Mining Argument Components in Essays at Different Levels

Roberto Demaria, Davide Colla, Matteo Delsanto, Enrico Mensa, Enrico Pasini and Daniele P. Radicioni

July 9, 2023

Mining Argument Components in Essays at Different Levels

Roberto Demaria$^{1[0009-0000-1555-1698]}$, Davide Colla$^{3[0000-0002-9999-0109]}$,
Matteo Delsanto$^{1[0000-0002-7582-9185]}$, Enrico Mensa$^{1[0000-0001-7743-4999]}$,
Enrico Pasini$^{2[0000-0002-4525-187X]}$, and Daniele P. Radicioni$^{1[0000-0003-0443-7720]}$

[1] Dipartimento di Informatica, Università degli Studi di Torino
[2] Istituto per il Lessico Intellettuale Europeo, ILIESI/CNR – Roma
[3] Dipartimento di Studi Storici, Università degli Studi di Torino

Abstract. The research of arguments in student essays has long been the subject of automatic approaches to argument mining. The task has been mostly modeled as a sequence tagging problem, where the text is either analyzed in its entirety or split into smaller homogeneous units, such as sentences or paragraphs. However, previous research has highlighted how the various essay sections may fulfill different functions, and thereby how the position of specific argument components obeys precise structural dependency criteria. Based on such underpinning we propose an approach that exploits such structural information: in this work we present a hybrid training approach that takes into account the specific structural components of the essays, in order to be able to mine different types of argument components at different levels. Our hybrid approach achieves an improvement over essay-level and paragraph-level training, in particular in the extraction of some specific argument components.

Keywords: Argument Mining · Argument Component Classification · Persuasive essays · Natural Language Processing · Transformers · Machine Learning.

1 Introduction

Argumentation is a linguistic realization of the human reasoning [15], and is employed to justify a viewpoint about a controversial issue [35]. One fundamental problem with the definition and formal description of argumentation and argumentative paths is that there is no agreement among theorists about a universal an uniquely accepted theory. As Van Eemeren et al. [13] state in their recent survey of the field:

> *As yet, there is no unitary theory of argumentation that encompasses the logical, dialectical, and rhetorical dimensions of argumentation and is universally accepted. The current state of the art in argumentation theory is characterized by the coexistence of a variety of theoretical perspectives and approaches, which differ considerably from each other in conceptualization, scope, and theoretical refinement.*

While the lack of theories covering all possible argumentative structures affects computational applications, where we observe a gap between theoretical and computational models, a taxonomy of argumentation models has been proposed addressing three different categories —micro-level models (or monological models), macro-level models (or dialogical models), and rhetorical models—, intended to formalize conversations such as discussions, debates, or negotiations by introducing rules on how arguments interact [7].

Argument Mining (AM) is a specific area of Natural Language Processing aimed at mining arguments from natural language texts [21]. AM initially started with the aim at analyzing structured texts in the legal domain, and at a later time it was extended to more heterogeneous and unstructured sources from the web. When dealing with well-structured texts, we are in a paradigm called *closed-domain discourse-level AM* [33]. This task has been typically arranged into three main sub tasks [31]:

- *Argument Identification*, that is, the recognition and localization of arguments within a text;
- *Argument Classification*, which is concerned with the categorization of arguments and their argument components;
- *Structure Identification*, targeted to the reconstruction of the relations connecting the arguments.

All such tasks have been extensively studied in AM, and state-of-the-art approaches adopt supervised learning and transformer-based architectures such as BERT or LONGFORMER [11,6].

Among the many kinds of structured texts, we single out persuasive essays. A persuasive (or argumentative) essay is a text written to argue about a controversial topic while following a particular structure. This makes such kind of texts an excellent playground to test AM tasks [8]. An open issue, in this setting, is whether to treat arguments as a closed (or at least discrete) system with local fragments of text influenced by an isolated set of considerations, or to consider them as an open system within a broader spectrum of influence [28]. In analyzing student essays (a class of persuasive essays) this has resulted in considering a paragraph-level or an essay-level perspective when approaching the learning phase. In some cases the former perspective turned out to be preferable: for example, analyzing student essays at paragraph-level lead to better AM performances than essay-level [12], but contrasting evidence is also reported in literature [20], and there seems to be some intertwining with the model used.

