Shuffling-Based Data Augmentation for Argument Mining

Roberto **Demaria**[1], Matteo **Delsanto**[1], Davide **Colla**[1], Enrico **Mensa**[1], Enrico **Pasini**[2] and Daniele P. **Radicioni**[1]

[1]*Dipartimento di Informatica, Università degli Studi di Torino*

[2]*Istituto per il Lessico Intellettuale Europeo, ILIESI/CNR – Roma*

Abstract

Argument Mining is an emerging research area in natural language processing; it is concerned with extracting arguments and their structure from text documents. Deep neural networks and contextualized word embeddings have been recently obtaining state-of-the-art results on various classification and relation extraction tasks including argument mining, but they tend to learn spurious correlations and to memorize high-frequency patterns possibly undermining systems' predictions. In this paper we illustrate how transformers-based models fine-tuned on argumentative student essays are biased (and their performance are thereby affected) by their structure; additionally, we show that adopting data augmentation by shuffling sentences may be helpful in reducing structure dependency and to improve generalization.

Keywords

Natural Language Processing, Argument Mining, Data Augmentation, Transformers

1. Introduction

Argumentation is a verbal and social activity aimed at increasing or decreasing the acceptability of a controversial standpoint: it typically involves logical and linguistic competences going well beyond syntax, semantics and pragmatics. Whereas the study of argumentation was originally a philosophical subject, it is to date a comprehensive and interdisciplinary research field: it involves many other research areas including communication science, linguistics, psychology and computer science. Computational Argumentation (CA) is a recent research field in computational linguistics that focuses on the automatic analysis of arguments in natural language texts [1]. Until recently, the argumentation analysis has been mostly considered a manual task performed through a skilled and time-consuming process; its scientific relevance and high economic potential for innovative applications are making CA increasingly relevant. A relevant branch in CA is Argument Mining (AM), that is the automatic extraction of the argumentative structure of a text [2]: this typically includes the identification and classification of the arguments, their components and the relations intervening between them.

State-of-the-art approaches in AM adopt supervised learning and deep neural networks, but one of the main challenges is the lack of appropriately annotated corpora with argumentative

information to train and test Machine Learning systems [3]. The creation of a corpus for AM is indeed a difficult and time-consuming task, one requiring costly resources to obtain high quality annotations, as delivered by expert annotators. Furthermore, the annotation process itself can be subjective and to some extent controversial: even the annotations by expert annotators may be affected by poor inter-rater agreement [4], thus menacing the learning process. Another issue related to existing corpora is that they are typically small in size and domain-specific, thereby making harder the learning task. To face these problems different solutions have been proposed in AM, such as Transfer Learning (TL) and Cross-Corpora systems [5]. Another approach that has been proposed to deal with data scarcity and imbalance is Data Augmentation (DA): DA techniques can expand the amount and diversity of training examples without explicitly collecting new data in the wild, but manipulating or perturbing existing ones, also helping to reduce variance and overfitting in a simple and low-cost way [6]. Among classic and unconditional text DA, two directions have been extensively investigated: *adding noise* through text editing, for example by substituting/deleting words such as Easy Data Augmentation (EDA) [7]; and *back translation*, translating the original text into other languages (one or more), and then back to the original one [8]. To the best of our knowledge, none of these DA techniques have been investigated in AM.

Deep Neural Networks and transformer-based architectures such as the Bidirectional Encoder Representations from Transformers (BERT) [9] have been obtaining state-of-the-art results on various classification and relation extraction tasks including AM, but they also tend to learn spurious correlations and memorize high-frequency patterns that are difficult for humans to detect but influence predictions. These may amount to frequent patterns or to extracting specific linguistic structures that do not generalize well [10]. In this perspective, DA can act as a regularization strategy, whereby overfitting is contrasted by shuffling the particular forms of language and thus mitigating the influence of unwanted patterns.

Three steps are typically involved in the closed-domain investigation [1]:

- *Argument Identification* is concerned with categorizing argumentative and non-argumentative sections in a given text;
- *Argument Classification* addresses the *function* of argument components, classifying them into different types according to the linguistic model adopted;
- *Structure Identification* is to assign the relation type to directly connected arguments.

The main contributions of this paper are as follows: *i)* We adapted an existing BERT-based model originally developed for Information Extraction in the Medical Domain to perform Argument Identification and Classification; *ii)* We obtained results on par with state-of-the-art models, and analyzed the different performances at various levels of training and tags; *iii)* We propose a novel DA technique to reduce the dependency from the structure, and experimentally found beneficial effects on the accuracy of the employed model.

The paper is structured as follows: Section 2 surveys related work that precedes and inspires our research. Section 3 presents the dataset used in the experiments, along with its properties. In Section 4 we introduce the BERT-based model, and report the results obtained in the Argument Identification and Classification. In Section 5 we introduce the novel shuffling technique exploited for augmenting data and show its impact on the argument classification task. Section 6 contains conclusions and an outlook on future work.

2. Related Work

2.1. Background on Argument Mining

One pioneering approach to AM was proposed in the legal domain, aimed at argument identification [11]. The first corpus of argumentative student essays of attested high quality, composed by 90 essays, later expanded to 402, was compiled in [12]; on this basis AM applications could be tested and their results compared [1]. Historically, research on models dealing with multiple domain corpora required to solve some issues, such as the difficulty of merging such corpora, which were often annotated based on different argumentation schemes [13]. In the early phase, the three sub-tasks of AM (Argument Identification, Argument Classification and Structure Identification) were carried out separately; pipelines usually employed domain specific feature-based models, and the argumentative structure was mostly considered as a tree.

