End-to-End Argument Mining as Augmented Natural Language Generation

Anonymous ACL submission

Abstract

Argument Mining (AM) is a crucial aspect of computational argumentation, which deals with the identification and extraction of *Argumentative Components (ACs)* and their corresponding *Argumentative Relations (ARs)*. This work proposes *a unified end-to-end framework* based on a *generative paradigm*, in which the argumentative structures are framed into label-augmented text, called *Augmented Natural Language (ANL)*. Additionally, we explore the role of different types of markers in solving AM tasks. Through different marker-based fine-tuning strategies, we present an extensive study by integrating marker knowledge into our generative model. The proposed framework achieves competitive results to the state-of-the-art (SoTA) model and outperforms several baselines.

1 Introduction

Argument Mining (AM) (Lawrence and Reed, 2019) deals with the detection and classification of *Argumentative Components* (ACs) and their corresponding *Argumentative Relations* (ARs) from discourse dynamics. Figure 1 gives an illustrative example of ACs and ARs. AM is the fundamental process of computational argumentation (Dung, 1995) and is useful for debate analysis (Lawrence et al., 2017), automated essay scoring (Nguyen and Litman, 2018), customer review analysis (Chen et al., 2022), etc. AM task has been commonly subdivided into four key sub-tasks (Niculae et al., 2017): *(i) Component Segmentation* entails identifying fine-grained Argumentative Discourse Units (ADUs) (Peldszus and Stede, 2013), *(ii) Component Classification* involves categorizing ADUs into various ACs (Feng and Hirst, 2011), *(iii) Relation Identification* focuses on detecting argumentative relationships among two or more ADUs (Carstens and Toni, 2015), and *(iv) Relation Classification* deals with classifying these identified relations into different ARs (Jo et al., 2021).

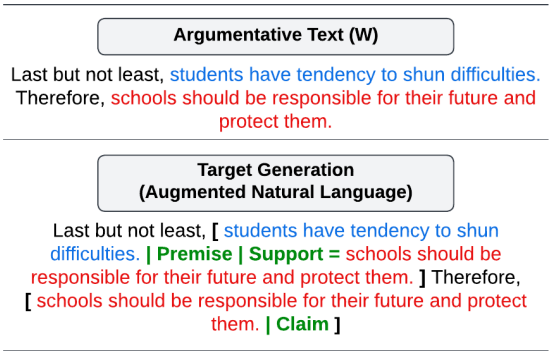

Figure 1: An overview of the proposed *generative end-to-end argument mining* task. Two ACs: *Claim* and *Premise* are marked in *red* and *blue* repectively. Their argumentative relation (AR) is *Support*. AC and AR labels in Generated Augmented Natural Language (ANL) is marked in *green*.

However, following the studies in Ye and Teufel (2021), Bao et al. (2022), Morio et al. (2022), we collectively term the initial pair of sub-tasks as Argument Component Extraction (ACE) and the subsequent pair as Argumentative Relation Classification (ARC). Consequently, end-to-end AM, referred to as ACRE in this work, involves jointly addressing both ACE and ARC tasks.

The primary challenge in any AM task lies in effectively handling the longer sequence length of ACs and their associated ARs. Defining boundaries for ACs is more intricate compared to tasks like Named Entity Recognition (NER) or Parts-of-Speech (POS) tagging, where the target text span consists of a few tokens only. Also, every AC has certain underlying contexts of argumentativeness and is related to another AC of the same context. Variations in argument representations across domains pose another challenge (Daxenberger et al., 2017). Given these complexities, we aim to explore an alternative end-to-end setup within the generative paradigm.

This work redefines the end-to-end AM task as a text-to-text generation problem by drawing

inspiration from successes in the generative approach for NER (Yan et al., 2021) and Joint Entity and Relation Extraction (Liu et al., 2022) tasks. The proposed framework takes plain text as an input and generates Augmented Natural Language (ANL) (Paolini et al., 2021) as an output with both ACs and ARs as label-augmented text (See Figure 1)(Athiwaratkun et al., 2020). The motivation behind choosing the ANL is its close resemblance to actual natural language. The model will interpret it as generating a different *"form"* of everyday language, which is relatively simpler compared to other forms of target generation.

Additionally, we explore the effectiveness of two types of markers for AM: *(a) Argumentative Markers* and *(b) Discourse Markers (DMs)* in our generative approach. Studies in (Gao et al., 2022; Clayton and Gaizauskas, 2022; Lawrence and Reed, 2015) indicate that *Argumentative Markers* strongly signal the presence of argumentative text. These are mostly a *span of tokens* like *"I strongly agree that", "But, I deny the point that", "However, this clearly proves that", etc.,* conveying argumentativeness of the discourse. In contrast, DMs are single-token connectives such as *"But", "And", "However", etc.,* representing the rhetorical structure of a language. But, in a broader sense, DMs are the subset of argumentative markers, as argumentative markers can sometimes also be single-token words depending on the context. Earlier, Kuribayashi et al. (2019) has created a list of markers by performing rule-based marker extraction from multiple datasets. Similar to them, this work also employs a simple rule-based extraction to prepare a list of argumentative markers from a single AM corpus. However, compared to Kuribayashi et al. (2019), we add an additional manual filtering step to remove non-markers containing topic information. To investigate influence of DMs, we consider *Discovery* corpus (Sileo et al., 2019), containing 174 DMs. Within the generative paradigm, the efficacy of both types of markers for end-to-end AM task has not been thoroughly investigated. To incorporate the knowledge of these markers into our proposed method, we introduce four distinct *fine-tuning strategies using markers* to familiarize the model with the marker distribution in the text. The resultant models from these strategies undergo additional fine-tuning for our proposed generation tasks.

