When Argumentation Meets Cohesion: Enhancing Automatic Feedback in Student Writing

Yuning Ding[1], Omid Kashefi[2], Swapna Somasundaran[2], Andrea Horbach[1,3]
[1]CATALPA, FernUniversität in Hagen, Germany
[2]Educational Testing Service (ETS),Princeton, NJ, USA
[3]Hildesheim University, Germany
{yuning.ding, andrea.horbach}@fernuni-hagen.de
{okashefi,ssomasundaran}@ets.org

Abstract

In this paper, we investigate the role of arguments in the automatic scoring of cohesion in argumentative essays. The feature analysis reveals that in argumentative essays, the lexical cohesion between claims is more important to the overall cohesion, while the evidence is expected to be diverse and divergent. Our results show that combining features related to argument segments and cohesion features improves the performance of the automatic cohesion scoring model trained on a transformer. The cohesion score is also learned more accurately in a multi-task learning process by adding the automatic segmentation of argumentative elements as an auxiliary task. Our findings contribute to both the understanding of cohesion in argumentative writing and the development of automatic feedback.

Keywords: Document Classification, Text Categorisation, Tools, Systems, Applications

1. Introduction

Argumentative writing is an important task type in education since it encourages critical thinking and civic participation (Andrews, 2009). By their nature, argumentative essays require a high degree of cohesion, defined as a network of semantic relationships that link together (Hasan, 2014), making an essay easier to follow and understand.

In recent years, the development of natural language processing (NLP) has opened new avenues for automating the assessment of text cohesion. Cohesion is usually measured by the presence or absence of certain linguistic cues that enable the reader to establish connections between the ideas within the text (Crossley et al., 2016a). These cues include explicit transitional words such as *because* (Hasan, 2014) and chains of related words that contribute to the continuity of lexical meaning (Morris and Hirst, 1991). These lexical chains are, for instance, the explicit chain "home"→"home" and the implicit chain "class"→"seminar" in the following example:

> *Students would benefit from attending **classes** from **home**. For example, they could join **seminar** online or do the assessment at **home**.*

Related work shows that including features like the number of transitional words (Chiang, 2003) and lexical chains (Somasundaran et al., 2014) improve the performance of automatic scoring of cohesion.

The recently released datasets in the Kaggle competition series "Feedback Prize"[1] written mostly by English native writers (L1), have both cohesion score and annotation of different argumentative elements. It provides us with a chance to assess cohesion directly linked to how well the argument is structured and how smoothly it progresses.

Figure 1 shows two example essays from the datasets, arguing whether the students should be able to attend classes from home. The font in different colors illustrates four categories of argumentative elements, including the Position, which is an opinion on the main question, the Claim that supports the position, the Evidence as ideas or examples that support claims and the Conclusion that restates the claims. Both essays contain cohesion features, including the highlighted transitional words and lexical chains, but have different cohesion scores. According to the scoring rubrics (1-5) provided by the competition host, getting a cohesion score of 4 means that:

> The organization generally well controlled, a range of cohesive devices used appropriately such as reference and transitional words and phrases to connect ideas and generally appropriate overlap of ideas were found in this essay.

On the other hand, an essay with a cohesion score of 2 has the following characteristics:

[1]https://www.kaggle.com/competitions?searchQuery=feedback+prize

LREC-COLING 2024, pages 17513–17524
20-25 May, 2024.

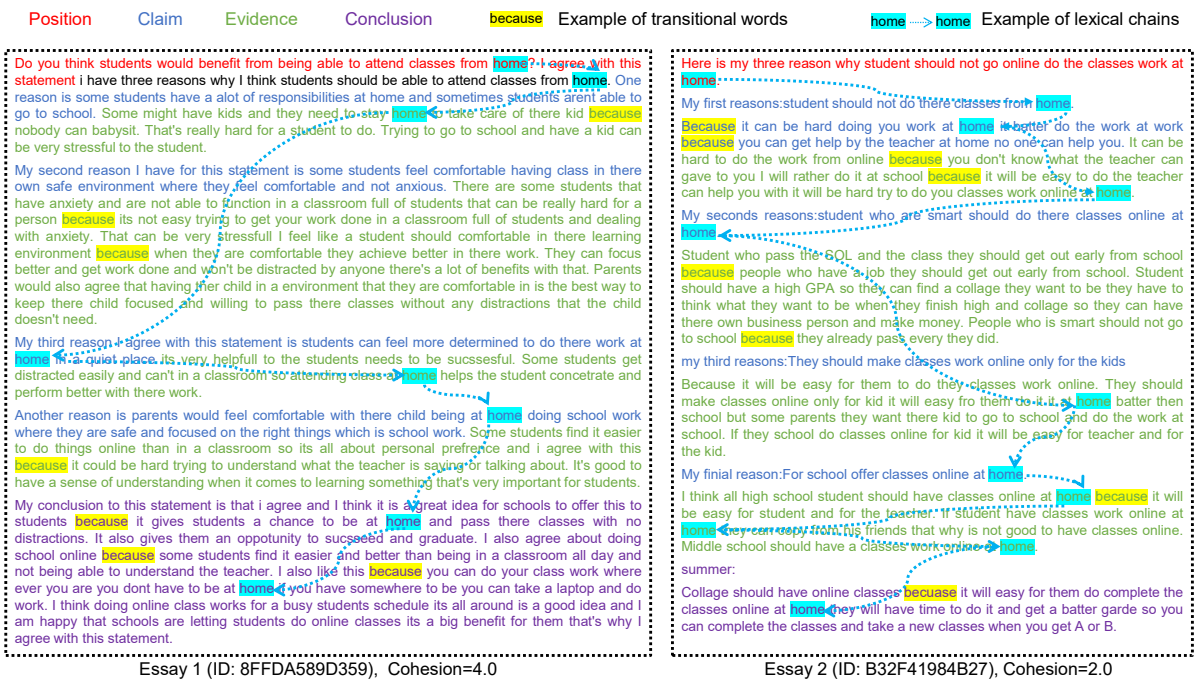

Figure 1: Example essays with argumentative elements, transitional words, and lexical chains. Text in different colors stands for four types of argumentative elements. The transition words are highlighted in yellow, while the lexical chains of the word "home" are illustrated with blue blocks and arrows.