In [14] sequence-to-sequence attention modeling was applied to structural prediction in discourse parsing tasks, and a joint model was developed to extend this architecture to simultaneously address the link extraction task and the classification of argument components. This work showed that joint optimization on both tasks is crucial for high accuracy results. The first results on end-to-end AM in student essays using a pipeline approach were presented in [15]; they also proposed a novel notion for scoring metrics specifically tailored for AM, that has been referred to as 'α level matching' [16]. In particular, this notion allows to distinguish between exact (100% level) and approximate (50% level) match. In the first case predicted and gold components must have exactly the same spans, whereas in the 50% level they need to share at least 50% tokens.

A trend in more recent AM approaches is to construct end-to-end models using contextualized world embeddings. The first neural end-to-end model, jointly addressing all sub-tasks and employing the Argument Annotated Essay Corpus (AAEC) was proposed in [16], and the results obtained by their LSTM-ER model [17] lasted as state-of-art results until recently. A new approach based on a modified biaffine dependency parsing was proposed by [18], by replacing its embedding layer and BiLSTM layer with a pre-trained BERT encoder and was able to reach a new state-of-the art in AM. The work in [19] showed that adding information about Part of Speech tagging (POS), Chunking and using advanced deep learning techniques lead to results favorably comparing with those obtained by state-of-the-art systems exploiting hand-crafted features. Another relevant work adopting BERT and the transfer learning approach included four corpora from different domains, including the AAEC, to construct a binary identification model ("Argument", "Non-Argument") [20]. The approach introduced by AMPERSAND (Argument Mining for PERSuAsive oNline Discussions) employs BERT to perform argument classification and relation extraction, bringing together both micro-level (intra-argument relations) and macro-level (inter-argument relations) models of argumentation [21]. Moreover, a neural transition-based model has been developed to handle both tree and non-tree structured argumentation schemes, and to incrementally construct an argumentation graph [22]. This model obtained the best accuracy on the argument classification task, experimenting on AAEC data. An approach to apply transfer learning across auxiliary AM corpora and to develop an end-to-end cross-corpus model using multi task learning called Multi-Task Argument Mining (MT-AM) has been introduced in [5]. This work obtained state-of-art accuracy for argument classification and

relation extraction on AAEC data. Argument classification on 'middle school students' (11-14 years old) essays is instead the focus of the work in [23]. However, such sort of essays has been acknowledged as rather different from the university students' essays that we are employing in the present work because poorly compliant with argumentative conventions, which makes such data more challenging and difficult to analyze.

Finally, a *sentence rearrangement* step was introduced in [24] to improve the quality and consistency of the essay at the level of textual surface, to retain the original meaning of the sentence and to track major claims, by reordering sentences, replacing pronouns with their referents, and removing or replacing inappropriate connectives. This research provided evidence about the potential to use well-written texts enriched with syntactic information, together with noisy texts, to increase the size of AM training data. This work can be compared to ours, in that we also investigate how to apply DA techniques with the purpose to mitigate the effects of structural features, possibly undermining the classification task. One chief difference is, nevertheless, that we do not employ deep semantic techniques (such as parsing and reference resolution), but rather a shallow sampling approach.

2.2. Background on Data Augmentation

DA is a simple and low-cost technique that has been successfully applied in some areas of AI, such as computer vision and speech recognition. Text Data Augmentation (TDA) was also successfully employed in some NLP tasks such as text classification, question answering and multi-turn dialogue [25]. More specifically, it has been used for classification problems where class boundaries are learned from label assignments [26], similar to the argument mining tasks. DA has been largely employed in low-resource tasks and domains, to tame data scarcity and imbalance, and has been employed as a regularization strategy [6]. As mentioned, AM corpora often suffer from lack of data, and models tend to overfit a particular domain or structure; DA has thus been profitably applied in tasks such as sequence tagging, parsing and dialogue systems.

There are essentially two classes of (TDA): classical TDA, also called *unconditional*, includes rule-based methods that have been applied in NLP without the use of neural generative methods, such as *back translation* [8] or *text editing*. EDA is a good example of this lexical technique, consisting of four simple but powerful operations of token-level random perturbation: synonym replacement, random insertion, random swap and random deletion [7]; EDA obtained good results in text classification tasks using small datasets, but it can also easily interface with pair-wise classification, extractive question answering, abstractive summarization, and chatbots [26]. Another class of DA methods is concerned with *conditional* neural text generation, that can be used to create new text or counterfactual examples [27].

Notwithstanding its merits, to the best of our knowledge DA remains largely unexplored in AM. It was used in multilingual AM, where the original English training data were augmented through machine-translated data of other languages: results show that this approach may be helpful in tracking the stance of the argument towards the topic [28]. The same study shows that for evidence detection (evidence relevant to the topic), adding data from the target language along with related languages also improves performances.
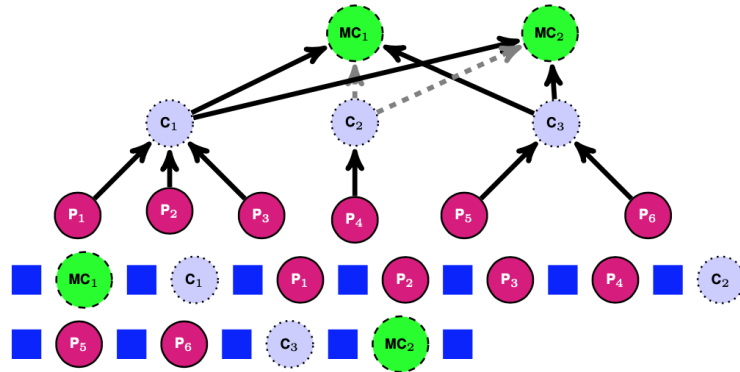
Figure 1: The general structure of an essay. In a linear perspective there are non-argumentative units (blue square) and argumentative components: Major Claims (MC), Claims (C) and Premises (P). Relations are represented as solid (support and for) and dashed (attack and against) arrows. Figure borrowed from [16].