Through extensive experimentation on ACRE task upon two structurally different standard benchmarks of AM literature, our proposed method achieves competitive results to the several important baselines in both benchmarks. In particular, compared to the only available current State-of-the-Art (SoTA) generative baseline (Bao et al., 2022), we achieve micro F1 improvement of up to 6.65 for the ACE task and up to 5.54 for the ARC task, affirming the effectiveness of our approach. The main contributions and findings of this paper are:

1. A generative task formulation for *End-to-End AM* along with *Component-only* and *Relation-only* variants to generate augmented natural language (ANL).

2. Investigation about contributions of different types of markers in solving AM tasks and associated four distinct marker-based fine-tuning strategies in the proposed formulation.

3. Surprisingly, being an exclusive feature of argumentative texts, the knowledge of markers doesn't contribute to the performance improvements of AM tasks in the generative paradigm.

4. The *Single-step* fine-tuned version shows superiority over *Two-step* versions in almost all AM tasks.

5. Analysis suggests that our proposed method can efficiently handle a diverse length of input text, spanning from shorter to longer paragraphs.

2 Related Work

2.1 Argument Mining

Most of the prior studies have focused on only a subset of the four AM sub-tasks. However, recent works (Eger et al., 2017; Morio and Fujita, 2018; Bao et al., 2022) are focusing more on joint formulation in an end-to-end manner. Persing and Ng (2016) followed a pipelined approach for ACE and ARC one after another and optimized the error propagation by performing joint inference using Integer Linear Programming (ILP). Eger et al. (2017) reformulated end-to-end AM task in four different ways: sequence tagging, dependency parsing, and multi-task tagging and relation extraction problem. Ye and Teufel (2021) proposed a biaffine network-based (Dozat and Manning, 2018) dependency parsing for end-to-end AM. Morio et al. (2022) identi-

fied the dataset scarcity in AM literature and proposed a cross-corpora multi-task formulation with a span-biaffine architecture. A span classifier generates BIO tags of spans and using average pooling, it generates span representations. Within the generative paradigm, Bao et al. (2022) framed it as a *text-to-sequence* generation task. In this generative framework, an array-like sequence is generated consisting of AC and AR types with the start and end indices of AC spans. However, to the best of our knowledge, no current literature has modeled end-to-end AM as a *text-to-text* generation task.

2.2 Markers

The literature delves into the significance of different types of markers across various NLP tasks (Pan et al., 2018; Nie et al., 2019; Sileo et al., 2020). Several studies have also shown that markers are crucial signals for ADUs (Stab and Gurevych, 2017; Kuribayashi et al., 2019; Dutta et al., 2022). Stab and Gurevych (2017) used markers as a lexical feature for classifying argument components with a feature-based multiclass classification. Later, Kuribayashi et al. (2019) extracted 1131 *argumentative markers* from different datasets to check the efficacy of AM tasks. They proposed an improved span representation utilizing the information of extracted markers. Dutta et al. (2022) also explored the contribution of markers for AM tasks in the Reddit social discussion thread. They extracted 69 Reddit-specific markers and performed *selective masked language modeling (sMLM)* by masking those markers for domain adaptation. Later, the resultant model was used for *Argument Component Identification*. A template-based approach was designed to predict the marker-like tokens in masked positions to predict the *Relation Type* between given components. However, within the current body of literature, the exploration of markers for AM tasks within a generative paradigm, employing an end-to-end AM framework, remains unexplored.

2.3 Augmented Natural Language (ANL) & Generative Paradigm

With the recent development of generative methods, most NLP tasks are being reformulated as generation problems. Generating label-augmented text *(a.k.a. ANL)* is one among various generative formulation strategies. It has been applied for several NLP tasks like NER (Athiwaratkun et al., 2020), sentiment analysis (Zhang et al., 2021), and relation extraction (Liu et al., 2022). Recently, Paolini et al. (2021) applied ANL to perform various structured prediction tasks like joint entity and relation extraction, event argument extraction, coreference resolution, by framing them as generative text-to-text translation problems. Despite ANL's growing popularity, its efficacy in argument mining to handle longer-span labels and longer-range relational dependencies remains unexplored. Our work aims to fill this research gap.

3 Task Formulation

We represent argumentative text as $W = w_1, w_2, w_3, ..., w_n$, where n is the total number of tokens in W. For a text-span $w_i, w_{i+1}, w_{i+2}, \ldots, w_j$ in W, we write it as $w_{i:j}$. We define a set of AC types as $T^c = \{t^c_1, t^c_2, t^c_3, ..., t^c_{n_c}\}$ and a set of AR types as $T^r = \{t^r_1, t^r_2, t^r_3, ..., t^r_{n_r}\}$, where n_c and n_r refer to a total number of possible AC and AR types respectively. Subsequent sections discuss different formulations of the AM task.

3.1 ACE task: Component-Only Variant

For any given argumentative text $w_{1:n}$, objective of the *ACE* is to extract a set of ACs as $C = \{C_i | C_i = (c_i, c^s_i, c^e_i)\}$, where C_i is the ith AC, $c_i \in T^c$, and c^s_i and c^e_i refer to the relative start and end token indices of c_i respective to W. Here, we generate the label-augmented text for ACs only such as, in Figure 1, for the head, AC is *[students have tendency to shun difficulties. | Premise]* and for tail AC is *[schools should be responsible for their future and protect them. | Claim]*.

3.2 ARC task: Relation-Only Variant

It is different in terms of the given input text and the target output text. Here, the input is also an ANL with an indication of ACs's boundaries without the corresponding type information. The target output is also an ANL of a specified format: *"Relationship between [ADU 1] and [ADU 2] is = Relation-Type"*. Notably, at any point in time, only 2 ADUs are taken to form an output. For example, in Figure 1, input ANL is *"Last but not least, [students have ... difficulties.]. Therefore, [schools should ... them.]"*. The corresponding output ANL is *"Relationship between [students have ... difficulties.] and [schools should ... them.] is = Support"*.