> The organization was only partially developed with a lack of logical sequencing of ideas and some basic cohesive devices are used but with inaccuracy or repetition.

The argument features, i.e., the counts of different argumentative elements in these two essays, are the same. Therefore, using these features directly to predict the cohesion score is probably not a good idea. However, argumentative elements carry semantic and rhetorical information. By providing the argument features, we expect the scoring models to better understand the meaning and organization of different ideas. This may enhance semantic understanding and can aid in recognizing cohesive features within argumentative essays. The open question here is **how** should argument features be provided.

Besides adding the absolute frequency of each argumentative element as features directly, we could also combine them with cohesion features, such as the transitional words and lexical chains mentioned above. The number of transition words and lexical chains could be extracted from the essays or in different argumentative elements and then concatenated to the argument features. Both methods could provide the model with cohesion and argument features. In the context of increasingly used multi-task learning in automatic essay scoring, we also propose training a model to segment argumentative elements as an auxiliary task. Predicting argumentative elements forces the model to better understand the structure and organization of argumentative essays. This understanding might help the model grasp the logical flow of ideas while identifying claims and evidence, which is crucial for cohesion. In summary, the following research questions are investigated in this paper:

RQ. 1 *To what extent does adding argument features enhance the performance of predicting cohesion scores in the context of automatic cohesion scoring?*

RQ. 2 *How does the inclusion of cohesion features, specifically transitional words and lexical chains, among argumentative elements impact the performance of automatic cohesion scoring when compared to the inclusion of these features alone?*

RQ. 3 *Does employing the automatic segmentation of argumentative elements as an auxiliary task result in better performance for predicting cohesion scores compared to the single-task approach?*

We answer these questions through two studies. In *Study 1*, we analyze the correlation between the cohesion score and different features. Results show that the lexical chain features extracted from argumentative elements correlate with the cohesion score. Secondly, we score the cohesion using logistic regression and large pre-trained language models. By including cohesion features, argument features, and their combination, we show that the combined features benefit the prediction performance of a transformer-based scoring model.

In *Study 2*, we learn the segmentation of argumentative elements as an auxiliary task of automatic cohesion scoring. We find that multi-task learning settings perform better than the single-task setup.

Having addressed these research questions, our work contributes to advancing both theoretical understanding and practical applications in the field of writing evaluation and education. To our knowledge, our study is the first to combine argumentation analysis with cohesion scoring in argumentative essays. In practical terms, our model's ability to assess cohesion and comprehend argumentative structures showcases its potential for providing feedback to learners.

2. Related Work

Starting with Page's seminal paper (Page, 1966), the field of automatic essay scoring has remained active for more than 50 years. Numerous studies have examined the subject of automatic essay scoring through comprehensive literature reviews Ke and Ng (2019); Beigman Klebanov and Madnani (2020). In this paper, we limited our related work to the automatic scoring of cohesion, argument mining in essays, and multi-task learning in the educational domain. Please note that instead of aiming to achieve state-of-the-art results in essay scoring, our goal is to fill the gap in the current literature by highlighting the significance of argument structure in understanding essay cohesion.

2.1. Automatic Cohesion Scoring

For scoring a particular dimension of the quality of the essay, such as cohesion, corpora with expert dimension-specific scores are essential. The cohesion score, as cited in the rubrics in Section 1, is also reflected in the rubrics named *coherence* and *organization* in other corpora. Table 1 lists these corpora, including their language, number of essays, essay type, writer's background, and score range. In our study, we use the ELLIPSE dataset released from the Kaggle competition, which is further described in Section 3.

Targeting a better performance of scoring the holistic score or giving specific feedback, many automatic essay scoring systems contain specific feature sets or rules to capture coherence and cohesion. Miltsakaki and Kukich (2004) found that adding the *transition features* defined in the Centering Theory (Grosz et al., 1995) to the e-Rater essay scoring system e-Rater (Attali and Burstein, 2006) contribute significantly to the scoring performance of learner texts. Burstein et al. (2010) explored how the *entity-grid* can be used to discriminate learner texts with different coherence. It is also tested in Yannakoudakis and Briscoe (2012)'s Support Vector Machine (Cortes and Vapnik, 1995) based system along with other features such as *'superficial' cohesive features* including *Part-of-Speech (POS) distribution, discourse connectives, word length* and *semantic similarity*. *Lexical chains related features* (LEX-1, LEX-2 in Somasundaran et al. (2014)) were also successfully used as an indicator of text cohesion, which are further investigated in our paper together with *transitional words* and their combination with argument features.

It should be noted that general automatic tools for analyzing text cohesion, such as Coh-Metrix (Graesser et al., 2004) and TAACO (Crossley et al., 2016b), fall outside the scope of our related work. However, some scoring systems use these tools to extract cohesion-related features for scoring, such as Crossley et al. (2013). They extracted a number of features (i.e., incidence of all connectives or conjuncts) from the Coh-Metrix and found that indices of global cohesion were significant predictors of the quality of the essay.

Besides the traditional feature engineering approaches, Ramesh and Sanampudi (2022) proposed a sentence-based embedding for training on the Long Short-Term Memory network (LSTM) and the Bidirectional LSTM (BiLSTM) to capture a text's coherence and cohesion. Tay et al. (2018) proposed *neural coherence features*, which is the similarity of a pair of positional outputs, which are collected from different time steps from an additional layer in their LSTM scoring model.

