores [15]. Our model, TransforMED, achieves the best overall performance, outperforming single entity model scores in Recall for Population and Intervention entities and in Precision and F1 for Outcome entities. What is more, BERT only achieves these results after 30 epochs of training, while EBM+ is trained for 20 epochs. Overall, our model requires less training to maintain the best overall performance due to transfer learning from the pretrained SciBERT model and the Transformer architecture.

TransforMED appears in first glance to have sub-optimal performance in terms of Recall for Outcome, and Precision for Population. By analyzing the predicting labels (Table A.8), we notice that in the case of Population our model achieves lower Precision because it identifies slightly longer spans of text than identified in the gold labels (Example 1). Displaying an opposite behaviour for Outcome entities, TransforMED manages to identify the key Outcome components while missing the non-medical terms (Example 3). This leads to sub-optimal Recall score, when the predicted labels are compared with the ground truth. For Intervention, we notice that the model suffers from the same issue as Outcome (Example 2). However, we also identify more entities, that should be identified within the dataset but are missed by the annotators which have a negative impact on both Precision and Recall (Example 3). This is attributed to disagreeing annotations between annotator groups in the EBM-NLP dataset, leading to terms such as "patients" not being consistently annotated in the gold labels as Population entities.

A comparative performance of the Argument Identification models is given in Table 2 in which we use $f1$ and F1 to denote the micro and macro averaged F1-scores respectively with C-F1 and E-F1 representing Claim and Evidence entitiy specific scores. Our model is compared with the best approaches in AMCT. AMCT-Sci refers to a finetuned version of SciBERT while AMCT-Bio refers to a finetuned version of BioBERT. Both models also use a Bidirectional Gated Recurrent Unit layer and a CRF. Our approach performs better overall, with average gains of 4.6% Claim identification and 2% in Evidence Identification across the three test

Table 2
Argument Identification models performance comparison.

Models	Neoplasm				Glaucoma				Mixed			
	$f1$	F1	C-F1	E-F1	$f1$	F1	C-F1	E-F1	$f1$	F1	C-F1	E-F1
AMCT-Sci	82.79	90.11	75.62	91.08	83.51	91.40	82.05	91.37	82.21	90.08	76.64	90.74
AMCT-Bio	79.88	89.22	73.60	89.91	83.33	91.95	83.55	91.17	80.84	90.93	77.93	91.25
TransforMED	**91.11**	**91.29**	**79.01**	**91.92**	**93.42**	**92.44**	**87.72**	**92.05**	**91.82**	**91.85**	**80.45**	**92.75**

Table 3
Argument Relation Classification model performance comparison in terms of F1 average.

Models	Neoplasm	Glaucoma	Mixed
TreeAM	37.00	45.00	40.00
ResidualAM	43.00	39.00	45.00
Roberta	67.00	66.00	69.00
AMCT-Sci	68.16	62.28	69.43
TransforMED	**69.96**	**69.72**	**71.82**

sets.

From the reported performance we also note an average 9% increase in terms of micro-f1 across all three datasets. This shows that our model is significantly better in identifying more tokens which are part of the argument, missing a lot less information compared to previous approaches. As a result, it provides the complete arguments as output which in turn has an impact on the performance of the ArgRC model.

The comparative performance of different models, using the same identified relations is showcased in Table 3. TransforMED performs consistently better than the best performing AMCT model, with a 3% increase in F1-score on the Mixed dataset and an 8% gain in F1 score on the Glaucoma dataset, which are both used only for evaluation. This is indicative of the ability of TransforMED to capture the latent characteristics of arguments and generalize past the given vocabulary in the Neoplasm dataset.

When comparing TransforMED with AMCT, at argument class level (Tables B.9, B.10 and B.11), it becomes apparent that AMCT's overall performance is not indicative for all labels. In the heavily under represented *Attacking* class, TransforMED achieves 5%, 22% and 5% increase in performance in Neoplasm, Glaucoma and Mixed test sets respectively, while maintaining a performance increase across both *Supporting* and *No Relation* classes.

4.1. Ablations

In order to evaluate the performance gains and quantify the effects of the proposed changes in methodology and architecture, we performed an ablation study on the various components of our model.