In previous work we reported about differences stemming from learning at such different levels [10]; more specifically, we showed that, when employing a BERT-based model, the essay-level approach is preferable in order to deal with argument identification, whilst the paragraph-level approach is better when categorizing arguments. This boils down to the conclusion that mining arguments at a fine-grained level also needs a fine-grained learning approach. But we are not sure that the difficulty in mining argument components at essay-level does not actually depend on BERT limitations, in particular on the size of its memory window when considering long texts: so we presently employ LONGFORMER as well, to investigate whether it may be beneficial in overcoming such limitation.

Another point, then, is that the argument classification task is not scale-independent [37], since different argument components operate at different levels. Even though essay- and paragraph-level are popular partitions, these are not, in principle, the only ones that can be taken into account. Our hypothesis is that such approaches are too simplistic, and fail to capture some intrinsic and relevant structural characteristics of the argumentative essays, and that a hybrid-level separation during the training phase might be more suited and efficient to classify argument components. We explore such hypothesis by using both a BERT-based classifier and LONGFORMER-based classifier. Furthermore, using LONGFORMER we also improved the essay-level classification, which was particularly lacking with respect to the paragraph-level classification based on BERT, showing that in this case the two approaches have analogous accuracy. Finally, we implemented two variants of the hybrid-level (called *hybrid+* and *hybrid++*), covering some shortcoming of the basic approach and we registered further improvements when testing using LONGFORMER.

These are the main contributions of this paper: *i)* We report evidence that employing LONGFORMER leads to better results when training at essay-level, and show that a larger window may be helpful in mitigating performance differences compared to performing training at the paragraph-level; *ii)* We introduce a novel hybrid-level approach for learning, showing that it is possible to increase the performance of Argument Classification by mining the argument components at different levels; *iii)* Finally, we show that there exists a different model dependency among the three learning approaches and that not only the hybrid one is better, but it also reduces model dependency.

The paper is structured as follows: in Section 2 we survey related work that precedes and inspires our research. Section 3 provides more details on the Argument Classification task. In Section 4 we present our result and discuss them along with their implications. Section 5 contains conclusions and an outlook on future work.

2 Related Work

This paper mainly lays its foundations in the AM research. AM on structured texts has a long history: among the different application domains we mention news articles [5], scientific articles [1], legal documents [21], healthcare [17] and student essays [31]. Most relevant to our work are those approaches that focus on the classification of argument components in natural language texts. The first approach to identify the argument microstructure were carried out by [21]. They chose the simplest definition of argument as “a set of propositions, being all of them premises, except maximum one, which is a conclusion”. So they used premises and conclusions as argumentative units. Research has continued uninterrupted, also with the help of the advances in machine learning and deep neural architectures: former approaches focused on feature-based models [23,24], but with the advances in machine learning and deep neural network techniques, new approaches were proposed using contextualised word embeddings [29] and adopting the transformers architecture [17] that alleviate the burden of developing *ad hoc* feature selection steps.

In our research we used the Argument Annotated Essay Corpus (AAEC) developed by Stab and Gurevych, containing 402 student essays annotated with argumentative information [31]. The argumentative structure is represented here as a tree, which is a simplified but realistic and useful abstraction for computational applications. This corpus has been extensively studied in subsequent research, that has attempted to improve the performances in AM tasks by using more advanced techniques, and also qualitative accounts have been considered in literature [32,9]. Essays are acknowledged to have a recurrent structure [30,36,38], and there are also proper guidelines to annotate them [31]. It is also important to note that essays considered in the AAEC are written by university students. As demonstrated by [3], 'middle school students' (11-14 years old) essays are quite different due to shortcomings in argumentation quality and conventions.

Eger et al. developed a neural end-to-end model addressing all the AM subtasks using the AAEC corpus [12], and this LSTM-ER model remained the state-of-the-art for a long time [19]. In their work they also compared the essay- and paragraph-level approach, showing that the paragraph-level was able to obtain better results in an easier way, which is also consistent with our own previous results [10]. However, the fact that text sequences are much longer when training at essay-level could also be a shortcoming when dealing with systems who struggle to keep a long memory on these long sequences of text. By contrast, paragraphs are shorter and contain an argumentative integrity that can be at least partly analyzed separately like a watertight compartment, since the argumentation structure in this case is completely contained within a paragraph. We will show in fact that using LONGFORMER when training at essay-level substantially dampens this disparity.