3. Dataset

Similar to most surveyed literature, we used the AAEC [1], which contains university level student essays annotated with argumentative information at the token-level. An essay is a structured text divided into paragraphs, where a specific and controversial topic is discussed; it typically begins with an introduction, followed by a series of body paragraphs and ends with a conclusion. Based on this structure, Stab and Gurevych [1] developed annotations guidelines to deal with these types of text and chose three argument components:

1. *Major Claim*: the stance of the author with respect to the essay's topic;
2. *Claim*: a statement that are either directly for or against the major claims;
3. *Premise*: a statement giving reasons for claims or other premises and either support or attack them.

Such components are then structured as a tree and connected by appropriate relations. Relations follow a precise structure in this model: we have *outgoing relations* (*ORs*) and *incoming relations* (*IRs*) that link the various components. Each premise has one OR and none or several IRs, e.g., coming from further premises. A claim may have one or more IRs coming from different premises and one OR going towards the major claim, which in turn exhibit one or more IRs but no ORs. The relation of each argument to the major claim, or the direct relation between a claim and the major claim is indicated by a stance attribute which can either be *for* or *against*, while the relationship between two premises or between a premise and a claim is marked either as *support* or *attack*. The dataset consists of 402 essays, spanning over 7, 116 sentences divided into 1, 833 paragraphs. Regarding the argument components, 6, 089 instances have been annotated, where 73% are premises, 18% claims and 9% major claims. Additionally, 5, 338 relations have been identified, where 68% are *support* and 4% are *attack* relations while 23% are *for* and 5% are *against* relations.

The argumentation structure related to a claim (i.e., premises either attacking or supporting

```
TITLE: Should students be taught to compete or to cooperate?

It is always said that competition can effectively promote the development of economy. In order to surv
ive in the competition, companies continue to improve their products and service, and as a result, the
whole society prospers. However, when we discuss the issue of competition or cooperation, what we are c
oncerned about is not the whole society, but the development of an individual's whole life. From this p
oint of view, I firmly believe that we should attach more importance to cooperation during primary educ
ation.
 First of all, through cooperation, children can learn about interpersonal skills which are significant
in the future life of all students. What we acquired from team work is not only how to achieve the same
goal with others but more importantly, how to get along with others. During the process of cooperation,
children can learn about how to listen to opinions of others, how to communicate with others, how to th
ink comprehensively, and even how to compromise with other team members when conflicts occurred. All of
these skills help them to get on well with other people and will benefit them for the whole life.
 On the other hand, the significance of competition is that how to become more excellence to gain the v
ictory. Hence it is always said that competition makes the society more effective. However, when we con
sider about the question that how to win the game, we always find that we need the cooperation. The gre
ater our goal is, the more competition we need. Take Olympic games which is a form of competition for i
nstance, it is hard to imagine how an athlete could win the game without the training of his or her coa
ch, and the help of other professional staffs such as the people who take care of his diet, and those w
ho are in charge of the medical care. The winner is the athlete but the success belongs to the whole te
am. Therefore without the cooperation, there would be no victory of competition.
 Consequently, no matter from the view of individual development or the relationship between competitio
n and cooperation we can receive the same conclusion that a more cooperative attitudes towards life is
more profitable in one's success.
```

Figure 2: Example of an essay from the AAEC with argument components annotated. Major Claims are highlighted in red, Claims in blue and Premises in green.

it) is completely contained within the paragraph including that particular claim. For this reason some approaches perform the predictions at the paragraph-level instead of essay-level. Which of the two methods is better is however controversial and may depend on the model. Another important issue relies to the fact that arguments do not necessarily cover an entire sentence. Stab and Gurevych [1] identified some preceding text units, called "shell language", that can help recognizing the argument components and their type, but should not be marked as arguments. Furthermore, a sentence might include more than one argument component. They also found that this shell language can be very useful to identify the boundaries of argument components.

The general structure of an essay is shown in Figure 1 [16]. The argumentative components can be reconstructed in a tree structure. Here solid arrows represent *support* and *for* relations and dashed arrows represent *attack* and *against* relations. In a linear perspective an essay can be seen as a sequence of non-argumentative units, represented as dark blue squares, and argument components. Since there may be several major claims (typically two), each claim potentially connects to multiple targets, violating the tree structure. However, there are some tricks one can use to uniquely reconstruct a tree. Since all major claims within an essay are considered to be equivalent in meaning, they can be treated as a single special root node [16].

Figure 2 shows an example of essay from the AAEC where argument components are annotated.

4. Experimentation

In this Section we report the results of a preliminary experimentation on Argument Identification and Classification.

Both Argument Identification and Argument Classification were addressed by exploiting an existing library originally devised for information extraction in the medical domain [29]. This model was originally conceived to extract named entities from clinical texts, framing the

O	O	O	O	B	I	I	I
I	firmly	believe	that	we	should	attach	more

importance	to	cooperation	during	primary	education	.
I	I	I	I	I	I	O

Figure 3: The output of the model for the sentence *I firmly believe that we should attach more importance to cooperation during primary education.* Here the token *we* is marked as the beginning of an argumentative unit, followed by the subsequent tokens. The full stop is classified as not belonging to the unit.

problem as a span detection task. Considering the similarity between the problem for which this model was designed and the task addressed in our experiments, and also given the versatility of the library, we fine-tuned for 15 epochs the model (using an early-stop condition on 5 epochs) on both Argument Classification and Argument Identification by opting for BERTbase [9].