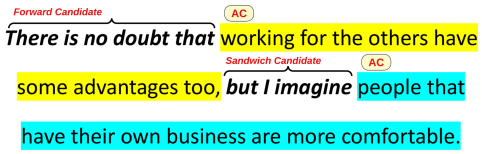

Figure 2: An example sentence from AAE corpus describing both ways of marker extraction. Here, extracted ***marker candidates*** are in bold italics, and ACs are highlighted with colors.

3.3 ACRE task: End-To-End Argument Mining

The proposed end-to-end formulation, jointly frame the *ACE* and *ARC* tasks in the following manner. We define ARs as $R = \{R_i | R_i = (r_i, r_i^h, r_i^t)\}$, where R_i is the ith AR corresponding to AR type $r_i \in T^r$, and r_i^h and r_i^t refer to the head and tail ACs respectively. r_i^h and r_i^t are connected with relation type r_i. If two components $C_p, C_q \in C$ are related with $R_k \in R$ as $r_k^h = C_p$ and $r_k^t = C_q$, then for the token spans of C_p i.e. $w_{c_p^s:c_p^e}$ and C_q i.e. $w_{c_q^s:c_q^e}$, the model will generate augmented labels as $[w_{c_p^s:c_p^e}|c_p|r_k = w_{c_q^s:c_q^e}]$ and $[w_{c_q^s:c_q^e}|c_q]$ respectively. Here $S = \{[,], =, |\}$ is a set of symbol tokens with "[", "]", " = ", "|" should be placed for the start of a component, end of a component, relation assignment and separation of labels respectively. The rest of the tokens in W will be rewritten as it is. We refer to this joint formulation as *ACRE*. Figure 1 illustrates the *ACRE* formulation.

4 Methodology

We consider T5-base (Raffel et al., 2020), an encoder-decoder model, as the base model. We propose two fine-tuned model variants: (i) *Single-step* fine-tuning involves directly fine-tuning the T5-base for the proposed generation task without any additional fine-tuning for markers. (ii) *Two-step* fine-tuning includes initial fine-tuning with marker strategies, followed by additional fine-tuning for the proposed generation task.

We describe the *Single-step* model variant in 4.1 and the marker extraction steps in 4.2. Subsequently, we discuss the *Two-step* fine-tuned model variants that are designed with different marker-based fine-tuning strategies. Finally, we will describe the decoding steps of generated ANL to get a cleaned text for evaluation.

4.1 Single-Step Model Variant

The single-step model variant is fine-tuned for directly generating ANL with plain text given as an input. This specific model variant is designed to be applicable across all versions, end-to-end, component-only, and relation-only, of the proposed method.

4.2 Argumentative Marker Extraction

An argumentative marker typically signals the beginning of an AC. However, not every marker is always followed by an AC, and an AC may not always be preceded by a marker. Considering this phenomenon, we extract two types of (See Figure 2) potential marker candidates from any argumentative text: *(i) Forward candidates:* by extracting tokens from the start of a sentence to the beginning of an AC and *(ii) Sandwich candidates:* by extracting tokens after the end of an AC until the start of another AC in the same sentence. Previously, Kuribayashi et al. (2019) also used a similar rule-based extraction strategy to prepare a marker set from AAE corpus (Stab and Gurevych, 2017) and Penn Discourse TreeBank 2.0 (Prasad et al., 2008). But, rule-based extraction also yielded some spans, which were topic-dependent. For example, "*(i) In spite of the importance of* ***sports activities***" or "*(ii) Nevertheless, opponents of* ***online-degrees*** *would argue that*" are having topic-specific tokens. These spans lacked generalizability across diverse topics, being tailored solely to their respective subjects. For that reason, different from Kuribayashi et al. (2019), we implemented an additional layer of manual filtering to eliminate these topic-dependent spans from the pool of extracted marker candidates. Initially, we identified 2925 marker candidates from a *single* AM corpus, AAE. After manual filtering (although we may miss a few) and removing duplicates, we refined the list to 1072 standard *argumentative markers*. These markers and their corresponding start and end token indexes were incorporated into the JSON-formatted AAE corpus.

4.3 Two-step Model Variant

This variant strategically divides the proposed generation task into two significant steps. The initial step leverages the extracted markers to execute marker-based fine-tuning. This involves implementing four distinct generative fine-tuning strategies, each utilizing varied input and output combinations (See Table 1). The objective is to ac-

Strategy	Input Sequence	Target Generation Sequence
A-MKT	Last but not least, students have ... difficulties.	**[Last but not least, | marker]** students have ... difficulties.
SM-MKT	*<extra_id_0> <extra_id_1> <extra_id_2> <extra_id_3> <extra_id_4>* students have ... difficulties.	*<extra_id_0>* **Last** *<extra_id_1>* **but** *<extra_id_2>* **not** *<extra_id_3>* **least** *<extra_id_4>* **,** *<extra_id_5>*
E-MKT	Last but not least, students have ... difficulties.	[**-1,-1,-1,-1,-1,**0,0,0,0,0,0,0]
N-MKT	Motivations for playing cricket are vastly different. It is a well crafted game.	Motivations for playing cricket are vastly different. **Truly,** it is a well crafted game.

Table 1: Description of various marker-based fine-tuning strategies. Example sentences for A-MKT, SM-MKT, and E-MKT are taken from *AAE* corpus. For N-MKT, the example is drawn from the *Discovery* corpus. Markers are marked in bold.

quaint the model with the nuanced representations of markers within the argumentative text. Notably, neither the *markers* from the test data nor the *test data* of the AAE corpus was used during this initial fine-tuning process.

In the second step, the models derived from the first step undergo additional fine-tuning specifically tailored for the proposed generation task. Below are the details of different marker-based fine-tuning strategies performed as the first step:

Argumentative Marker Knowledge Transfer (A-MKT): This strategy takes plain text input and fine-tunes the model to generate ANL where only *markers* are augmented. ACs and ARs are not augmented.