Mayer et al. [18] annotated randomized controlled trials for clinical decision making, and used the same components as Stab and Gurevych [31] but with a different logic: while major claims are usually defined as a stance of the author in the AM literature, here they are defined more as general/introductory claims about properties of treatments or diseases (a general hypothesis to be tested or an observation of a previous study to be confirmed), which is supported by a more specific claim, which is instead a concluding statement made by the author about the outcome of the study. Finally, a premise/evidence is an observation or measurement (observed facts, empirical evidence or comparisons) in the study, which supports or attacks another argument component (usually a claim). In this setting also the absence of change in outcomes plays an important role for clinical decision making, and is thus considered as an evidence in favour of the argumentation.

Bao et al. proposed a transition-based model [4] which can perform argument classification and relation identification simultaneously, increasingly constructing an argumentation graph [4]. The best F1-score were obtained by testing at the token-level on the argument classification task experimenting on the AAEC, while other relevant results include those obtained with the Multi-Task Argument Mining approach [20]. What emerges from the structural analysis of essays is that different types of argument components work at different levels within an essay. Based on this observation, the authors of the work in [37] argue that different types of argumentation components should be mined at different levels: this model obtained a significant improvement on mining major claims and claims with respect to previous models that only worked at essay- or paragraph-level for all the components. Our hybrid-level approach was developed by elaborating on this intuition.

Finally, we have to mention that state-of-art models are cast in a supervised learning fashion; however, some unsupervised approaches have been devised to cope with under-resourced settings. Persing and Ng [25] recently obtained interesting results compared with state-of-art supervised models: this research was concerned with avoiding argument-annotated data, and makes use of heuristics to bootstrap a small set of labels to self-train a model. These findings are relevant, and suggest to reconsider the unsupervised approach, also in the light of how difficult and expensive it may be to handcraft annotated data.

3 Methodology

In this paper we propose a novel learning approach for the *Argument Component Classification* (ACC) task, which is central to the field of Argument Mining. Specifically, ACC consists in the detection of specific argument components in an argumentative text. It is often treated as a supervised text classification problem: given a taxonomy describing the argumentative components, an annotated dataset is exploited to train a system that will perform their automatic recognition on previously unseen data. The kind of argumentative texts together with the adopted components taxonomy can affect the shape of the task. Concerning the taxonomy, most approaches in literature adopt a simplified claim-premise model [22], while other works rely on more complex component definitions [14], such as those by Toulmin [34].

In this work we take in consideration the Argument-Annotated Essays Corpus (AAEC) developed by Stab and Gurevych [30,31], which is to date one of the most widely adopted corpora to experiment on this task. The authors adopt a model that includes major claim, claim and premise to classify argument components in persuasive student essays.

3.1 The Hybrid Approach

In order to explain the intuition behind our approach, we take into consideration the prototypical structure of a student essay, shown in Figure 1. An essay usually begins with an *Introduction*, that describes the controversial topic of the argumentation, and as such is not argumentative itself. The introduction often illustrates the ‘Major Claim’, which is the author’s stance towards the topic of

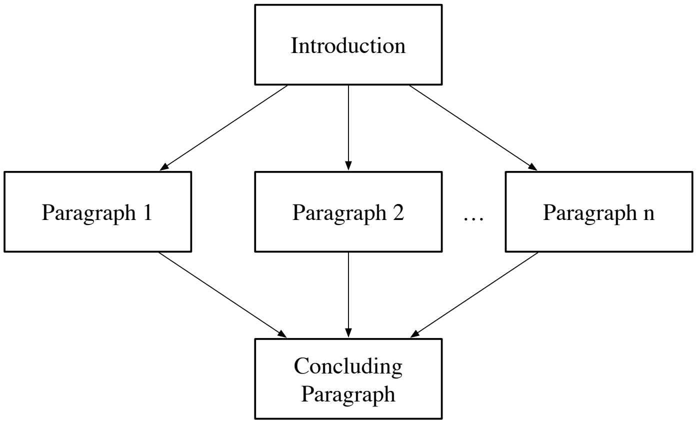

Fig. 1: The general structure of an essay. An essay typically starts with an *Introduction* where the Major Claim (MC) is stated; a set of *Body Paragraphs* follow, containing Premises (P) and Claims (C); finally a *Concluding Paragraph* possibly containing a restatement of the MC, summarizes and ends up the essay.

the argumentation. The actual argumentation thus begins after the introduction, and is developed in a set of *Body Paragraphs*, each containing an argument in favour or against the major claim. Each such paragraph has an internal structure containing a 'Claim', which is the central component of each argument, and one or more 'Premises', either supporting or attacking the claim. Finally, we typically have a *Concluding Paragraph*, which often summarizes the highlights of the essay, restating the major claim and sometimes providing recommendations for future directions (which are also not argumentative themselves). The importance of the structure is also highlighted by other studies on other corpora, such as [36]: authors herein found that, similarly to Stab and Gurevych, the first paragraph usually begins with non argumentative sentences and contains an introduction together with the major claim (called thesis in their research). They also highlight the special roles of the first and last paragraph of an essay.