Data representation: BIO labelling. We cast both Argument Identification and Argument Classification to span classification problems: that is, we aim at detecting the boundaries of each argumentative unit within the essays. We employed a sequence labeling strategy using the BIO labeling schema [30]: each token in a sentence is labeled with one of the B-I-O tags, where B indicates the first token of the argumentative unit, I is used to label tokens within a unit, and O marks tokens that do not belong to any argumentative unit. The adopted model allows to reshape the problem as token classification task that, given a sentence $W = w_1 w_2 \ldots w_n$, amounts to labeling each word w_i with one of the B-I-O tags. Figure 3 reports an example of the system output for the sentence from Figure 2. Considering the example, we can see that the token *we* is correctly tagged as the beginning of the new argumentative unit, while the token *education* is labeled as the last token within the same unit.

Given that we are interested in dealing with both tasks with the same architecture, we devised two different sets of tags according to the task definition. More precisely, for the Argument Identification task we adopted the classic B-I-O tags during the training to identify the boundaries of each argumentative unit regardless of its type. At test time, in addition to the classic B-I-O tags, we also employ an 'A' tag (to arguments, and thus including both B and I tags); however, not being part of the B-I-O labels, this is only a metrics to assess the accuracy of the system, and not actually part of the B-I-O encoding schema. Since the Argument Classification task requires to recognize different unit types, we adopted a different set of tags: here, extending the B-I-O tags logic, each token may be labeled with [B,I]-MC, [B,I]-C, [B,I]-P, or O tags for Major Claim, Claim, Premise or Other, respectively. At test time, similar to the identification task, we also use a synthetic (or compressed) labeling representation, including 4 tags instead of 7, to grasp the ability of the system to categorize components independently of B and I.

Data Partitioning. We employed a BERT-based classifier, and the experiments were carried out in 5-fold cross-validation. For the identification task we used three different training schemes: at sentence-level, at paragraph-level and at essay-level, while during classification we only trained at the paragraph and essay-level.

Table 1

Results of experiments in the Argument Identification task, performed at the sentence, paragraph and essay level. For each experimental setting columns report Precision (P), Recall (R), their harmonic mean F1(P,R), and accuracy (Acc.).

(a) Results at sentence, paragraph and essay level. Best results for each metric and for a given tag are marked in bold, for example, considering the Precision (P) on the I-tag, the highest score is obtained by the essay-level identification while the best precision for B-tag is obtained at sentence-level. This means that for each metrics we found four best results (one for each tag). The dashed line separates the scores obtained with the B-I-O tags and the synthetic A tag.

Level	Tag	P	R	F1	Acc.
Sentence	B-tag	**74.27**	86.42	79.85	98.17
	I-tag	89.83	92.10	90.91	88.04
	O	82.92	75.43	78.76	87.61
	A	89.74	92.76	91.18	87.61
Paragraph	B-tag	73.16	89.41	**80.44**	**98.18**
	I-tag	92.91	94.41	93.64	91.66
	O	88.64	82.69	85.52	91.46
	A	92.63	95.28	93.92	91.46
Essay	B-tag	72.34	**89.96**	80.18	98.13
	I-tag	**93.53**	**95.47**	**94.48**	**92.73**
	O	**90.98**	**83.98**	**87.33**	**92.60**
	A	**93.20**	**96.39**	**94.76**	**92.60**

(b) Comparison against the state of the art.

Model	F1
CRF ([1])	86.70
ACD ([19])	88.70
LSTM-ER ([16])	**90.84**
Wambsganss et al. ([20])	85.19
ours	87.33

Evaluation Metrics. We assessed Argument Identification results through Precision, Recall, Accuracy and F1 score, that are standard Information Retrieval metrics. In Argument Identification these are computed by counting true positive, true negative, false positive and false negative at the token-level, while in Argument Classification we also used the 'α level matching' method proposed in [15], considering matching of spans (instead of tokens); we only considered Precision, Recall, and F1 score as evaluation metrics. The corpus is quite imbalanced, and thus Accuracy is not so meaningful. This approach is coherent with surveyed literature [16, 5], and involves considering both exact and approximate (over 50%) matches. In this setting, two text spans i and j are considered an *exact match* if they have exactly the same boundaries, whereas 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 beings' agreement, which is not always full in complex tasks, such as the present one [15].

4.1. Argument Identification

Our results in the Identification task are presented in Table 1. Results tend to improve as the granularity of training data becomes rougher, that is from the sentence-level to the essay-level. The reason is probably due to the fact that paragraphs and essays contain some complete arguments, while sentences do not (or equivalently, arguments seem to often span across sentences). But paragraphs can show different behaviors and roles within an essay: for example, the first and last paragraph generally contain fewer argumentation elements, or none.

Table 1b reports a comparison with results from the literature: our F1-score has been computed as the arithmetic mean among B-I-O tags, as in [1]; and best results are reached at the essay-level. (please refer to Table 1a). The 'A' tag measures whether the argumentative units (without distinction between B and I) were correctly discriminated from the not argumentative units.

4.2. Argument Classification

As earlier mentioned, this task is to recognize different unit types and mark them as Major Claim, Claim, Premise or Other. We did not perform the Argument Classification in a pipeline: rather, the Argument Classification task was performed independently from the Argument Identification, using the same BERT-based model directly for the classification. Furthermore, we adopted the BIO notation with 7 labels, and the more synthetic notation with 4 labels (that is, without distinction between B and I tags).