2.2. Argument Mining in Essays

The most widely accepted model of arguments in student essays is a variant of Toulmin's argument model (Toulmin, 1958) proposed by Stab and Gurevych (2014b), consisting of four categories: *major claim, claim, premise* and *non-argumentative*. Their annotation guidelines led to a high level of agreement in an annotation study conducted on 90 persuasive essays in English (Stab and Gurevych, 2014a). Persing and Ng (2015) extended this framework to annotate the International Corpus of Learner English (Granger et al., 2009). They then developed classification models to recognize argument components and incorporated them as features to predict argumentative scores in essays (Persing and Ng, 2016).

The dataset for argument mining used in this work, i.e., the FB-Arguments described in Section 3, uses an adaptation of the Toulmin model (Nussbaum et al., 2005; Stapleton and Wu, 2015) with seven argumentative elements, namely *lead, position, claim, counterclaim, rebuttal, evidence, and concluding statement*. Ding et al. (2022) trained a sequence tagging model using a pre-trained Longformer (Beltagy et al., 2020) on different subsets of this corpus and reported an F1 score of .55. Their

Corpus	Lan.	# Essays	Type	Writers	Score	Range
ICLE, created by Granger et al. (2009), subset scored by Persing et al. (2010)	EN	1,003	argumentative	university undergraduates	Organization	1-4
SkaLa (Horbach et al., 2017)	DE	2,020	summary&discussion	university undergraduates	Coherence	1-6
ASAP++(prompt 1, 2) (Mathias and Bhattacharyya, 2018)	EN	3,585	argumentative	students in grades 7 to 10	Organization	1-6
Essay-BR (Marinho et al., 2021)	PT	4,570	argumentative	high school students	Cohesion & Coherence	0- 200
Song et al. (2020)	ZH	1,220	argumentative	high school students	Organization	Great/Medium/Bad
ELLIPSE (The Learning Agency Lab, 2022)	EN	3,911	argumentative	students in grades 8 to 12	Cohesion	1-5

Table 1: Overview of corpora with cohesion-related scores.

framework was used in our experiments to extract argument features.

Based on the argument components, argumentation features are computed and used in automatic essay scoring. Ghosh et al. (2016) proved that argumentation features related to argument components (e.g., the number of claims and premises), argument relations (e.g., the number of supported claims), and the typology of argumentative structure (chains, trees) are good predictors of holistic scores of persuasive essays. Wachsmuth et al. (2016) and Nguyen and Litman (2018) also improved the system performance on scoring an essay's organization and argument strength by adding features capturing the composition of types of units in arguments. However, to our knowledge, there has been no research on using argumentative features to better predict cohesion scores.

2.3. Multi-Task Learning

Multi-task learning involves methods that simultaneously learn various tasks using the same dataset and a unified loss function. For instance, a multi-task BiLSTM was employed in previous research by Rei (2017) to train on grammatical error detection and automated essay scoring tasks. In their experiments, essay scoring performance was improved by error detection, while the error detection task did not.Another variation of hierarchical BiLSTMs was used by Song et al. (2020) to identify discourse elements and evaluate the organization of Chinese student argumentative essays. This joint model achieved better performance in both tasks. Muangkammuen and Fukumoto (2020) took a similar approach by combining word and sentence-level BiLSTMs into a hierarchical model for predicting essay scores and sentiment classes of individual words. Their result showed that sentiment analysis was beneficial for essay scoring. In line with this methodology, we employ diverse annotations on the same essays and define our joint tasks as detecting argument spans, predicting their type, and assessing their quality. Ding et al. (2023) trained a model for automated argument detection, classification, and effectiveness prediction, outperforming sequential approaches. Their multi-task learning architecture was modified and used in our study.

3. Datasets

In this study, we used two distinct data sets, both originating from Kaggle competitions: "Feedback Prize - Evaluating Student Writing" (FB-Arguments) and "Feedback Prize - English Language Learning" (FB-Score). These datasets represent subsets of the broader PERSUADE corpus (Crossley et al., 2022) and the ELLIPSE corpus, respectively, as detailed in the descriptions provided by The Learning Agency Lab[2].

The FB-Arguments dataset comprises a collection of 15,600 argumentative essays written by students in grades 6-12 in the United States. These essays contain expert annotation encompassing seven categories of argumentative elements, annotated with an overall inter-rater reliability of .73. An expert rater adjudicated all disagreements. Based on the definition of these elements, we merge the counterclaim and rebuttal into the claim and get a more simplified categorization with five labels, namely *Lead, Position, Claim, Evidence* and *Conclusion*. The resultant taxonomy facilitated the training of an argument labeling model, enabling us to extract features related to arguments in the essays.

The FB-Score dataset has 3.911 argumentative essays crafted by English language learners in grades 8-12. These essays were annotated by human raters using a five-point scoring rubric that comprised both holistic and analytic scales, including cohesion, syntax, phraseology, vocabulary, grammar, and conventions. Two raters scored each essay. If their scores differed by 1, the final score would be the average of the two scores. For score differences greater than 1, the raters would discuss the difference and come to a final adjudicated score. The exact inter-rater agreement was not reported but described as "strong"[3]. Our research predominantly focuses on the cohesion scores, which range from 1.0 to 5.0 in increments of 0.5.

As shown in Table 2, there is an intersection between the datasets above (numbers in green cells), which encompasses a subset of 452 essays.