Systems proposed in literature are trained either at *essay-level* or at *paragraph-level*. In the former case the text of the essay is given in input to the system, and the model is concerned with recognizing all possible argumentative components. In this case the model has access to the the entire structure of the essay, to its tags, and context. In the latter case, instead, paragraphs are the input unit for the tagger, which has to recognize argumentative components within much shorter sequences, and without knowing the entire context of the essay: in fact, if the paragraph is the unit, the model cannot distinguish between paragraphs from different essays, and can only access structure, tags and context within individual paragraphs.

We conjecture that both approaches can be improved by exploiting the structural information available in student essays. In fact, the essay-level training has to deal with the greater variability in the location of the components and may encounter difficulties arising from structural components (in particular MC and C); the paragraph-level, conversely, may fail to recognize the specific role played by the individual paragraphs (paragraphs are not structurally equivalent), e.g., missing the specificity of introduction and conclusion, that intrinsically differ from the central paragraphs. Our hypothesis is therefore that we might guide the system to better trace these components developing an approach to capture the best of the two previous approaches. We call this approach *hybrid-level* training. We have considered three variants: in the basic one (referred to simply as *hybrid*), we have arranged the essay into three parts and then trained the system by feeding it with these three blocks: the introduction, the set of body paragraphs, and the concluding paragraph. Then, in the first variant (referred to as *hybrid+*) we have guided the system to recognize also the boundaries between the internal body paragraphs by inserting a separator —tagged with a specific label— between each of them; in the second variant (referred to as *hybrid++* we have inserted another separator (also with its own specific tag) between introduction and body paragraphs and body paragraphs and conclusion, which allows the essay to be seen again in its entirety as at the essay-level (and no longer split into three blocks), but with structure boundaries this time.

4 Experiments and Discussion

In previous work we investigated the task of Argument Identification and Argument Classification [10]; more precisely, we fine-tuned a BERT-model originally devised by [39]. We presently extend that approach by also employing LONGFORMER [6] to investigate the impact of a longer window for the subword tokenizer and test the effectiveness of competing learning approaches on different models. We cast the task to a span classification problem, using the BIO labeling system as sequence labeling strategy [27]. In this setting, every token is labeled according to the position within or outside an argument component: the tag *'B'* indicates the first token of the argument component, *'I'* is used to label tokens included within a component, and *'O'* is used to mark tokens outside argument components. Since the ACC task involves recognizing different unit types, the B-I-O tags are associated to each component: thus [B,I]-MC, [B,I]-C, [B,I]-P, and O tags for Major Claim, Claim, Premise and Other, respectively. Nevertheless, when testing we only consider 4 tags (one for each component) and we do not distinguish between B and I when calculating accuracy metrics, since they both identify the same component. This means that, at evaluation time, B and I tags are interchangeable for identifying a given component.

Three different training schemes were employed essay-level, paragraph-level and hybrid-level, and experiments were carried out in 5-fold cross-validation; a randomly-chosen 80% of the corpus was used for training and 20% for testing. We recorded F1-scores using both a token-level and the 'α-level matching' method proposed in [24]; this methods considers the matching of spans instead of tokens, and allows considering both exact (100% $\alpha-$level) and approximate (over 50%) matches. In this setting, two text spans are considered an exact match if they

Table 1: Results (F1-scores) on the Argument Classification Task using BERT.

		Essay	Paragraph	Hybrid
Major Claim	Token	65.36	70.74	**76.09**
	α 50%	77.20	79.11	**84.33**
	α 100%	51.59	59.78	**70.71**
Claim	Token	50.93	**58.64**	57.55
	α 50%	56.10	**65.31**	63.43
	α 100%	38.89	**51.80**	51.16
Premise	Token	86.28	87.25	**87.44**
	α 50%	87.74	**89.59**	88.91
	α 100%	75.82	76.00	**76.16**
Other	Token	**88.18**	85.99	87.78
	α 50%	**96.28**	95.07	95.09
	α 100%	**93.54**	90.82	91.08

are featured by same boundaries, whilst they are considered as an approximate match if they share over half tokens. This more lenient evaluation metrics is customarily used also to assess human annotators agreement, which is not always full in complex tasks, such as in the present one.