Results in the Classification task are provided in Table 2. In Table 2a and Table 2b we report results for the training at paragraph-level, which leads to the best results in general, while in Table 2c we also reported the averaged F1 score at essay-level for comparison (we averaged using the 4-tags notation). More specifically, Table 2a shows the results obtained using BIO tags for each component (such as MajorClaim B-tag or MajorClaim I-tag) and also using synthetic tags (such as MajorClaim tag). In the first case we have 7 labels: B and I for each of the three components and the O component, while in the second case labels reduce to 4 since we do not distinguish between B and I. Results for the approximate and exact match are reported in Table 2b. From these sets of results, we observe that claims are more difficult to identify, maybe because, compared to the other components, they exhibit an higher degree of variability in the essay structure. When the training is performed at the paragraph-level, however, accuracy significantly increases in respect to the setting in which the training is performed at essay-level. Paragraphs may vary in length: hence at essay-level it may be harder to learn high-frequency positional features related to claims within an entire essay. Finally, in Table 2c we provide a comparison with other baselines found in literature. Since models have been trained in different perspectives (basically, at essay- and paragraph-level), we also report information on training to allow for better comparison.

4.3. Argument Identification through Classification

Finally, as customary, we also performed Argument Identification through Classification. In this case we carried out Argument Classification, and then mapped the tagged classes onto the two required in the Identification task. This technique results in an improvement of the results on the Argument Identification, for both A and O classes, as illustrated in Table 3. Experiments

Table 2
Results on the Argument Classification Task.

(a) Accuracy at token-level (paragraph-level training). The dashed line separates the scores obtained with the BIO tags (7 in total) and the synthetic tags (4 in total).

	P	R	F1	Acc.
MajorClaim B-tag	34.17	61.07	43.67	99.18
MajorClaim I-tag	69.41	68.39	68.59	95.58
MajorClaim tag	67.96	70.61	68.97	95.18
Claim B-tag	31.01	58.16	40.35	98.20
Claim I-tag	59.49	55.93	57.39	87.99
Claim tag	58.09	58.30	57.96	86.88
Premise B-tag	59.70	80.37	68.47	98.05
Premise I-tag	86.48	87.02	86.71	88.37
Premise tag	85.97	88.17	87.02	87.84
Other tag	89.60	84.51	86.95	92.28

(b) Accuracy at α-level 50% and α 100% (paragraph-level training).

	α 50%			α 100%		
	P	R	F1	P	R	F1
Major Claim tag	83.82	72.99	77.81	82.89	52.66	63.90
Claim tag	68.37	59.65	63.44	67.39	46.67	54.69
Premise tag	91.28	86.46	88.76	90.08	70.19	78.85
Other tag	97.72	94.05	95.85	97.68	89.42	93.36

(c) Comparison against the state of the art. For each model at stake we report whether it was trained at the paragraph-level (par), or at the essay-level (ess).

Models	F1 (Precision and Recall)		
	Token-Level	α-Level (50%)	α-Level (100%)
ILP (par) [1]	82.60	73.35	62.61
Potash et al. (par) [14]	84.90	-	-
Mensonides et al. (par) [19]	72.57	-	-
Bao et al. (par) [22]	**88.40**	-	-
Morio et al. (par) [5]	86.82	-	76.48
Morio et al. (ess) [5]	-	-	**76.55**
LSTM-ER (par) [16]	-	77.19	70.83
LSTM-ER (ess) [16]	-	-	66.21
STagBLCC (par) [16]	-	74.08	66.69
STagBLCC (ess) [16]	-	-	63.23
Wang et al. (par) [31]	-	76.7	69.3
ours (ess)	72.06	**79.09**	63.79
ours (par)	75.22	**81.47**	72.70

where training was conducted at paragraph-level show that the accuracy in identification does not benefit from the higher accuracy in the classification task, while the training at essay-level leads to an improvement. This trend is also confirmed in the binary argument identification we performed before: at essay-level, almost all metrics improve for O and A respect to the classic argument identification.

Thus a more fine grained argumentation analysis aimed at identifying the argument components, paired with essay-level training, also improves the general identification (regardless

Table 3

Accuracy on Argument Identification through Classification: we report precision, recall, their even harmonic mean (F1 score), and accuracy, along with the adopted training strategy (ess: at the essay-level) and (par: at the paragraph-level).

Tag	P	R	F1	A
O Tag (Standalone AI) (ess)	**90.98**	83.98	87.33	92.60
O Tag (AI through AC) (par)	89.60	84.51	86.95	92.28
O Tag (AI through AC) (ess)	90.59	**85.71**	**88.08**	**92.96**
A Tag (Standalone AI) (ess)	93.20	**96.39**	94.76	92.60
A Tag (AI through AC) (par)	93.34	95.61	94.50	92.28
A Tag (AI through AC) (ess)	**93.89**	96.12	**94.99**	**92.96**

of argument types). This is probably due to an improved capacity to identify major claims in introduction and conclusion, since now they are explicitly fed to the system with a proper label. This was in fact one of the major sources of errors in Argument Identification.

To sum up, we observe that testing at the essay-level seams to produce improved results for the argument identification, while the paragraph-level for the argument classification.

5. Data Augmentation Through Shuffling

In what follows we report the results of a preliminary experimentation concerning a novel DA technique, specifically performed through sentence shuffling.

By analyzing the misclassified sentences in the previous experiments, we realized that a fraction of such cases were located in recurrent positions in the essays. An hypothesis stemming from this observation is that our model may have been misled by the structural patterns in the input text, rather than focusing on the actual wording and the semantic content of the sentences themselves. In particular, the identification of the major claim (MC) appears to be more difficult, since the introduction and conclusion –where MC is typically located– show greater variability, both in length and structure.