Table 4 exhibits the changes in performance on the EBM data. Specifically, we highlight the importance of using a Transformers encoder, in comparison to using a simple projection layer and in comparison to using a Bidirectional LSTM. From the results, it is obvious that the Transformer encoders perform significantly better than the baseline models, with all architectural changes contributing to the overall improvement. The addition of a Backward Encoder is also contributing to better identify the entities in all classes, boosting the overall performance. Although all previous models utilize a CRF to handle predictions [32,14,7], for reasons of completeness we also studied the effects of a CRF layer in this sequence tagging task. As CRF is shown to improve performance in all metrics on our baseline model, we use it in all further experiments.

To investigate the effects of the EBM model on Argument Identification and Argument Relation Classification, we also evaluate both models without the use of EBM annotations. Specifically, for the Argument Identification model, in Table 5, we highlight the effective changes from the different architectural decisions. We notice that, in comparison to the EBM model, the Backward Encoder resulted in worse performance as it added complexity to the model without providing any advantages in terms of better identifying arguments which attests to our previous design choice. Additionally, from the inclusion of EBM annotations the most important improvements come in the form of a 1–2% increase in Claim F1-score across all datasets, while Evidence only sees a 1–2% increase in F1 score in Neoplasm and Mixed, with no effects on the Glaucoma dataset. The removal of the imbalance towards the Before tag has made the most substantial contribution of 8–9% in micro F1 across all datasets. We also evaluated the use of a CRF layer to handle prediction, which has proven to have a positive impact in the performance of the ArgId model, similarly to our EBM model.

Table 6
Ablations Study on the Argument Relation Classification model.

Model	Neoplasm	Glaucoma	Mixed
TransforMED	**69.96**	**69.72**	**71.82**
-Encoder + Decoder	69.35	64.41	69.65
Encoder	69.19	67.57	70.32

Table 4
Ablations study on the EBM model.

Model	Population			Intervention			Outcome			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TransforMED	76.97	**91.24**	**83.49**	69.04	**79.24**	**73.79**	**84.38**	60.18	70.25	**79.32**	**80.33**	**79.82**
-Backward Encoder	75.54	90.49	82.34	67.46	78.15	72.41	84.00	60.77	70.52	78.61	79.56	78.35
-Encoders+BiLSMT	**79.20**	75.46	77.29	**72.18**	68.64	70.36	74.35	69.26	**71.30**	77.86	75.59	76.68
SciBert	78.88	74.73	76.75	71.20	68.69	69.92	73.19	69.36	71.22	76.97	75.69	76.32
SciBert -CRF	78.83	72.93	75.77	70.06	65.30	67.59	83.41	58.15	68.53	73.80	73.42	73.61

Table 5
Ablations study on the Argument Identification model.

Model	Neoplasm				Glaucoma				Mixed			
	*f*1	F1	C-F1	E-F1	*f*1	F1	C-F1	E-F1	*f*1	F1	C-F1	E-F1
TransforMED	**91.11**	**91.29**	**79.01**	**91.92**	**93.42**	**92.44**	87.72	**92.05**	**91.82**	**91.85**	80.45	**92.75**
+Backward Encoder	91.06	91.02	78.40	91.42	93.13	92.28	**86.73**	92.02	91.82	91.54	**81.38**	92.47
-EBM	91.08	91.17	78.92	91.42	93.12	92.24	85.62	92.03	91.46	91.33	80.03	92.21
-Encoders +BiGRU	90.34	90.29	78.32	91.14	93.11	92.44	85.63	92.05	91.17	91.11	79.03	92.16
-EBM-Encoders +BiGRU +IO Scheme change	90.99	90.84	76.84	91.20	92.82	92.32	84.92	92.10	91.13	90.86	79.67	91.63
-EBM-Encoders +BiGRU +IO Scheme change -CRF	90.51	90.30	77.23	91.66	91.67	91.93	82.04	91.87	90.50	90.42	77.63	91.50
AMCT	82.79	90.11	75.62	91.08	83.51	91.40	82.05	91.37	82.21	90.08	76.64	90.74

Table 7
Ablations Study on the Pooled Argument Relation Classification model.