Let us start by introducing the results obtained when employing the BERT-based model. The results on the ACC task obtained by training at essay-, paragraph- and hybrid-level are illustrated in Table 1. In this case, for the sake of brevity, we have only considered the hybrid approach in its basic form; the models *hybrid+* and *hybrid++* were only tested using LONGFORMER as performances are generally better than BERT. The three metrics essentially reveal the same pattern. The MC is the component that benefits more from the hybrid approach, revealing that separating introduction and conclusion from the body paragraphs during the training helps in classifying such component. We obtain a 5% improvement with respect to the paragraph-level in classifying MC at token level, and a 5% and 10% improvement at the 50% and 100% α-level, respectively. Conversely, the C classification only loses 1% with respect to the paragraph-level, and also the classification of P registers the best results (by a reduced margin, though) in terms of F1 Score in 2 out of 3 metrics. The essay-level is less appropriate in classifying MC and C: our results unveil the difficulty of the BERT-based model to handle the whole essay, while it is surprisingly effective in classifying O. In general we observe that C is the hardest component to classify, probably because it varies to a greater extent (please also refer to results in [10]. In fact, training at the paragraph-level is the most suitable perspective for C, since we have a smaller degree of variability within a single paragraph

Table 2: Results (F1-scores) on the Argument Classification Task using LONGFORMER.

		Essay	Paragraph	Hybrid	Hybrid+	Hybrid++
Major Claim	Token	77.50	75.78	78.49	**78.97**	78.27
	α 50%	82.71	83.83	85.48	**86.21**	85.15
	α 100%	70.57	71.31	72.88	73.75	**74.11**
Claim	Token	57.51	61.24	60.89	**63.97**	60.12
	α 50%	62.64	67.23	66.62	**68.48**	65.58
	α 100%	53.81	58.77	57.02	**61.01**	57.19
Premise	Token	88.23	88.12	88.55	**89.29**	88.83
	α 50%	89.62	90.10	89.85	**90.60**	90.15
	α 100%	80.13	78.10	78.0	79.29	**81.03**
Other	Token	**89.74**	87.05	88.55	88.88	89.71
	α 50%	**96.07**	94.60	95.11	95.45	96.01
	α 100%	**93.72**	90.61	90.75	91.66	93.41

with respect to an essay or all the body paragraphs gathered together. This is also supported by literature: e.g., in [36] regularities were found in the argumentation flow within body paragraphs, showing that students tend to first state a claim and then argue for it; also, it was showed that there is a tendency to state the central claim of a paragraph in the very first sentence, followed by the end of the text [22]. Such tendency to state the central claim at the beginning of a paragraph seems to be a peculiarity of the English language (and of Anglo-Saxon cultures, more in general), since other studies show that in documents authored by Asian people the claim is mostly found at the end [26,16]. Even MC can be either posited at the beginning of the essay or pushed into the middle, mostly when it contains background information about the discussion topic. In this case, having the introduction separated and more identifiable from the other paragraphs gives to the model less ambiguity to identify the MC using semantic and syntactic information.

Table 3: Averaged results (F1-scores) on Argument Classification using BERT and LONGFORMER.

	BERT			LONGFORMER				
	Essay	Paragraph	Hybrid	Essay	Paragraph	Hybrid	Hybrid+	Hybrid++
Token	72.69	75.66	**77.21**	78.25	78.05	79.12	**80.28**	79.23
α 50%	79.33	82.27	**82.94**	82.76	83.94	84.26	**85.26**	84.22
α 100%	64.96	69.60	**72.28**	74.56	74.70	74.68	76.37	**76.44**