To test this hypothesis, we propose a novel DA technique based on the shuffling of the sentences in the essay. We were of course aware that such process can be highly detrimental to the argument classification task, since reordering statement units might yield as a result an incoherent text containing wrong references and tangled connective expressions [32]. For these reasons two different augmentation approaches were explored. In the *open DA* approach sentences were shuffled at the essay-level, and moved all throughout the essay (in this setting sentences may be moved also outside their original paragraph). On the other hand, in the *closed DA* approach we implemented the shuffling only at the paragraph-level, thereby allowing for more local variations (in this setting sentences cannot be moved outside their original paragraph). In both cases, to systematically analyze how this method impacts on the argument classification task, we have considered different shuffling steps: varying percentages of essays were shuffled, in particular, we used a $+10\%$, $+20\%$, $+30\%$, $+40\%$ and $+50\%$ incremental augmentation. Only sentences in the training set were shuffled (test data were not altered),
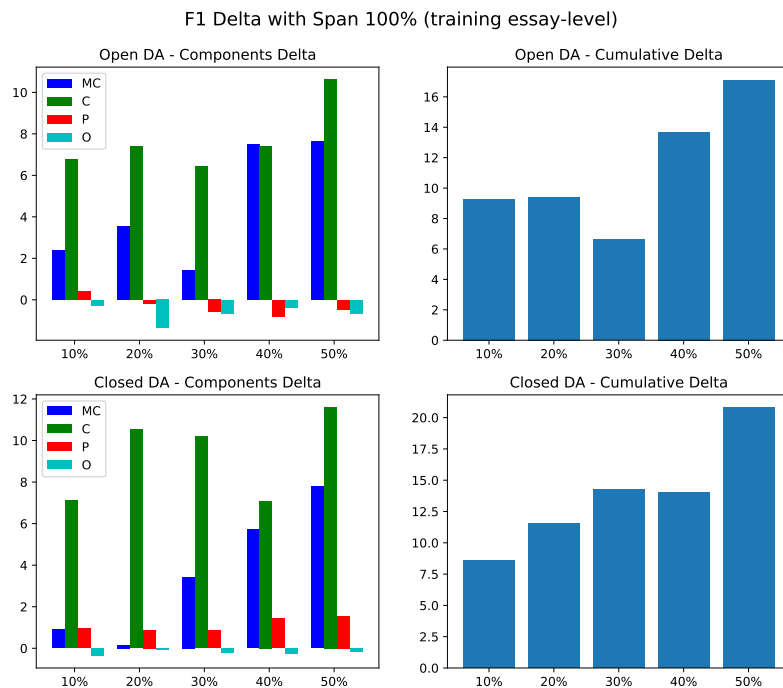
F1 Delta with Span 100% (training essay-level)

Open DA - Components Delta

Open DA - Cumulative Delta

Closed DA - Components Delta

Closed DA - Cumulative Delta

Figure 4: Absolute gain/loss of F1 in ordinates at varying levels of DA (from 0% to 50%), measured at α 100% on the argument classification task (training at essay-level). The zero values for each component are MC: 52.46, C: 33.95, P: 75.29, O: 93.45.

and the training process was performed both at the essay- and paragraph-level; the results were evaluated both through the α-100% and α-50% evaluation metrics, and using the same cross-validation set-up to ensure the comparability to previous results. Furthermore, since the open domain makes no sense when training is conducted at paragraph level, we employed this setting only when training was conducted at essay-level.