Span-Masked Marker Knowledge Transfer (SM-MKT): It is a self-supervised denoising fine-tuning strategy by *masking the span of markers*. We replace every token of the marker-span with sentinel tokens. Here, the target generation sequence is formed by concatenating the sentinel tokens and the corresponding marker tokens.

Marker Knowledge Transfer through Encoding (E-MKT): Unlike the above strategies, which are based upon text-to-text generation, it is a *text-to-sequence* generation strategy. Here, we generate the labels of marker tokens in terms of a numeric sequence of 0's and -1's, where, -1 and 0 are replacing the markers and non-markers of the input text respectively.

Normal Marker Knowledge Transfer (N-MKT): To check the effectiveness of single-token DMs over multiple-token markers, we propose this fine-tuning strategy. Using sentence pairs from *Discovery* corpus, we generate a target sequence like this: *(Sentence 1 + DM + Sentence 2)*, where the concatenated input is *(Sentence 1 + Sentence 2)*. Thus the model can achieve the capability of generating the probable *"connective"* between a pair of sentences.

4.4 ANL Decoding

This step is common for both single-step and two-step fine-tuned model variants. After generating the target ANL, we post-process the sequence by removing the symbol tokens to get a cleaned text. Following (Paolini et al., 2021), we employ dynamic programming for token-level alignment. Finally, for a comprehensive evaluation, AC and AR tuples are created, including their types and corresponding boundaries.

5 Experimental Setup

5.1 Dataset

We evaluate our proposed method with two AM benchmark datasets: `Argument Annotated Essay (AAE)` (Stab and Gurevych, 2017) and `Consumer Debt Collection Practices (CDCP)` (Niculae et al., 2017). We use the `Discovery` (Sileo et al., 2020) corpus for the sole purpose of the initial fine-tuning of N-MKT version only.

AAE: This dataset contains 402 student essays annotated at the segment (span) level. Every essay is divided into multiple paragraphs. A total of 1833 paragraphs are annotated with three AC types, $T^c = \{Claim, MajorClaim, Premise\}$, and two AR types, $T^r = \{Support, Attack\}$. AAE contains a large number of argumentative markers, with almost every AC beginning with one. We extract argumentative markers only from this dataset and transfer this knowledge to all experiments irrespective of dataset used.

CDCP: This dataset contains 731 user comments collected from the Consumer Financial Protection Bureau (CFPB) website. It is also annotated at the span level with five AC types, $T^c = \{Fact, Testimony, Reference, Policy, Value\}$, and two AR types, $T^r = \{Reason, Evidence\}$. CDCP mainly contains single-token DMs and only a few argumentative markers.

Discovery: Extracted from the *DepCC* web corpus (Panchenko et al., 2018), it features 1.74 mil-

lion pairs of adjacent sentences *(Sen1, Sen2)* with 174 DMs, consolidating 10k pairs per DM. All DMs occur at the beginning of *Sen2*.

5.2 Training Details

We use identical hyperparameter settings for CDCP and AAE benchmarks in the AREC task. We optimize them based on the best results from dev sets. Our setup includes Nvidia A100 GPU with a batch size of 8, AdamW optimizer (Loshchilov and Hutter, 2019), and a learning rate of 0.0005. Input/output sequence lengths are capped at 512 tokens. We run the end-to-end variant for 100 epochs for CDCP and 200 epochs for AAE. Results are averaged over 5 test runs, each taking around 6 hours of GPU time. During inference, we employ beam search with a length of 8. The same set of hyperparameters are used for Component-only and Relation-only variants. For AAE Relation-only variants, we run the model for only 50 epochs. We use the following libraries: (i) TANL framework[1] (Paolini et al., 2021), and (ii) HuggingFace's Transformers[2] (Wolf et al., 2019).

Marker Fine-Tuning: A-MKT and SM-MKT are trained for 200 epochs with batch size of 16, and 0.0005 as learning rate. N-MKT is trained for 5 epochs with batch size of 32, and learning rate of 0.0002. E-MKT is trained for 200 epochs with batch size of 4, and learning rate of 0.0005. In each case, AdamW optimizer is used with sequence length of 512 tokens except for N-MKT, where 128 tokens are considered.

5.3 Baselines

We take several important SoTA baselines to investigate the efficacy of our end-to-end AM formulation. For the AAE benchmark, we consider the following baselines. **ILP** (Persing and Ng, 2016): Rich feature based approach to perform *joint inference* over the AM sub-tasks optimized by *Integer Linear Programming (ILP)*. **BLCC** (Eger et al., 2017): Based upon *Bi-LSTM-CNN-CRF (BLCC)* to formulate this task as a sequence tagging problem. **LSTM-ER** (Eger et al., 2017): An adapted version of an end-to-end relation extraction model with sequential LSTM (Miwa and Bansal, 2016). **LSTM-Parser** (Eger et al., 2017): A dependency parsing approach built on stacked LSTM (Dyer et al., 2015). **BiPAM** (Ye and Teufel, 2021): Another dependency parsing approach with customized biaffine operation based upon BERT-base (Devlin et al., 2019). **BiPAM-syn** (Ye and Teufel, 2021): An enhanced version of BiPAM with the inclusion of *syntactic* information. **BART-B** (Bao et al., 2022): A generative approach to *text-to-sequence generation* with Bidirectional and Auto-Regressive Transformer (BART) (Lewis et al., 2020). **RPE-CPM** (Bao et al., 2022): An enhanced version of BART-B with *reconstructed positional encoding (RPE) and constrained pointer mechanism (CPM)*.

BiPAM, **BART-B**, **RPE-CPM**, and **CPM-only** (without RPE) are used as baselines for the CDCP benchmark. We compare with the best results obtained by Morio et al. (2022) for both the benchmarks.