Model	Neoplasm	Glaucoma	Mixed
AMCT + ArgId + PICO	69.75	67.83	70.10
AMCT + ArgId	68.79	62.75	69.59
AMCT	68.16	62.28	69.33

Going further, we investigate (Tables 6 and 7) the effects of better Argument Identification and the architectural changes in Argument Relation Classification performance as well as the advantages of using EBM annotations in different forms. We distinguish our results from using pooled representation of the arguments and EBM vectors (Table 7) and sequence representations (Table 6). In the pooled approach, we extend the baseline model (AMCT) by concatenating the pooled vector representations and forward through a feed-forward layer with a softmax activation. In the sequence approach, which we use for our model, we study the changes in performance from the architecture we introduced in Section 3.4. On the baseline model, there is an immediate gain in performance stemming from better argument identification (ArgId), that is augmented further by the use of EBM annotations. Most notably, the addition of EBM annotations contributes to a 4% increase on the Glaucoma dataset. On our approach, we always use the output of the ArgId model. With the addition of the Encoder model, there is a slim fluctuation in performance, with a slight increase in the Glaucoma dataset that is not part of the training data and a slight decrease in Neoplasm dataset, on which the model is trained on.

5. Discussion

In this work, we introduced a state-of-the-art Evidence-Based Medicine model, alongside a state-of-the-art Medical Argument Mining model that handles both Argument Identification and Argument Relation Classification. What is more, all these models work in pipeline for a better overall Medical Argument Mining performance.

Our EBM model, beats the previous state-of-the-art systems by a significant margin, when predicting all PICO entities at once. It also requires less training time as it is only trained for 3 epochs. While the QA-PICO is trained for 2 epochs, its performance can only be achieved by first pre-training to the SQUAD dataset and the EBM+ model requires 20 epochs of training.

Table A.8
Example predictions of EBM model.

Example 1	
Gold labels	In patients taking blood pressure or lipid-lowering treatment for the prevention of cardiovascular disease, text messaging improved medication adherence compared with no text messaging.
TransforMED	In patients taking blood pressure or lipid-lowering treatment for the prevention of cardiovascular disease, text messaging improved medication adherence compared with no text messaging.
Example 2	
Gold labels	Conclusions: The laparoscopic surgery in combination of QYJDR could effectively improve clinical symptoms of EMs patients of blood stasis and toxin accumulation syndrome, promote negative conversion of EMAb, lower serum CA125 levels, and elevate the clinical pregnancy rate.
TransforMED	Conclusions: The laparoscopic surgery in combination of QYJDR could effectively improve promote clinical symptoms of EMs patients of blood stasis and toxin accumulation syndrome, promote negative conversion of EMAb, lower serum CA125 levels, and elevate the clinical pregnancy rate.
Example 3	
Gold labels	Objective: The study objective was to identify factors associated with death and cardiac transplantation in infants undergoing the Norwood procedure and to determine differences in associations that might favor the modified Blalock-Taussig shunt or a right ventricle-to-pulmonary artery shunt.
TransforMED	Objective: The study objective was to identify factors associated with death and cardiac transplantation in infants undergoing the Norwood procedure and to determine differences in associations that might favor the modified Blalock-Taussig shunt or a right ventricle-to-pulmonary artery shunt.

Arguably, the QA-PICO model can achieve better performance when predicting individual PICO entities. However, in doing so there is an increased complexity that needs to be resolved in case of conflicting PICO annotations from the different models in order to use them in the pipeline. Furthermore, such a complexity is not resolved even when the model predicts all labels at once as it can return the same span of text as an answer to more than one questions. In a sequence tagging setup, such an issue is avoided with the use of a CRF layer. What is more, in order to use the QA-PICO in such a pipeline, a significant pre-processing and post-processing would be required to work in unison with the other models. Overall, our straightforward approach, which handles all PICO elements at once and does not need a lot of data transformations, is the best option for such a system.

With the inclusion of PICO annotations from our EBM model, we provide further semantic information about the source documents which benefited both components of the Argument Mining pipeline. In Argument Identification, we notice that 91% of the arguments in the dataset, regardless of their argument type, contain at least one PICO entity. As a result, the improvements in the performance from the PICO annotations were minimal but not negligible. In Argument Relation Classification however, better Argument Identification immediately translated into better Relation Classification performance. Moreover, the importance of providing semantic information to the model in the form of PICO annotations is clear, especially in the Glaucoma dataset which the model is not trained on.Our approach has proven exceptionally good to identifying *Attacking* argument relations, compared the previous models. Due to the scarcity and nature of *Attacking* arguments, it is crucial that such information is not missed. Given the very low number of occurrences in the AMCT corpus, which is indicative of the expected rate of attacking arguments in medical literature, TransforMED is a more appropriate Argument Mining system.