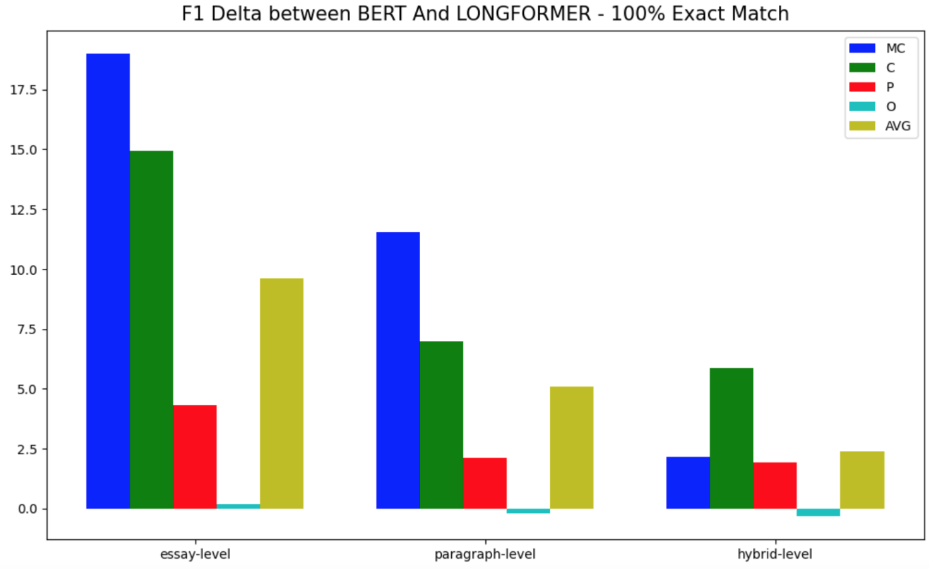

Fig. 2: Difference in F1-score when passing from BERT to LONGFORMER (F1-delta) using the three different approaches (case exact match).

The results recorded by employing LONGFORMER (presented in Table 2) show improved results with respect to those of BERT. Even in this case the hybrid-level brings a significant improvement for MC, and is in general more favorable than the other two approaches, as also confirmed in Table 3. More specifically, we can see that the results obtained by employing LONGFORMER improve over those obtained through the BERT-based models around 2% at token-level, 1.5% at 50% α-level and 2.5% at 100% α-level for the hybrid-level. Even more consistent is the improvement for the essay-level which increase 6%, 3.5% and 10% respectively, revealing the relevance of long-memory cutting. Also the paragraph-level registers improvements in the order of 3.5%, 1.5% and 5%. In general, the exact-matching is the perspective which benefits the most from passing from BERT to LONGFORMER, probably due to the fact that it is easier to improve a lower performance.

Figure 2 shows the F1-delta scores by passing from BERT to LONGFORMER in the case of exact match; for the sake of brevity, we report the figures for exact match condition only testing with the the basic hybrid level; token- and approximate-levels also reveal the same trend. Higher bars illustrate experimental conditions where LONGFORMER ensures higher improvements with respect to BERT. This plot also shows that the hybrid approach, in addition to higher accuracy, also exhibits a lower dependence on the model, since the F1-delta is the lowest one, while it is clear how the essay-level is highly dependent on the model in this case. That is, the reduced difference between LONGFORMER and BERT when adopting the hybrid approach may be explained by a simple effect:

the shorter the text excerpts being processed, the lesser the benefits deriving from employing a longer memory window (in such a case, almost all texts could be processed through BERT with little loss). The results obtained experimenting with LONGFORMER at essay-level are also of interest: the classification of MC is improved by 12%, 5.5% and 19% for token-level, approximate and exact match, respectively, when passing from BERT to LONGFORMER; the classification of C is improved by 6.5%, 6.5% and 15% instead. Such results illustrate how the larger memory window of LONGFORMER impacts on the ACC of argumentative essays.

To complete the assessment, we need to mention that the three approaches and the two models are also featured by different computational properties. LONGFORMER requires more computational resources than BERT and the same holds for the paragraph-level, since there are more chunks of text to analyse.[4] This involves that using the hybrid-level is also beneficial in saving computational resources and, since it has less model dependency, using BERT hybrid-level also ensures a good computational gain with a loss in accuracy in the order of 2%. Finally, guided by the encouraging results obtained with the basic hybrid approach, we developed two variants called *hybrid+* and *hybrid++* in order to overcome some shortcoming of the basic hybrid approach, and whose results are reported in Tables 2 and 3. In the first case (*hybrid+*) we still have a partitioning of the essay in introduction, body paragraphs and conclusion but we have informed the system about the boundaries of the internal paragraphs through the insertion of a specific separator (tagged as *P-Sep*); thanks to this arrangement we were able to reach a consistent improvement on each component and in particular on C. In the second case (*hybrid++*) we used the same separators as in *hybrid+*, but instead of considering a partitioning into three blocks we returned to consider the essay in its entirety (since this allows to save computational resources) and we inserted another separator, tagged as *IC-Sep*, between the introduction and the body paragraphs and the body paragraphs and the conclusion (thus two of these separators for each essay). The last setting is particularly beneficial when considering the exact matching; it also helps improving the classification of O, which seams particularly good when considering the whole essay.