[2] https://the-learning-agency-lab.com/the-feedback-prize-overview/

[3] https://www.kaggle.com/competitions/feedback-prize-english-language-learning/discussion/348973#1936418

	Cohesion Score									Only in FB-Arg.
	1.0	1.5	2.0	2.5	3	3.5	4.0	4.5	5.0	
#Essays	**0**	**4**	**28**	**67**	**160**	**119**	**60**	**11**	**3**	**15,148**
Lead	0	1	7	25	97	72	42	7	2	9,060
Position	0	3	17	51	126	79	39	7	2	14,967
Claim	0	6	58	175	433	313	161	31	10	58,394
Evidence	0	15	71	207	477	348	168	37	11	44,315
Conclusion	0	4	21	56	151	114	63	11	2	13,124
#Essays only in FB-Score	**10**	**23**	**287**	**723**	**936**	**869**	**474**	**114**	**23**	- -

Table 2: Data distribution in FB-Arguments and FB-Score. Green fields stand for the intersection.

These essays are unique because they possess gold-standard annotations for argumentative elements and cohesion scores. This subset is our test set for argument labeling and essay scoring models. We noticed no essay with a cohesion score of 1.0 in this test set. The remaining essays only in the FB-Score (in yellow cells) were allocated to distinct validation and training sets for cohesion model training, distributed according to a 9:1 ratio. With the essays only in FB-Arguments (in blue cells), we trained an argument mining model to extract argument features.

4. Study 1: Scoring of Cohesion

This study first establishes three scoring baselines in Section 4.1 utilizing both linear regression and pre-trained deep learning models. Subsequently, we introduce and analyze the cohesion features and argument features in Section 4.2 and apply them to our baseline systems. The experiment results are discussed in Section 4.3.

4.1. Baseline Setup

Our Linear Regression baseline comes from scikit-learn (Pedregosa et al., 2011) using 500 term frequency–inverse document frequency (tf-idf) weighted uni- to trigrams. In the deep learning approaches, we modified the pre-trained BERT[4] and Longformer[5] models with a regression head, then trained them for 10 epochs with an Adam optimizer (Kingma and Ba, 2014) and Mean Squared Error (MSE) as the loss function. The max length of BERT is 512, and its batch size is 8, while the Longformer was trained with a max length of 1024 and a batch size of 2. For this study, all models were trained in 32 hours on a single GPU. The original regression output is evaluated with MSE, and the rounded classification results are evaluated with the F1 score and the Quadratic Weighted Kappa (QWK).

[4] https://huggingface.co/bert-base-uncased

[5] https://huggingface.co/allenai/longformer-base-4096

4.2. Additional Features

Based on the discussion in Section 1, we look at four types of features in student essays that may be related to the cohesion scores:

Our first set of features is the argument features (**ARG**). It includes ten features related to the argumentative elements. The features are the absolute and relative frequencies of the five types of argumentative elements (lead, position, claim, evidence, and conclusion) found in the essays. These features come from a trained argument mining model. To obtain this model, we use only the essays from the FB-Arguments dataset (as shown in the blue cells in Table 2). This model achieves a performance of F1 score at .66 on the test data with both gold-standard annotation of arguments and scores (green cells in Table 2), taking the predictions as having an overlap of more than 50% with the ground truth as true positive, unmatched predictions as false positive, and unmatched ground truths as false negative.

The **Transitional Words** feature set contains the absolute and relative frequencies of the connective words of different transitions, such as “so that” for the transition *belief*, “according to” for *consequence*, the aforementioned “because” for *evidence* and so on. These features are extracted by a feature service described in Madnani et al. (2018).

This service also provides us with the **LEX-1** feature set, which includes 38 features. It covers the counts, corresponding percentages, and normalized versions of different types of lexical chains, such as the total number of chains in the essay, the average chain size, and the number of large chains (lexical chains with more than four nodes). To reflect the diversity of the text, additional features such as the count (and corresponding percentage) of chains containing more than one word/phrase type were also employed. Furthermore, the nature of the links between chains was also considered. It encompasses various aspects, such as the count and proportion of each link type (extra strong, strong or medium) and their distribution within both large and small chains.

The **LEX-2** feature set captures the interactions between discourse transitions using a discourse cue tagger described in Burstein et al. (2001). With the help of patterns and syntactic rules, the tagger automatically detects words and phrases functioning as indicators within the discourse and assigns them a specific discourse tag. Each tag consists of two parts: a primary component that indicates whether an argument or topic is being introduced (arg-init) or elaborated (arg-dev), and a secondary component specifying the precise type of discourse introduction (e.g., CLAIM, SUMMARY) or development (e.g., CLAIM, CONTRAST). Each discourse cue was replaced with its tag, and the number of

Feature Set	Average $\lvert r\rvert$	Max. Positive Correlated Feature	r	Max. Negative Correlated Feature	r
ARG	**.114**	Absolute Frequency of Concluding Statement	.201	Relative Frequency of Evidence	-.175
Transitional Words	.004	Absolute Frequency of Transitional Words	.005	Relative Frequency of Transitional Words	-.003
TW in Claims	**.112**	Absolute Frequency of Transitional Words	.123	-	-
TW in Evidence	**.120**	-	-	Relative Frequency of Transitional Words	-.125
LEX 1	.066	Percentage of Large Chains with Extra Strong Links	.135	Number of Large Chains with Medium Links	-.160
LEX 1 in Claims	.092	Total Number of Links	**.208**	Percentage of Small chains	-.027
LEX 1 in Evidence	.084	Number of Small chains with a variety of words	.092	Total Number of Links in Large Chains	**-.260**
LEX 2	.051	Number of Chains start after argument expression	.174	Percentage of Chains continue over an initialized argument	-.200
LEX 2 in Claims	.031	Number of Chains end before argument expression	.199	Number of Chains start after an initialized argument	-.040
LEX 2 in Evidence	.045	Percentage of Chains start after an developed detail	.104	Number of Chains continue over an initialized argument	-.217

Table 3: Partial correlation (r) between additional features and cohesion score with the effect of essay length being controlled. $|r|$ is the absolute value of r.

chains that (i) start after it, (ii) end before it, and (iii) continue over it (chains having nodes before and after the tag) was counted as features, resulting in a total of 138 features.