Corpus	Method	C-F1	R-F1
AAE	LSTM-Parser	58.86	35.63
	ILP	62.61	34.74
	BLCC	66.69	39.83
	LSTM-ER	70.83	45.52
	BiPAM	72.90	45.90
	BiPAM-Syn	73.50	46.40
	BART-B	73.61	47.93
	RPE-CPM	75.94	50.08
	$T5_{Single-step}$	75.93	50.56
	Morio-MT-All	**75.66**	**55.17**
CDCP	BiPAM	41.15	10.34
	BART-B	56.15	13.76
	RPE-CPM	57.72	16.57
	CPM-only	58.13	15.11
	$T5_{Single-step}$	64.78	20.65
	Morio-MT-All	**68.81**	**33.74**

Table 2: Experiment results of ACRE task with the comparable baselines. Best scores are marked in bold. Here, C-F1 is Component F1, and R-F1 is Relation F1.

5.4 Evaluation Metrics

Following (Eger et al., 2017; Bao et al., 2022), we evaluate the results with *micro-F1* score for both ACE and ACRE tasks, where an exact match with the gold label is considered as a true label. But, for the ARC task in *Relation-only* variant, as we are already giving the ADU spans (without component types) in both input and output, we only calculate the micro-F1 score for the generated AR labels, namely "*Rel-F1*". Following (Paolini et al., 2021), if the generated AR labels are not in the pre-defined set, we determine the correct label by considering the log-likelihood of all pre-defined class scores.

6 Results and Discussion

Table 2 compares the ACRE task performance of the proposed model with the baseline models. Among the proposed model variants, we report the

[1] https://github.com/amazon-science/tanl
[2] https://github.com/huggingface

Method	Model	AAE	CDCP
Comp-Only **(C-F1)**	$T5_{Single-step}$	**70.37**	65.59
	$T5_{E-MKT}$	66.48	57.51
	$T5_{A-MKT}$	67.44	60.66
	$T5_{SM-MKT}$	68.71	62.26
	$T5_{N-MKT}$	69.84	**65.95**
Rel-only **(Rel-F1)**	$T5_{Single-step}$	96.27	97.13
	$T5_{E-MKT}$	**96.77**	96.77
	$T5_{A-MKT}$	96.34	97.35
	$T5_{SM-MKT}$	96.32	**97.42**
	$T5_{N-MKT}$	96.47	97.27

Table 3: Performance comparison of different *(Single-step and Two-step)* models for *Component-only* and *Relation-only* variants.

Corpus	Model	C-F1	R-F1
AAE	$T5_{Single-step}$	75.93±0.60	**50.56±1.13**
	$T5_{E-MKT}$	73.06±0.51	45.89±1.75
	$T5_{A-MKT}$	74.22±0.77	48.01±1.15
	$T5_{SM-MKT}$	75.91±1.00	49.08±1.44
	$T5_{N-MKT}$	**76.45±0.80**	49.91±1.01
CDCP	$T5_{Single-step}$	64.78±0.52	**20.65±0.80**
	$T5_{E-MKT}$	54.82±0.49	8.02±1.04
	$T5_{A-MKT}$	59.85±0.24	13.04±0.80
	$T5_{SM-MKT}$	62.63±0.44	16.40±1.70
	$T5_{N-MKT}$	**64.90±0.68**	19.90±1.19

Table 4: Performance comparison of *Single-step* vs *Two-step* model variants for ACRE task.

results of the highest average of C-F1 and R-F1 over 5 runs in this table. Interestingly, the *Single-step* model variants outperform all the *Two-step* variants in both benchmarks. This underscores the effectiveness of formulating end-to-end AM in a generative approach with ANL, showcasing its superiority over other methods even without the knowledge of markers. In both benchmarks, our proposed method outperforms several significant baselines. In particular, as compared to the only generative baseline by Bao et al. (2022), the results are competitive in the AAE benchmark. In the CDCP benchmark, our approach outperforms them by 6.65% in C-F1 and 5.54% in R-F1. However, the model-variants proposed by Morio et al. (2022) remains the best performing models. Unlike most baselines that rely on explicit feature information in addition to raw text, our method performs decently using only plain text as inputs, without any extra information. Bao et al. (2022) modifies the model architecture, while our approach surpasses it without any architectural changes to the vanilla T5-base.

The ***Component-Only*** variant shows (See Table 3) 6.08% decrease in F1 scores on the best performing model (*Single-step*) for the ACE task in the AAE benchmark as compared to the ACRE task variant. But, surprisingly, in the CDCP benchmark, the performance of the ACE task shows a 1.05% increase in F1 over the best-performing model (*N-MKT*) than the ACRE task variant. This signifies that in the AAE benchmark, ACs and ARs benefit from mutual feature information, suggesting synergy in an end-to-end setup. Whereas, in CDCP, even without relational information, the ACE task performance is not dropped and is comparable with the end-to-end setup.

The ***Relation-Only*** variant (See Table 3) proves highly effective in predicting correct relations between provided AC spans. All model variants exhibit comparable F1 scores, consistently around 96% and 97% in AAE and CDCP benchmarks respectively. This highlights the capability of the T5-base to detect accurate relations when presented with well-defined text spans and proper input/output templates.

6.1 Effect of Fine-tuning Strategies with Markers

Table 4 shows the ACRE task performance comparison of *Single-step* and *Two-step* model variants for both benchmarks. The result suggests that the model does not gain much from transferred marker knowledge for the ARC task. However, for the ACE task, marker knowledge proves beneficial for both benchmarks, with the *Two-step* variant N-MKT yielding the best results. This is interesting and indicates that the model benefits more from the knowledge of *single-token* DMs than *span-of-token* argumentative markers for ACE task in an end-to-end setup. Except for N-MKT, ACRE task performance of *Two-step* variants drops as compared to the *Single-step* variant in both benchmarks. Notably, even the source corpus (AAE) of argumentative markers doesn't prove to be beneficial for the ACRE task by the marker-based fine-tuning.