Figure 4 illustrates the F1 gain/loss in the argument classification task by varying the percentage of shuffled sentences and measured with α 100%. In this setting the training was performed at essay-level. The base condition is that employing no DA (the corresponding F1 values are reported in the caption). Results for both open and closed approaches were recorded, together with detailed results on each component (MC, C, P and O) on the left; a cumulative sum of all differences between the augmented training and the base condition is presented on the right of the Table. We note a beneficial effect of DA both in the open and closed DA (plots in the right side of the Figure); but the closed domain ensures consistently higher accuracy. By training at essay-level even the best DA approach (closed DA, at 50% DA) yields a F1 score of 68.99, which is still significantly lower than the performance obtained by training at paragraph-level without any DA (72.70). In other words, when training at essay-level, DA is always able to improve performances for 100% α level (compared to the case without DA), but this is not sufficient to
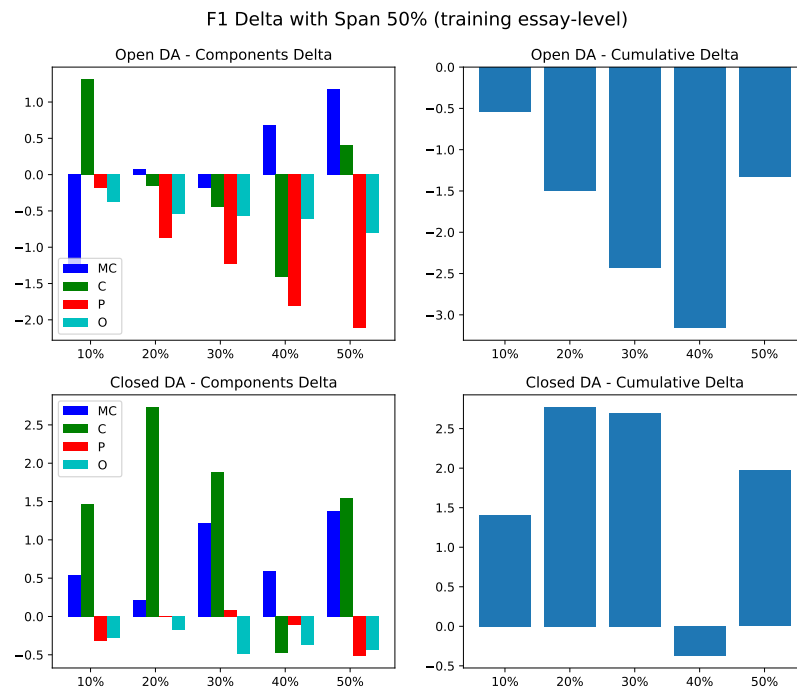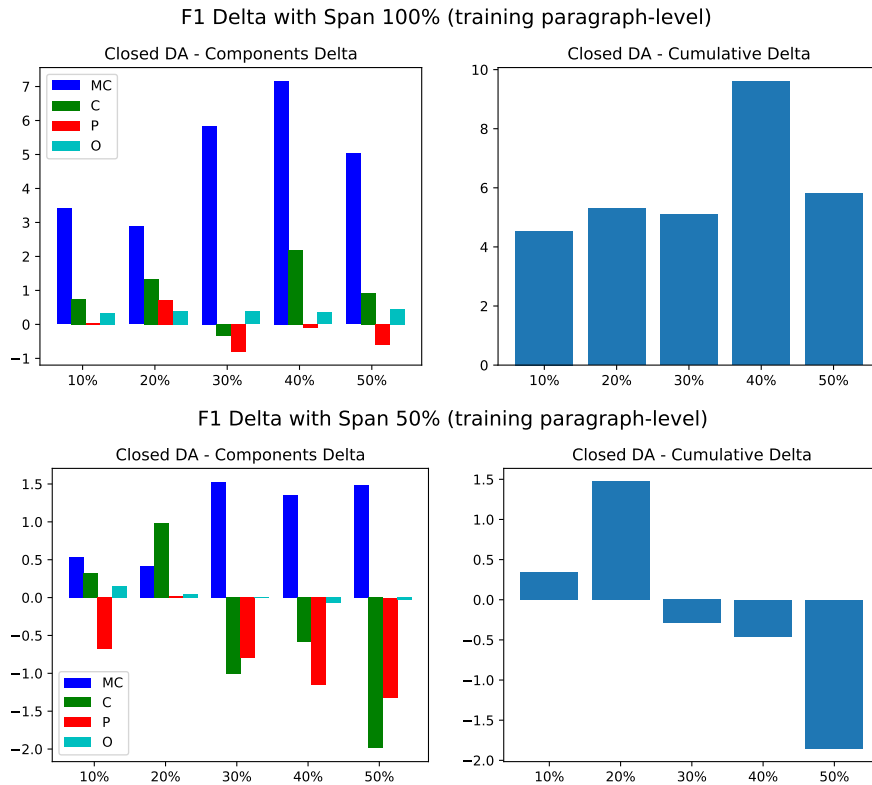
Figure 5: Absolute gain/loss of F1 in ordinates at varying levels of DA (from 0% to 50%), measured at α 50% on the argument classification task (training at essay-level). The zero values for each component are MC: 76.64, C: 55.65, P: 87.78, O: 96.27.

outperform the paragraph-level training.

Figure 5 illustrates the F1 gain/loss in the argument classification task by varying the percentage of shuffled sentences, with results recorded with α 50%. In this setting the training was also performed at essay-level. In this case, the open approach seems to undermine the accuracy of the system, even at a minimum DA of 10%. Using the closed DA and employing a small amount of shuffling seems helpful to preserve the semantics of the inputs: in the conditions with 20% and 30% sentences shuffled we obtained a small though significant improvement in accuracy by applying DA. Such improvements were recorded on claims and major claims. Interestingly enough, any level of open DA appears to be detrimental when evaluating at α level 50%, while it was not at α level 100%. Somehow, open DA creates such a variability beneficial to refine the exact match of some components, but it does not increase the general ability to find components in the $50\%\alpha$ level setting. This might be due to the fact that the exact match yields absolute lower figures which are easier to improve respect to the approximate match. More in general, we observe that testing at paragraph (essay) level after having trained with DA at paragraph (essay) level improves with respect to cases in which training did not make use of DA.

Figure 6 illustrates the F1 gain/loss in the argument classification task by varying the percentage of shuffled sentences and measured with α 100% and α 50%. In this setting the training

Figure 6: Absolute gain/loss of F1 in ordinates at varying levels of DA (from 0% to 50%), measured at α 100% and α 50% on the argument classification task (training at paragraph-level). The zero values for each component at α 100% are MC: 63.91, C: 54.69, P: 78.85, O: 93.36, while at α 50% are MC: 77.81, C: 63.44, P: 88.76, O: 95.85

was performed at paragraph-level and we only tested the closed DA approach (in that shuffling at essay level is pointless when the system is only trained on paragraphs).

For α 100% we obtain the best F1 score of 75.10 with 40% DA and for α 50% the best F1 score of 81.83 is reached with 20% DA; higher levels of DA appear to be detrimental. By focusing on the single components delta (graphs on the left) we can observe that some components (e.g., Major Claims) enjoy greater variability then others (e.g., Premises). In fact, higher percentages of DA allow for the better classification of Major Claims, however, the overall performance is worsened by an increased difficulty in Premises and Claims classification. Since DA was beneficial at essay-level, we also expected them to be beneficial when training at paragraph-level. In fact, using DA when training at paragraph-level lead us to the best results.

6. Conclusions

In this paper we adapted a BERT-based model to perform argument mining; this model obtains competitive results on the standard tasks of argument identification and classification. We have shown that training at essay-level is more suited for argument identification, while training at

paragraph-level produces the best results in argument classification. We then used this model to test a new model-free DA method, developed to mitigate the difficulty to deal with structural variations in the essays during argument classification.