Based on Toulmin's argument model (Toulmin, 1958), we projected the Position, Claim and Concluding Statement into Claims, then extracted the transitional words and lexical chain features among Claims and Evidence separately. It is considered another method to combine arguments and cohesion features together. The feature sets are noted as **X in Claims** or **X in Evidence**.

These datasets are first investigated in a correlation analysis using the Pearson correlation coefficient (r). Table 3 shows the *partial correlation* (Baba et al., 2004) between additional features and cohesion score with the effect of essay length being controlled. The correlation results show that the argument feature set has one of the greatest three absolute correlation coefficients ($r = .114$) to the cohesion score. This answers the **RQ. 1** to a certain extent, that argument features are related to cohesion score in argumentative essays.

The vanilla feature sets have nearly no correlation to the cohesion score for the transitional words. However, the correlations are slightly improved once we extract them separately in claims and evidence. The same phenomenon can be observed on the LEX 1 feature set but not on LEX 2. By further looking into the most positive and negative correlated features, we notice that the total counts of lexical chains found in claims and total counts of links in large chains found in evidence correlate most to the cohesion score. These results indirectly answer our second research question (**RQ. 2**): considering cohesion features in different argument elements is more helpful for cohesion scoring than considering these characteristics alone.

The most positive and negative correlated features also demonstrate that in argumentative essays, the lexical connection among claims holds greater significance for overall cohesion, with diversity and divergence expected in the evidence. This conclusion is consistent with our intuition and can also be observed in students' essays. Looking back to the example essays in Figure 1, we could see more lexical overlaps in the position, claims, and conclusions in Essay 1 compared to Essay 2.

However, it is too early to conclude since the linear correlation is overall not strong. We must add these features to our scoring baseline models to answer our research question better. Besides the feature sets described above, we also include **LEX** that combines LEX-1 and LEX-2. For the linear regression model, the features were concatenated to the tf-idf vectors. We passed the features through a feature network into a dense feature representation and concatenated the essay embedding for the deep learning models. The combined representation was used to train the classification layer and predict the cohesion score.

4.3. Results

Table 4 shows the results of the automatic scoring of cohesion with additional features. For deep learning approaches, we reported the evaluation of test predictions made by the models with the lowest MSE on validation data among ten training epochs. Comparing these baselines, we see that the BERT model has the best F1 and QWK scores, although the Longformer makes the predictions with lower MSE. The argument features (+ARG) can only improve the F1 score of the linear regression model by .01 but decrease the F1 scores of deep learning models. It is aligned with our observation in the examples that essays with different cohesion scores may contain a similar number of different argumentative types. Therefore, adding argumentative features directly into the scoring model may not help to improve the cohesion scoring.

Comparing the results on the other bold lines (+Transitional Words, +LEX-1 etc.), we see that adding cohesion-related features is not always beneficial, at least not by adding them alone. For the effect of adding the combination of cohesion and argument features, we expect larger green numbers on the second to the fourth lines than on the first line

Setting	Linear Regression			BERT			Longformer		
	MSE	F1	QWK	MSE	F1	QWK	MSE	F1	QWK
Baseline	.33	.18	.43	.23	**.28**	.57	.21	.21	.54
+ARG	+0	+.01	+0	-.01	-.07	+.01	+0	-.03	+.03
+ Transitional Words	+0	+0	-.01	+.02	-.03	+0	+.02	+.04	+.06
+TW +ARG	+0	+0	+.02	-.01	-.09	-.02	+.01	+.10	+.05
+TW in Claims	+0	+.01	+.01	+0	-.05	+0	+0	+.07	+.04
+TW in Evidence	+0	+.01	+0	+0	-.05	+0	+0	+.08	+.05
+LEX-1	+0	+.01	+.01	-.01	-.07	-.01	-.01	+0	+.06
+LEX-1 +ARG	+.01	-.01	+.01	+0	-.07	-.01	+.01	+.01	+.05
+LEX-1 in Claims	+0	+.01	-.01	-.01	-.06	-.02	+.01	**+.11**	+.05
+LEX-1 in Evidence	-.01	-.01	+0	-.01	-.06	-.02	+0	+.07	+.08
+LEX-2	+.02	+0	+0	-.01	-.03	+.02	+0	+.05	+.03
+LEX-2 +ARG	+.01	-.01	+0	+0	-.04	+.01	+0	+.10	+.05
+LEX-2 in Claims	+.02	+.01	-.01	+0	-.08	-.02	+0	+.03	+.04
+LEX-2 in Evidence	+.02	+.02	-.01	+0	-.03	+0	+.01	+.10	+.04
+LEX	+.01	+0	+.02	+0	-.07	-.01	+0	-.01	+.02
+LEX +ARG	+.01	-.01	+0	+0	-.06	-.01	+0	+.08	+.07
+LEX in Claims	+.03	+.02	-.02	+0	-.08	-.03	+0	+.02	+.03
+LEX in Evidence	+.02	+.01	-.01	+0	-.08	-.03	+0	+.03	+.04
+TW +LEX-1	+0	+.01	+.01	-.01	-.06	+.01	+0	+0	+.04
+TW +LEX-1 +ARG	+.01	-.01	+.01	-.01	+.02	+.03	+0	+.01	+.02
(+TW +LEX-1) in Claims	+0	+.01	-.01	-.01	-.09	+.01	+0	+.09	+.06
(+TW +LEX-1) in Evidence	+0	-.01	+.01	-.01	-.08	+.01	+0	+.10	**+.10**
+TW +LEX-2	+.01	+0	+0	+0	+0	+.01	+0	-.05	-.03
+TW +LEX-2 +ARG	+.01	+0	-.01	+0	-.09	-.01	+0	+.10	+.05
(+TW +LEX-2) in Claims	+.02	+.02	+0	+.01	-.08	-.01	+0	+.08	+.06
(+TW +LEX-2) in Evidence	+.02	+.02	+.01	-.01	-.08	-.01	+.01	+.07	+.08
+TW +LEX	+.01	+0	+0	+.02	+0	-.05	+0	-.02	-.06
+TW +LEX +ARG	+.02	-.01	+0	-.01	-.08	-.03	+0	+.04	+.05
(+TW +LEX) in Claims	+.03	+.02	-.03	+.01	-.09	-.02	-.01	+.02	+.07
(+TW +LEX) in Evidence	+.02	+.01	-.01	+.01	-.09	-.01	-.01	+.08	+.08