The addition of Transformer layers in both Argument Identification and Argument Relation Classification has helped the better generalization of the model. In our experiments, we observed a relatively equal performance in the Neoplasm split of the dataset, with the majority of the improvements happening in Glaucoma and Mixed data. Due to the nature of these improvements, we argue that this can be further extended to contribute gains to inter-document MAM.

6. Conclusions

In this paper we have introduced a state-of-the-art, EBM enhanced, end-to-end Transformer MAM pipeline. To that end, we have created a state-of-the-art EBM model, which is also an end-to-end Transformer model, that predicts all PICO entities at once and infused it with both the Argument Identification model and the Argument Relation Classification model.

Our approach leverages the latent semantic information from the EBM model and the novel Transformer architecture to achieve better performance in unseen train data of a different medical domain than the one the models are trained on. With this contribution we aim to improve MAM and medical document retrieval. Our EBM model can be used to enable fine-grained searched of available medical literature using PICO elements as search parameters. What is more, we provide a tool for medical practitioners to quickly extract key arguments from the literature that will grant them the ability to select all relevant literature faster and empower them to design evidence-informed treatment plans sooner.

CRediT authorship contribution statement

Nikolaos Stylianou: Conceptualization, Methodology,

Table B.9
Detailed results on Neoplasm test set.

Model	Attacking			Supporting			No Relation		
	P	R	F1	P	R	F1	P	R	F1
TransforMED	58.18	53.33	55.65	55.45	52.40	58.72	96.00	95.03	95.51
AMCT	51.85	46.67	49.12	55.61	64.90	59.99	96.11	94.80	95.45

Table B.10
Detailed results on Glaucoma test set.

Model	Attacking			Supporting			No Relation		
	P	R	F1	P	R	F1	P	R	F1
TransforMED	81.82	32.97	47.00	58.00	74.75	65.32	96.80	96.96	96.88
AMCT	100.00	14.14	24.78	56.15	76.34	64.71	96.75	93.77	95.24

Table B.11
Detailed results on Mix test set.

Model	Attacking			Supporting			No Relation		
	P	R	F1	P	R	F1	P	R	F1
TransforMED	76.47	54.17	63.41	50.26	66.22	57.14	96.31	93.53	94.90
AMCT	70.59	49.22	58.00	49.60	62.50	55.31	95.89	93.66	94.76

Investigation, Software, Writing - original draft. **Ioannis Vlahavas:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is co-financed by Greece and the European Union (European Social Fund - ESF) through the Operational Programme “Human Resources Development, Education and Lifelong Learning” in the context of the project “Strengthening Human Resources Research Potential via Doctorate Research” (MIS-5000432), implemented by the State Scholarships Foundation (IKY).

Appendix A. EBM prediction examples

We manually selected three characteristic examples from the EBM-NLP corpus and compare the gold labels with our model’s predictions. These example highlight both where our proposed model performs well and the cases in which it fails. We use the previously defined coloring scheme to identify entities in text, i.e. Red for Populations, Green for Interventions and Light Blue for Outcomes.

Table A.8

Appendix B. Detailed argument relation classification results

Detailed results in terms of Precision (P), Recall (R) and F1-score (F1) for the Argument Relation Classification model for *Attacking*, *Supporting* and *No Relation* arguments (see Tables B.9–B.11).

References

[1] H. Bastian, P. Glasziou, I. Chalmers, Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med. 7 (9) (2010) e1000326.

[2] X. Huang, J. Lin, D. Demner-Fushman, Evaluation of pico as a knowledge representation for clinical questions, in: AMIA annual symposium proceedings, vol. 2006, American Medical Informatics Association, 2006, p. 359.

[3] I.E. Allen, I. Olkin, Estimating time to conduct a meta-analysis from number of citations retrieved, Jama 282 (7) (1999) 634–635.

[4] L. Longo, L. Hederman, Argumentation theory for decision support in health-care: A comparison with machine learning, in: International Conference on Brain and Health Informatics, Springer, 2013, pp. 168–180.

[5] N. Green, E. Cabrio, S. Villata, A. Wyner, Argumentation for scientific claims in a biomedical research article, in: ArgNLP, 2014, pp. 21–25.

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

[7] T. Mayer, E. Cabrio, S. Villata, Transformer-based argument mining for healthcare applications, in: 24th European Conference on Artificial Intelligence (ECAI2020), 2020.