5 Conclusions and Future Work

In this work we have been experimenting on the Argument Component Classification task: we introduced a novel level (the hybrid level) to train the models, and compared and commented results obtained through models based on BERT and LONGFORMER. In so doing, different strategies were employed to train the

[4] Experiments were performed on machinery provided by the Competence Centre for Scientific Computing [2]; nodes employed were equipped with 2x Intel Xeon Processor E5-2680 v3 and 128GB memory. Running experiments with LONGFORMER took 11.5 hours to complete 15 epochs at essay-level, while BERT only took 4 hours. At paragraph-level instead, LONGFORMER took 28 hours, BERT 9.5.

models: in the essay-level setting we used entire essays, and in the paragraph-level one we arranged the essays into their paragraphs, following two popular approaches from the literature. We then introduced a novel strategy named *hybrid-level* and two variants: in the basic form we differentiate introduction, set of body paragraphs, and conclusion; the first variant also considers separators between internal body paragraphs; the second variant abandons the partition into three blocks and considers the entire essay, but with two specific separators to better mark the structure of the essay and speed up the computation. It is noteworthy that such hybrid approaches all have in common the goal to better fit the essay structure of components when mining them. We found that this learning perspective is beneficial with respect to the classical essay- and paragraph-level when performing argument classification, using both BERT and LONGFORMER. Then, comparing the two transformers models, we found that the results obtained through LONGFORMER consistently improve on those obtained with BERT, in particular when training at the essay-level: this fact shows that the longer memory-window of LONGFORMER ensures better results when analysing text sequences. Finally, we provided experimental evidence supporting the intuition that argumentation, in particular within structured texts like argumentative essays, typically follows a particular and recurrent structure that can be exploited to facilitate the learning phase. Since our hybrid-level strategy is a model-free solution, we hope that these findings can be helpful for further research.

Future directions will consider different aspects. For example, in the classification of claims we recorded lower accuracy with respect to the other components, showing that this step is harder and still needs further efforts. Furthermore, to enhance the robustness of this technique there is also the necessity to test the hybrid approach on state-of-the-art systems and on further types of argumentative texts featured by an underlying recurring structure. This is a first exploratory step in this direction which has shown encouraging prospects.

References

1. Accuosto, P., Saggion, H.: Mining arguments in scientific abstracts with discourse-level embeddings. Data & Knowledge Engineering **129**, 101840 (2020)
2. Aldinucci, M., Bagnasco, S., Lusso, S., Pasteris, P., Rabellino, S., Vallero, S.: OCCAM: a flexible, multi-purpose and extendable HPC cluster. Journal of Physics: Conference Series **898**(8), 082039 (2017)
3. Alhindi, T., Ghosh, D.: " sharks are not the threat humans are": Argument component segmentation in school student essays. arXiv preprint arXiv:2103.04518 (2021)
4. Bao, J., Fan, C., Wu, J., Dang, Y., Du, J., Xu, R.: A neural transition-based model for argumentation mining. 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). pp. 6354–6364 (2021)
5. Basile, P., Basile, V., Cabrio, E., Villata, S.: Argument mining on italian news blogs. In: Third Italian Conference on Computational Linguistics (CLiC-it 2016)

& Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016) (2016)

6. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)
7. Bentahar, J., Moulin, B., Bélanger, M.: A taxonomy of argumentation models used for knowledge representation. Artificial Intelligence Review **33**(3), 211–259 (2010)
8. Cabrio, E., Villata, S.: Five years of argument mining: a data-driven analysis. In: IJCAI. vol. 18, pp. 5427–5433 (2018)
9. Carlile, W., Gurrapadi, N., Ke, Z., Ng, V.: Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 621–631 (2018)
10. Demaria, R., Delsanto, M., Colla, D., Mensa, E., Pasini, E., Radicioni, D.P., et al.: Shuffling-based data augmentation for argument mining. In: CEUR WORKSHOP PROCEEDINGS. pp. 1–17. CEUR (2022)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
12. Eger, S., Daxenberger, J., Gurevych, I.: Neural end-to-end learning for computational argumentation mining. arXiv preprint arXiv:1704.06104 (2017)
13. van Haaften, T.: Frans h. van eemeren, bart garssen, erik cw krabbe, a. francisca snoeck henkemans, bart verheij and jean hm wagemans: Handbook of argumentation theory: Springer, dordrecht, 2014, isbn: 978-90-481-9472-8 (hardback), isbn: 978-90-481-9473-5 (ebook), isbn: 978-90-481-9474-2 (print and electronic bundle) (2016)
14. Habernal, I., Gurevych, I.: Argumentation mining in user-generated web discourse. Computational Linguistics **43**(1), 125–179 (2017)
15. Hinton, M.: Evaluating the language of argument, vol. 37. Springer (2020)
16. Kaplan, R.B.: Cultural thought patterns in inter-cultural education. Language learning **16**(1-2), 1–20 (1966)
17. Mayer, T., Cabrio, E., Villata, S.: Transformer-based argument mining for healthcare applications. In: ECAI 2020, pp. 2108–2115. IOS Press (2020)
18. Mayer, T., Marro, S., Villata, S., Cabrio, E.: Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials. Artificial Intelligence in Medicine p. 102098 (May 2021). https://doi.org/10.1016/j.artmed.2021.102098, `https://hal.science/hal-03264761`
19. Miwa, M., Bansal, M.: End-to-end relation extraction using lstms on sequences and tree structures. arXiv preprint arXiv:1601.00770 (2016)
20. Morio, G., Ozaki, H., Morishita, T., Yanai, K.: End-to-end argument mining with cross-corpora multi-task learning. Transactions of the Association for Computational Linguistics **10**, 639–658 (2022)
21. Palau, R.M., Moens, M.F.: Argumentation mining: the detection, classification and structure of arguments in text. In: Proceedings of the 12th international conference on artificial intelligence and law. pp. 98–107 (2009)
22. Peldszus, A., Stede, M.: An annotated corpus of argumentative microtexts. In: Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon. vol. 2, pp. 801–815 (2015)
23. Peldszus, A., Stede, M.: Joint prediction in mst-style discourse parsing for argumentation mining. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 938–948 (2015)

24. Persing, I., Ng, V.: 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. pp. 1384–1394 (2016)
25. Persing, I., Ng, V.: Unsupervised argumentation mining in student essays. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 6795–6803 (2020)
26. Putra, J.W.G., Teufel, S., Tokunaga, T.: Parsing argumentative structure in english-as-foreign-language essays. In: Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 97–109 (2021)
27. Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Natural language processing using very large corpora, pp. 157–176. Springer (1999)
28. Reed, C.: Argument technology for debating with humans (2021)
29. Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., Gurevych, I.: Classification and clustering of arguments with contextualized word embeddings. arXiv preprint arXiv:1906.09821 (2019)
30. Stab, C., Gurevych, I.: Identifying argumentative discourse structures in persuasive essays. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 46–56 (2014)
31. Stab, C., Gurevych, I.: Parsing argumentation structures in persuasive essays. Computational Linguistics **43**(3), 619–659 (2017)
32. Stab, C., Gurevych, I.: Recognizing insufficiently supported arguments in argumentative essays. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 980–990 (2017)
33. Stab, C., Miller, T., Gurevych, I.: Cross-topic argument mining from heterogeneous sources using attention-based neural networks. arXiv preprint arXiv:1802.05758 (2018)
34. Toulmin, S.: The uses of argumentcambridge university press. Cambridge, UK (1958)
35. Van Eemeren, F.H., Grootendorst, R., Johnson, R.H., Plantin, C., Willard, C.A.: Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments. Routledge (2013)
36. Wachsmuth, H., Al Khatib, K., Stein, B.: Using argument mining to assess the argumentation quality of essays. In: Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers. pp. 1680–1691 (2016)
37. Wang, H., Huang, Z., Dou, Y., Hong, Y.: Argumentation mining on essays at multi scales. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 5480–5493 (2020)
38. Wang, X., Lee, Y., Park, J.: Automated evaluation for student argumentative writing: A survey. arXiv preprint arXiv:2205.04083 (2022)
39. Yang, X., Bian, J., Hogan, W.R., Wu, Y.: Clinical concept extraction using transformers. Journal of the American Medical Informatics Association **27**(12), 1935–1942 (2020)