Table 4: Automatic cohesion scoring with additional arguments (ARG), transitional words (TW), and lexical chain (LEX) features in different combinations. The green numbers show an improvement in the scoring performance compared to the baselines, while the red ones show a decrease.

in the seven four-line groups. This is seen as a pattern in the results of the Longformer model. Adding the LEX-1 features found in claims achieves the best F1 score at .32, while the Transitional Words and LEX-1 features found in evidence help it get the best QWK score at .64. It proves our hypothesis that adding cohesion features among different argumentative elements contributes more than adding cohesion features alone in automatic cohesion scoring. Unfortunately, we did not see the same pattern on the other two models. The only feature set that benefits the BERT model on all the evaluation metrics is the transitional words and LEX-1 found in claims. We cannot see any improvements in the linear regression model by adding feature combinations.

Through the observation of classification of essays in each cohesion group, we notice that our Longformer baseline model has difficulty with the lower-frequency classes. The models with adding LEX-1 and Transitional Words found in Claims and Evidence as additional features are relatively better at handling these classes. A possible explanation of the better performance of BERT than Longformer is the Next Sentence Prediction (NSP) objective of BERT. Through this objective, BERT indirectly learns the cohesion and argumentative flow of text, potentially resulting in a more effective representation (Devlin et al., 2018).

5. Study 2: Multi-Task Learning

Since adding the combined feature improves the performance of cohesion scoring on Longformer-based model, we further experiment with another method to learn the argumentative features. In Section 5.1, we integrate the segmentation of argumentative elements and cohesion score prediction into a multi-task learning framework with two different weighting strategies. The results are discussed in Section 5.2.

5.1. Experimental Setup

In the multi-task learning approach, we modified the sequence tagging architecture in Ding et al. (2023) as shown in Figure 2, which trains the input essay with Longformer and labels the sequence output of each token with Inside–Outside–Beginning (IOB) -Tags for the argument elements by a classification head, such as “B-Claim”, “I-Claim” or “O”, and the pooled output of each essay with a cohesion score by a regression head. The predicted IOB-Tags are merged into token indexes for each argumentative element. The cohesion score is rounded to 1.0 to 5.0 in increments of 0.5, exactly the same as the setting described in Section 4.1. For this study, all models were trained for about 8 hours on a single GPU.

For the argument mining task, we see the performance of the model trained on FB-Arguments dataset with gold standard annotations of argumentative elements as the upper bound. As mentioned in Section 4.2, it is evaluated by F1 score at .66. However, for the cohesion scoring, we only have the smaller FB-Score dataset with gold standard annotations of cohesion score and the predicted argumentative elements. Therefore, we set the all cohesion score to 0 and get a baseline with an F1 score of .63.

Meanwhile, we set all the argument labels as “O” and let the model only learn the cohesion score. This baseline performance of cohesion scoring is measured with F1 score (.21) and QWK (.54), which is aligned with the baseline Longformer results reported in Section 4.1. We see the Longformer+ARG as another baseline for cohesion scoring since it learns the argumentative features and the pre-trained embeddings at the same time, which corresponds to the design of our multi-task learning approach.

For multi-task learning, we also experiment with two different settings of weighting strategy. In the **linear** setting, the total loss is a simple average of the sum of the losses for each task:

$$L_{total} = \frac{1}{t}\sum_{t} L_t \quad (1)$$

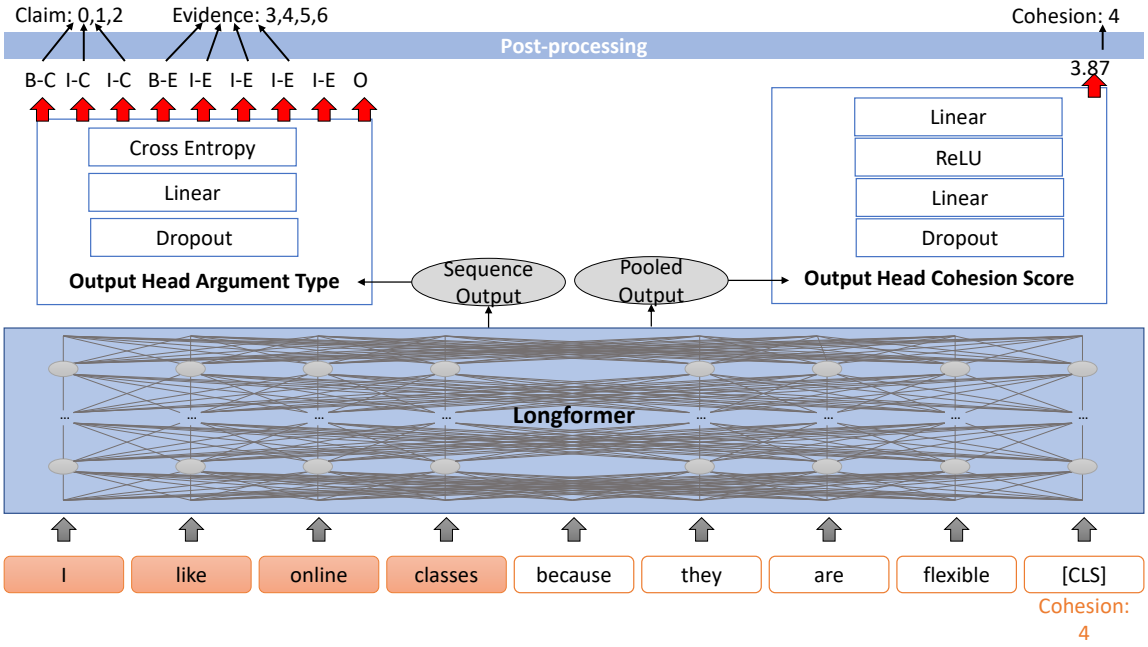

Figure 2: Multi-task learning architecture of argument mining and cohesion scoring.