Among all *Two-step* variants, a similar phenomenon is observed for the ACE task in ***Component-only*** variant, where in both benchmarks, *N-MKT* is proven superior as compared to other *Two-step* variants (See Table 3). The ***Relation-only*** version performs equally well for relation identification in both *Single-step* and *Two-step* variants, as the AC spans are already provided. No significant variations in results are observed across different fine-tuning strategies with markers.

These counter-intuitive results led to the following two important research questions: (i) Why do

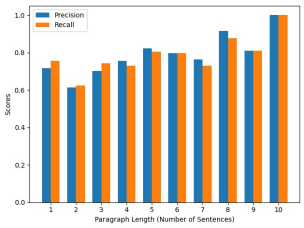

Figure 3: Performance of *End-to-End variant on ACRE task for extraction of ACs* in terms of precision and recall.

Corpus	Model	IT	IC	IF
AAE	$T5_{Single-step}$	2.95	4.51	1.11
	$T5_{E-MKT}$	5.45	6.62	3.39
	$T5_{A-MKT}$	2.61	5.4	1.05
	$T5_{SM-MKT}$	3.06	5.23	1.39
	$T5_{N-MKT}$	**2.39**	**4.9**	**0.8**
CDCP	$T5_{Single-step}$	11.33	**4.93**	7.6
	$T5_{E-MKT}$	27.33	15.2	**5.73**
	$T5_{A-MKT}$	13.6	8.93	7.33
	$T5_{SM-MKT}$	12.53	6.4	7.2
	$T5_{N-MKT}$	**9.6**	**4.93**	7.06

Table 5: Error analysis of different model variants of ACRE task. **IT, IC,** and **IF** refer to *Invalid Token, Invalid Component,* and *Invalid Format* respectively.

DMs prove to be effective over argumentative markers? (ii) Despite being an exclusive feature of the argumentative text, why do they fail to contribute to the performance improvement?

Firstly, the resultant model from different fine-tuning strategies struggles to grasp the nuanced context of span-of-markers due to its varied length. It can range from a single token to up to 20 tokens depending upon the context. Whereas, being only a single-token length, the knowledge of DMs is well-generalized by the model. This underscores the need for a more sophisticated fine-tuning approach to effectively incorporate the information conveyed by longer argumentative markers into the model. *Secondly,* the initial fine-tuning of the N-MKT version was performed on a non-argumentative large-sized dataset as compared to the AAE dataset, upon which the other strategies (A-MKT, SM-MKT, E-MKT) are built. This way, N-MKT learns the cross-domain marker knowledge representation when it is again fine-tuned in an argumentative dataset for the target task. *Lastly,* among A-MKT, SM-MKT, and E-MKT, the knowledge of the relative position of markers doesn't seem beneficial as the E-MKT performs poorly in almost all variants. For the A-MKT version, as the tasks are similar in both steps of fine-tuning, the model is likely to suffer from the *catastrophic forgetting* (Luo et al., 2023). As a result, the gained knowledge of markers is partially forgotten after the target task fine-tuning. But in case of SM-MKT, as the tasks are different in both steps of fine-tuning, the effect of catastrophic forgetting is minimized. Hence, the performance is better as compared to E-MKT and A-MKT.

6.2 Performance Evaluation based on Input Text Length

We assess the performance of our best-performing model (*Single-step*) on the ACRE task for extraction of ACs, on the AAE benchmark based on the number of input text sentences. Figure 3 illustrates that the performance in terms of both precision and recall does not deteriorate with the increasing input length. This underscores the efficacy of our method in effectively handling paragraphs of longer lengths.

6.3 Error Analysis

The generative methods sometimes produce invalid outputs as the generation is uncontrollable. We identified the following three major types of erroneous generation: *(i) Invalid Token:* The generated ANL consists of some out-of-vocabulary tokens or out-of-context text spans (*Hallucinations*). *(ii) Invalid Format:* The invalid ANL format includes mismatched brackets, symbols, or corrupted text. *(iii) Invalid Component:* The tail component connected with the relation in ANL is invalid if it is a span of text from the non-component regions. Results in Table 5 indicate that N-MKT is superior in terms of generating error-free ANL. E-MKT generates more erroneous ANL than others. Importantly, erroneous generations are discarded as negative results without undergoing any additional post-processing.

7 Conclusion

In this work, we reformulate the end-to-end AM task in a generative paradigm. We focus on the effectiveness of utilizing ANL as a target generation text for producing argumentative structures. Using the extracted markers from the AAE corpus and DMs from the Discovery corpus, we investigate the effectiveness of different types of markers in the proposed formulation. Additionally, we compare different formulations of AM sub-tasks to evaluate the need for an end-to-end approach. Our extensive empirical experiments demonstrate the efficiency of our generative approach for end-to-end AM task.

8 Limitations and Future Scope

There are certain limitations of this study. Firstly, in all our experiments, we consistently use a single ANL format, which produces commendable results. But, there may be some other ANL formats, that could potentially enhance performance even further. Secondly, our experiments are based on a single-corpus setting. It is also worth exploring, how this generative method performs in a multi-corpora setup. Third, we put our efforts into invoking the markers' knowledge using four distinct *Two-step* fine-tuning strategies, but got counter-intuitive outcomes. Thus, there is merit in investigating superior strategies capable of enhancing performance within a generative paradigm, utilizing nuanced marker knowledge. Fourth, our proposed method uses standard AM corpus for fine-tuning, which is not noisy. In real-world scenarios, however, data tends to be noisy. Hence, evaluating our system's performance within a noisy environment presents an intriguing avenue of inquiry. Fifth, our current methodology adopts the default input sequence length of T5-base, set at 512 tokens for both input and output sequences. But our current ANL output sequence contains redundant texts sometimes; such as, if multiple premises support a single claim, then the same claim is repeated multiple times with each unique premise over and again, which is eating the limit of 512 token lengths. It demands exploration of some other ANL formats, which are shorter in length and reduce the redundant repeating text in the output target generation. Lastly, we use the potential of T5-base in all our experiments. It will be interesting to see how other encoder-decoder models (e.g. BART, LLaMA) perform with this task setup.