[8] H. Scells, G. Zuccon, B. Koopman, A. Deacon, L. Azzopardi, S. Geva, A test collection for evaluating retrieval of studies for inclusion in systematic reviews, in, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 1237–1240.

[9] F. Boudin, J.-Y. Nie, J.C. Bartlett, R. Grad, P. Pluye, M. Dawes, Combining classifiers for robust pico element detection, BMC Med. Informat. Decision Making 10 (1) (2010) 29.

[10] C. Gulden, M. Kirchner, C. Schüttler, M. Hinderer, M. Kampf, H.-U. Prokosch, D. Toddenroth, Extractive summarization of clinical trial descriptions, Int. J. Med. Informat. 129 (2019) 114–121.

[11] K.-C. Huang, C.C.-H. Liu, S.-S. Yang, F. Xiao, J.-M. Wong, C.-C. Liao, I.-J. Chiang, Classification of pico elements by text features systematically extracted from pubmed abstracts, in: 2011 IEEE International Conference on Granular Computing, IEEE, 2011, pp. 279–283.

[12] B.C. Wallace, J. Kuiper, A. Sharma, M. Zhu, I.J. Marshall, Extracting pico sentences from clinical trial reports using supervised distant supervision, J. Mach. Learn. Res. 17 (1) (2016) 4572–4596.

[13] B. Nye, J.J. Li, R. Patel, Y. Yang, I.J. Marshall, A. Nenkova, B.C. Wallace, A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, vol. 2018, NIH Public Access, 2018, p. 197.

[14] N. Stylianou, G. Razis, D.G. Goulis, I. Vlahavas, Ebm+: Advancing evidence-based medicine via two level automatic identification of populations, interventions, outcomes in medical literature, Artif. Intell. Med. 108 (2020) 101949.

[15] L. Schmidt, J. Weeds, J. Higgins, Data mining in clinical trial text: Transformers for classification and question answering tasks, arXiv preprint arXiv:2001.11268 (2020).

[16] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805 (2018).

[18] A. Peldszus, M. Stede, From argument diagrams to argumentation mining in texts: A survey, Int. J. Cogn. Informat. Nat. Intell. (IJCINI) 7 (1) (2013) 1–31.

[19] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. Simari, M. Thimm, S. Villata, Towards artificial argumentation, AI Magazine 38 (3) (2017) 25–36.

[20] E. Cabrio, S. Villata, Five years of argument mining: a data-driven analysis., in: IJCAI, vol. 18, 2018, pp. 5427–5433.

[21] J. Lawrence, C. Reed, Argument mining: A survey, Comput. Linguistics 45 (4) (2020) 765–818.

[22] C. Stab, I. Gurevych, Parsing argumentation structures in persuasive essays, Comput. Linguistics 43 (3) (2017) 619–659, https://doi.org/10.1162/COLI_a_00295, https://www.aclweb.org/anthology/J17-3005.

[23] R. Bar-Haim, I. Bhattacharya, F. Dinuzzo, A. Saha, N. Slonim, Stance classification of context-dependent claims, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 251–261. https://www.aclweb.org/anthology/E17-1024.

[24] M. Miwa, 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, vol. 1: Long Papers, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1105–1116. doi:10.18653/v1/P16-1105. https://www.aclweb.org/anthology/P16-1105.

[25] S. Eger, J. Daxenberger, I. Gurevych, Neural end-to-end learning for computational argumentation mining, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 11–22. doi:10.18653/v1/P17-1002. https://www.aclweb.org/anthology/P17-1002.

[26] P. Potash, A. Romanov, 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, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1364–1373, https://doi.org/10.18653/v1/D17-1143, https://www.aclweb.org/anthology/D17-1143.

[27] A. Galassi, M. Lippi, P. Torroni, Argumentative link prediction using residual networks and multi-objective learning, in: Proceedings of the 5th Workshop on Argument Mining, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1–10. doi:10.18653/v1/W18-5201. https://www.aclweb.org/anthology/W18-5201.

[28] J.S. Schuman, B. Horwitz, N.T. Choplin, R. David, D. Albracht, K. Chen, A 1-year study of brimonidine twice daily in glaucoma and ocular hypertension: a controlled, randomized, multicenter clinical trial, Arch. Ophthalmol. 115 (7) (1997) 847–852.

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

[30] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450 (2016).

[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[32] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, arXiv preprint arXiv:1603.01360 (2016).

[33] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).

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