where L stands for the loss, and t denotes different tasks. The **dynamic** weight average calculates multiple losses by considering the homoscedastic uncertainty of each task (Kendall et al., 2018) as shown in Equation 2, where σ denotes the loss variation of each task (default as 0.5).

$$L_{total} \approx \sum_{t} (L_t \frac{1}{2\sigma_t^2} + log\sigma_t) \quad (2)$$

5.2. Results

As shown in Table 5, our multi-task learning approach with linear weighting strategy achieves the best performance on cohesion scoring while preserving the performance of argument mining within predicted spans. The weighted strategy also exceeds the baseline performance on cohesion scoring, but its performance on segmentation of argumentative elements drops slightly.

By comparing the confusion matrices in Table 6, we can observe that the multi-task learning model can yield a better performance on the essays with lower cohesion scores (<2.5). However, it also produce more incorrect score predictions in the majority class.

Setting	Span F1	Cohesion F1	Cohesion QWK
Arguments Span Only (with goldstandard)	.66	-	-
Arguments Span Only (with predicted span)	.63	-	-
Cohesion Only	-	.21	.54
Longformer+ARG	-	.18	.57
MTL_{linear} (with predicted span)	.63	**.30**	**.60**
$MTL_{dynamic}$ (with predicted span)	.62	.23	.59

Table 5: Automatic cohesion scoring in multi-task setting using segmentation of argumentative elements as an auxiliary task.

6. Conclusion

This paper explores argumentative element spans as a component in the automatic scoring of cohesion. In order to answer **RQ. 1** and **RQ. 2**, we explored the correlation between different feature sets and the cohesion score. We found that the lexical cohesion between claims plays a more significant role in overall cohesion, while diversity and divergence are expected in the evidence. Moreover, the experimenters with argumentative features demonstrated that augmenting the feature set with argumentative elements and lexical chains can improve the performance of transformer-based automatic cohesion scoring. Lastly, to answer the **RQ 3**, our experiments showed that jointly learning the segmentation of argumentative elements as an auxiliary task can improve the performance of cohesion score prediction. These studies demonstrate that integrating argumentative and cohesion features can enhance the performance of automatic essay

	1.5	2	2.5	3	3.5	4	4.5	5
1.5	0	2	2	0	0	0	0	0
2	0	0	17	9	2	0	0	0
2.5	0	1	30	32	4	0	0	0
3	0	0	29	90	36	5	0	0
3.5	0	0	6	62	41	10	0	0
4	0	0	0	18	33	8	1	0
4.5	0	0	0	0	7	2	2	0
5	0	0	0	0	1	2	0	0

	1.5	2	2.5	3	3.5	4	4.5	5
1.5	2	2	0	0	0	0	0	0
2	0	6	12	8	2	0	0	0
2.5	0	7	27	23	9	1	0	0
3	0	2	45	67	38	8	0	0
3.5	0	1	8	43	54	13	0	0
4	0	0	0	8	37	15	0	0
4.5	0	0	0	0	3	8	0	0
5	0	0	0	0	1	2	0	0

Table 6: Confusion matrix between gold standard (rows) and predictions (columns) of Longformer baseline model (upper) and MTL_{linear} model (down).

scoring and provides educators and researchers with a practical solution to improve cohesion scoring accuracy.

7. Discussion and Future Work

Recently, a larger version of the PERSUADE and ELLIPSE corpus[6] has been published, which has almost doubled the amount of available data in FB-Score with 26K essays, FB-Argumnent with 6.5K essays, and their overlap with 762 essays, as shown in Table 7.

	Cohesion Score									Only
	1.0	1.5	2.0	2.5	3	3.5	4.0	4.5	5.0	**in FB-Arg.**
#Essays	**1**	**4**	**43**	**121**	**268**	**199**	**101**	**22**	**3**	**25,234**
Lead	0	0	16	46	125	110	68	18	3	14,712
Position	1	4	42	121	268	200	101	22	3	24,907
Claim	6	11	169	460	1,107	940	499	122	22	95,999
Evidence	5	10	124	334	804	648	319	70	12	73,262
Conclusion	0	1	25	98	215	181	95	22	3	21,643
#Essays only in FB-Score	**12**	**40**	**486**	**1,154**	**1,612**	**1,414**	**787**	**180**	**35**	-

Table 7: Data distribution in the newly released larger version of FB-Arguments and FB-Score.