References

Ben Athiwaratkun, Cicero Nogueira dos Santos, Jason Krone, and Bing Xiang. 2020. Augmented Natural Language for Generative Sequence Labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 375–385.

Jianzhu Bao, Yuhang He, Yang Sun, Bin Liang, Jiachen Du, Bing Qin, Min Yang, and Ruifeng Xu. 2022. A Generative Model for End-to-End Argument Mining with Reconstructed Positional Encoding and Constrained Pointer Mechanism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10437–10449.

Lucas Carstens and Francesca Toni. 2015. Towards relation based Argumentation Mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34.

Zaiqian Chen, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo, and Joonsuk Park. 2022. Argument mining for review helpfulness prediction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8922.

Jonathan Clayton and Rob Gaizauskas. 2022. Predicting the presence of reasoning markers in argumentative text. In *Proceedings of the 9th Workshop on Argument Mining*, pages 137–142.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Conference on Empirical Methods in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Subhabrata Dutta, Jeevesh Juneja, Dipankar Das, and Tanmoy Chakraborty. 2022. Can unsupervised knowledge transfer from social discussions help argument mining? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7774–7786.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural End-to-End Learning for Computational Argumentation Mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996.

Yingqiang Gao, Nianlong Gu, Jessica Lam, and Richard H.R. Hahnloser. 2022. Do Discourse Indicators Reflect the Main Arguments in Scientific Papers? In *Proceedings of the 9th Workshop on Argument Mining*, pages 34–50.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739.

Tatsuki Kuribayashi, Hiroki Ouchi, Naoya Inoue, Paul Reisert, Toshinori Miyoshi, Jun Suzuki, and Kentaro Inui. 2019. An Empirical Study of Span Representations in Argumentation Structure Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4691–4698.

John Lawrence, Joonsuk Park, Katarzyna Budzynska, Claire Cardie, Barbara Konat, and Chris Reed. 2017. Using argumentative structure to interpret debates in online deliberative democracy and erulemaking. *ACM Transactions on Internet Technology (TOIT)*, 17:1 – 22.

John Lawrence and Chris Reed. 2015. Combining Argument Mining Techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Tianyu Liu, Yuchen Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. In *Conference on Empirical Methods in Natural Language Processing*, pages 993–1005.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yuechen Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *ArXiv*, abs/2308.08747.

Makoto Miwa and Mohit Bansal. 2016. 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)*, pages 1105–1116.

Gaku Morio and Katsuhide Fujita. 2018. End-to-End Argument Mining for Discussion Threads Based on Parallel Constrained Pointer Architecture. *Proceedings of the 5th Workshop on Argument Mining*, pages 11–21. Conference Name: Proceedings of the 5th Workshop on Argument Mining Place: Brussels, Belgium Publisher: Association for Computational Linguistics.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. 2022. End-to-end Argument Mining with Cross-corpora Multi-task Learning. *Transactions of the Association for Computational Linguistics*, 10:639–658.

Huy V. Nguyen and Diane J. Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *AAAI Conference on Artificial Intelligence*.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument Mining with Structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Boyuan Pan, Yazheng Yang, Zhou Zhao, Yueting Zhuang, Deng Cai, and Xiaofei He. 2018. Discourse marker augmented network with reinforcement learning for natural language inference. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 989–999.

Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone P. Ponzetto, and Chris Biemann. 2018. Building a web-scale dependency-parsed corpus from CommonCrawl. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, 2021*.

Andreas Peldszus and Manfred Stede. 2013. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204.

Isaac Persing and Vincent Ng. 2016. 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*, pages 1384–1394.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. 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)*, pages 3477–3486.

Damien Sileo, Tim Van de Cruys, Camille Pradel, and Philippe Muller. 2020. DiscSense: Automated semantic analysis of discourse markers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 991–999.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A Unified Generative Framework for Various NER Subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822.

Yuxiao Ye and Simone Teufel. 2021. End-to-end argument mining as biaffine dependency parsing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 669–678.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards Generative Aspect-Based Sentiment Analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510.

A Differences between valid Markers and Topic-dependent Non-Markers

In this section, we show the potential differences between valid markers and topic-dependent non-markers (See Table 6). We use the instance examples which are both common in our marker set and the set provided by Kuribayashi et al. (2019).

Valid Markers
(i) Nevertheless, I believe that
(ii) Another supporting reason is that
(iii) People who hold different opinion may argue that
(iv) I strongly disagree with this affirmation because I believe
(v) In conclusion, the above stated reasons clearly outweigh the fact that
Manually Filtered-Out Topic-Dependent Marker Candidates
(i) In spite of the importance of ***sports activities***
(ii) Moreover, the proponents of ***globalization*** *idea point out*
(iii) Nevertheless, opponents of ***online-degrees*** *would argue that*
(iv) The ***official term*** *of it is named* ***"technological unemployment"***
(v) However, as the ***society*** *grows,* ***human rights*** *become more highly respected*

Table 6: Example of extracted valid markers and manually filtered-out topic-dependent marker candidates (non-markers) from AAE corpus, which are common in our extracted list of marker candidates and the marker list provided by Kuribayashi et al. (2019). Topic information is marked in bold.

B Precision and Recall scores in terms of the number of ADUs

We compared the performance of the ACRE task in terms of the number of ADUs with their precision and recall scores with the best performing *Single-step* model with AAE benchmark. Figure 4 shows that the performance does not degrade with the increasing number of ADUs present in a paragraph.

C Extracted marker examples corresponding to the paragraphs

We present the illustrations of extracted argumentative valid markers (See Table 7) from the AAE corpus. Notably, these examples are presented after manual-filtering steps.