Our approach is based on shuffling sentences and we have explored two approaches: an Open Domain (shuffling sentences within the whole essay) and a Close Domain (shuffling sentences only within paragraphs) and we have experimentally verified that Closed Domain shuffling always provides more accurate results. We have shown that DA techniques are always helpful with the metrics considering exact matches (α 100% condition); in 50% approximate match there is an improvement in F1 score when shuffling is limited to 10-20% sentences. In general we noticed that DA produces improvements in the results, especially in recognizing Claims and Major Claims. Regards as argument classification, DA techniques paired with training at essay-level are not enough to improve on the accuracy obtained by training at paragraph-level. Conversely, DA also improves results obtained through models trained at paragraph-level, leading to the best result of 75.10 (α 100% and +40% DA) and 81.83 (α 50% and +20% DA). Reported figures favorably compare to state-of-the-art results.

Further research is indeed required to assess the robustness of this DA technique, to try to refine shuffling techniques (e.g., by pairing shuffling with dependency parsing information), and to perform experiments on multiple and diverse datasets. Additionally, further developments may be envisaged aimed at applying differing levels of shuffling according to each component.

References

[1] C. Stab, I. Gurevych, Parsing argumentation structures in persuasive essays, Computational Linguistics 43 (2017) 619–659.

[2] R. M. Palau, M.-F. Moens, Argumentation mining: the detection, classification and structure of arguments in text, in: Proceedings of the 12th international conference on artificial intelligence and law, 2009, pp. 98–107.

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

[4] M. Lippi, P. Torroni, Argument mining: A machine learning perspective, in: International Workshop on Theory and Applications of Formal Argumentation, Springer, 2015, pp. 163–176.

[5] G. Morio, H. Ozaki, T. Morishita, K. Yanai, End-to-end argument mining with cross-corpora multi-task learning, Transactions of the Association for Computational Linguistics 10 (2022) 639–658.

[6] A. Hernández-García, P. König, Data augmentation instead of explicit regularization, arXiv e-prints (2018) arXiv–1806.

[7] J. Wei, K. Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks, arXiv preprint arXiv:1901.11196 (2019).

[8] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, arXiv preprint arXiv:1511.06709 (2015).

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

[10] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for nlp, arXiv preprint arXiv:2105.03075 (2021).

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

[12] C. Stab, I. Gurevych, Identifying argumentative discourse structures in persuasive essays, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 46–56.

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

[14] P. Potash, A. Romanov, A. Rumshisky, Here's my point: Joint pointer architecture for argument mining, arXiv preprint arXiv:1612.08994 (2016).

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

[16] S. Eger, J. Daxenberger, I. Gurevych, Neural end-to-end learning for computational argumentation mining, arXiv preprint arXiv:1704.06104 (2017).

[17] M. Miwa, M. Bansal, End-to-end relation extraction using lstms on sequences and tree structures, arXiv preprint arXiv:1601.00770 (2016).

[18] Y. Ye, S. Teufel, 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, 2021, pp. 669–678.

[19] J.-C. Mensonides, S. Harispe, J. Montmain, V. Thireau, Automatic detection and classification of argument components using multi-task deep neural network, in: 3rd International Conference on Natural Language and Speech Processing, 2019.

[20] T. Wambsganss, N. Molyndris, M. Söllner, Unlocking transfer learning in argumentation mining: a domain-independent modelling approach, in: 15th International Conference on Wirtschaftsinformatik, 2020.

[21] T. Chakrabarty, C. Hidey, S. Muresan, K. McKeown, A. Hwang, Ampersand: Argument mining for persuasive online discussions, arXiv preprint arXiv:2004.14677 (2020).

[22] J. Bao, C. Fan, J. Wu, Y. Dang, J. Du, R. Xu, 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), 2021, pp. 6354–6364.

[23] T. Alhindi, D. Ghosh, " sharks are not the threat humans are": Argument component segmentation in school student essays, arXiv preprint arXiv:2103.04518 (2021).

[24] J. W. G. Putra, S. Teufel, T. Tokunaga, Parsing argumentative structure in english-as-foreign-language essays, in: Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, 2021, pp. 97–109.

[25] P. Liu, X. Wang, C. Xiang, W. Meng, A survey of text data augmentation, in: 2020 International Conference on Computer Communication and Network Security (CCNS), IEEE, 2020, pp. 191–195.

[26] C. Shorten, T. M. Khoshgoftaar, B. Furht, Text data augmentation for deep learning, Journal of big Data 8 (2021) 1–34.

[27] M. Bayer, M.-A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, C. Reuter, Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers, International Journal of Machine Learning and Cybernetics (2022) 1–16.

[28] O. Toledo-Ronen, M. Orbach, Y. Bilu, A. Spector, N. Slonim, Multilingual argument mining: Datasets and analysis, arXiv preprint arXiv:2010.06432 (2020).

[29] X. Yang, J. Bian, W. R. Hogan, Y. Wu, Clinical concept extraction using transformers, Journal of the American Medical Informatics Association 27 (2020) 1935–1942.

[30] L. A. Ramshaw, M. P. Marcus, Text chunking using transformation-based learning, in: Natural language processing using very large corpora, Springer, 1999, pp. 157–176.

[31] X. Wang, Y. Lee, J. Park, Automated evaluation for student argumentative writing: A survey, arXiv preprint arXiv:2205.04083 (2022).

[32] R. Iida, T. Tokunaga, Building a corpus of manually revised texts from discourse perspective, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014, pp. 936–941.