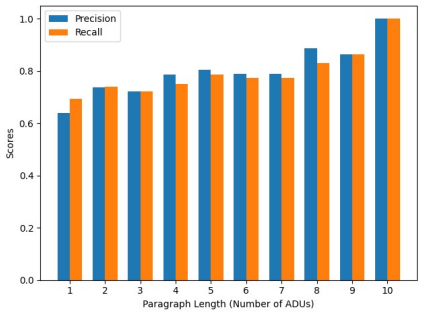

Figure 4: Performance of *End-to-End variant on ACRE task for extraction of ACs* in terms of precision and recall with number of ADUs present in a paragraph.

D Different types of errors and their example instances

In Table 8, we describe different types of ANL-related errors that our model experiences during the generation task.

SN.	Paragraph	Extracted Argumentative Markers
1	In conclusion, I strongly agree that we should give more priority to health education and preventative measures than to treatment. However, reasonable attentions should be paid to treatment so that our citizens are always looked after with the best services.	1. In conclusion, I strongly agree that (0,6) 2. However, (22,23)
2	First and foremost reason is that pursuit of nuclear technology one way or the other leads towards atomic weapons.	1. First and foremost reason is that (0,5)
3	First of all, I do support the idea that advertising alcohol, cigarettes, goods, and services with adult content should be prohibited because these kinds of ads will have a negative effect on our children. Fortunately, some countries take this issue seriously, and advertising alcohol, cigarettes, and materials with adult content is banned in those countries.	1. First of all, I do support the idea that (0,9) 2. Fortunately, (57,58)
4	All in a nutshell, workers over 50 have proven themselves subordinate to the following generation in all aspects. However, in my point of view, it is still of great importance to remain the harmonious coexistence of both sides for the smooth sailing of the company.	1. All in a nutshell, (0,4) 2. However, in my point of view, (20,27)
5	Firstly, by having CCTV cameras at the workplace, crimes such as robbery can be trimmed down. This is because when a shop is attacked by thieves, the clips taken by these cameras serve as a source of evidence to help the authorities trace the criminals. To illustrate this further, in Malaysia, policemen used the videos captured by the CCTV cameras to trace and successfully catch a group of robbers who broke into a Seven Eleven shop during the midnight. Not only that, as the employees know that they are under constant surveillance, they will be less likely to steal from the shops. Thus, it is clear that CCTVs must be put in all workplaces so that crime rates can be minimized.	1. To illustrate this further, (49,53) 2. Not only that, (86,89) 3. Thus, it is clear that (110,115)
6	To sum up, if all those factors are contemplated, it is obviously found that there are many solid reasons to support my viewpoint. So, from what has been discussed above, I strongly affirm the conclusion that people will spend less time in cooking food.	1. So, from what have been discussed above, I strongly affirm the conclusion that (26,40)

Table 7: Examples of extracted argumentative markers from the AAE dataset with corresponding *(Start Index, End Index)* relative to the paragraph.

Error Type	Ground Truth ANL	Predicted ANL	Error Description
Invalid Token	In summary, [spending budget for space exploring is not the waste budget | MajorClaim], because [it may helps to save civilization of humanity | Claim | Support = spending budget for space exploring is not the waste budget]. Although exploring space used for discovery of other planets, nowadays most of governments have used many rockets for military purposes and **therefore** the purposes of exploring space has altered.	In summary, [spending budget for space exploring is not the waste budget | MajorClaim], because [it may helps to save civilization of humanity | Claim | Support = spending budget for space exploring is not the waste budget]. Although [exploring space used for discovery of other planets | Claim | Attack = spending budget for space exploring is not the waste budget], [nowadays most of governments have used many rockets for military purposes | Premise | Support = the purposes of exploring space has altered]. **Therefore** [the purposes of exploring space has altered | MajorClaim].	Here, the ground truth ANL token is “therefore", but the predicted ANL token is “Therefore". So, casing mismatch is there. We consider casing mismatch also as an invalid token.
Invalid Format	In short, although both sides have their equally valid arguments, I am inclined to think that although [Internet can be a useful tool for education | Claim | Attack = at no time will the unique role of human teachers be replaced], [at no time will the unique role of human teachers be replaced | MajorClaim].	In short, although *[* both sides have their equally valid arguments | Claim | Attack = I am inclined to think that although *[* Internet can be a useful tool for education, at no time will the unique role of human teachers be replaced | MajorClaim], I am inclined to think that although *[* Internet can be a useful tool for education, at no time will the unique role of human teachers be replaced | Claim | Attack = I am inclined to think that although [Internet can be a useful tool for education, at no time will the unique role of human teachers be replaced | Premise | Attack = Internet can be a useful tool for education, at no time will the unique role of human teachers be replaced].	Here, in the predicted ANL, the brackets marked in italics are mismatched.
Invalid Component	In short, although both sides have their equally valid arguments, **I am inclined to think that although** [Internet can be a useful tool for education | Claim | Attack = at no time will the unique role of human teachers be replaced], [at no time will the unique role of human teachers be replaced | MajorClaim].	In short, although [both sides have their equally valid arguments | Claim | Attack = **I am inclined to think that although** [Internet can be a useful tool for education, at no time will the unique role of human teachers be replaced | MajorClaim], I am inclined to think that although [Internet can be a useful tool for education, at no time will the unique role of human teachers be replaced | Claim | Attack = I am inclined to think that although [Internet can be a useful tool for education, at no time will the unique role of human teachers be replaced | Premise | Attack = Internet can be a useful tool for education, at no time will the unique role of human teachers be replaced].	Here, although in the predicted ANL, the span marked in bold is not generated as a valid component, still it is related to some “claim". So, it is an invalid component error. Notably, in the ground truth ANL, it is not a valid span.

Table 8: Instances of different types of errors are highlighted in bold.