We reran the studies in this paper on the larger dataset and found that the improvement brought by feature augmentation and multi-task learning on Longformer does not persist (Table 8). This observation, however, is not coming as a surprise. Prior research has shown the critical role of the training

[6] https://github.com/scrosseye/persuade_corpus_2.0

Setting	MSE	F1	QWK	Setting	MSE	F1	QWK
Baseline	.20	.29	.62				
$MTLlinear$	+0	.02	.04	**+ Transitional Words**	.01	.07	.04
$MTLdynamic$	+0	.02	.05	+TW +ARG	.01	.01	.02
				+TW in Claims	.01	.08	.04
+ARG	.01	.02	.04	+TW in Evidence	.01	.08	.04
+LEX-1	+0	+0	01	**+TW +LEX-1**	.01	.06	.04
+LEX-1 +ARG	+0	.03	.02	+TW +LEX-1 +ARG	.01	.03	.02
+LEX-1 in Claims	.01	.02	.03	(+TW +LEX-1) in Claims	.01	.02	.06
+LEX-1 in Evidence	+0	.06	+0	(+TW +LEX-1) in Evidence	.01	.01	.04
+LEX-2	+0	.02	.03	**+TW +LEX-2**	+0	.06	.02
+LEX-2 +ARG	+0	.03	.03	+TW +LEX-2 +ARG	.02	.01	+0
+LEX-2 in Claims	+0	.03	.03	(+TW +LEX-2) in Claims	.01	.07	.01
+LEX-2 in Evidence	.01	.05	.03	(+TW +LEX-2) in Evidence	.01	.01	.01
+LEX	+0	.02	.02	**+TW +LEX**	+0	.04	.02
+LEX +ARG	+0	.07	.02	+TW +LEX +ARG	.02	.07	.02
+LEX in Claims	+0	.05	.04	(+TW +LEX) in Claims	.01	+0	.03
+LEX in Evidence	.01	.03	.03	(+TW +LEX) in Evidence	.01	.03	.01

Table 8: Automatic cohesion scoring with MTL and additional features using Longformer on the larger version of datasets.

corpora size in the discriminative power of the supervised learners (Banko and Brill, 2001; Crossley et al., 2017). Therefore, training with the larger version of the corpus results in a more robust and accurate baseline model, which may encounter a ceiling effect that makes further performance improvement challenging, even through feature augmentation.

Furthermore, it's important to note that larger datasets may not always be readily available in many problem domains. As shown in Table 1, the majority of available datasets for cohesion scoring have comparable sizes to our original setting. This suggests that the current study offers a practical solution for enhancing cohesion scoring accuracy by compensating for the lack of large training examples through exploiting external sources of information, i.e., argumentative and discourse feature augmentation.

8. Limitations

Large Language Models (LLMs) have recently grown in importance and prominence in the field. According to our preliminary study with LLMs (GPT 3.5 Turbo and Bard) for argumentative scoring using rubric-based prompting, the results were not in a ballpark range. Therefore, we did not include LLM experiments in this paper. However, it provides potential future research in this direction, i.e., how to use LLM to generate better feedback on students' argumentative essays.

9. Acknowledgements

This work was partially conducted at "CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics" of the FernUniversität in Hagen, Germany, and supported by a fellowship of the German Academic Exchange Service (DAAD) for a research visit at Educational Testing Service (ETS). We thank James Bruno, Matthew Mulholland and Nitin Madnani for their valuable feedback and discussions.

10. Bibliographical References

Richard Andrews. 2009. *Argumentation in higher education: Improving practice through theory and research*. Routledge.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. 2004. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 26–33.

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. 2001. Enriching automated essay scoring using discourse marking.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, 18(1):32–39.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, pages 681–684.

Steve Chiang. 2003. The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System*, 31(4):471–484.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Scott Crossley, Mihai Dascalu, and Danielle McNamara. 2017. How important is size? an investigation of corpus size and meaning in both latent semantic analysis and latent dirichlet allocation. In *The thirtieth international flairs conference*.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016a. The development and use of cohesive devices in l2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32:1–16.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016b. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48:1227–1237.

Scott A Crossley, Laura K Varner, Rod D Roscoe, and Danielle S McNamara. 2013. Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16*, pages 269–278. Springer.

Ronan Cummins and Marek Rei. 2018. Neural multi-task learning in automated assessment. *arXiv preprint arXiv:1801.06830*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yuning Ding, Marie Bexte, and Andrea Horbach. 2022. Don't drop the topic-the role of the prompt in argument identification in student writing. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 124–133.

Yuning Ding, Marie Bexte, and Andrea Horbach. 2023. Score it all together: A multi-task learning study on automatic scoring of argumentative essays. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13052–13063.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse.

Ruqaiya Hasan. 2014. *Cohesion in english*. 9. Routledge.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Filipe Sateles Lima, Aluizio Haendchen Filho, Hércules Prado, and Edilson Ferneda. 2018. Automatic evaluation of textual cohesion in essays. In *19th International Conference on Computational Linguistics and Intelligent Text Processing*.

Nitin Madnani, Aoife Cahill, Daniel Blanchard, Slava Andreyev, Diane Napolitano, Binod Gyawali, Michael Heilman, Chong Min Lee, Chee Wee Leong, Matthew Mulholland, et al. 2018. A robust microservice architecture for scaling automated scoring applications. *ETS Research Report Series*, 2018(1):1–8.

Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48.

Panitan Muangkammuen and Fumiyo Fukumoto. 2020. Multi-task learning for automated essay scoring with sentiment analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 116–123, Suzhou, China. Association for Computational Linguistics.

Huy Nguyen and Diane Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

E Michael Nussbaum, CarolAnne M Kardash, and Steve Ed Graham. 2005. The effects of goal instructions and text on the generation of counterarguments during writing. *Journal of Educational Psychology*, 97(2):157.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based Argument Mining and Automatic Essay Scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Isaac Persing and Vincent Ng. 2015. Modeling Argument Strength in Student Essays. 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 543–552.

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.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. Coherence based automatic essay scoring using sentence embedding and recurrent neural networks. In *International Conference on Speech and Computer*, pages 139–154. Springer.

Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. *arXiv preprint arXiv:1704.07156*.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical papers*, pages 950–961.

Christian Stab and Iryna Gurevych. 2014a. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2014b. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-Topic Argument Mining from Heterogeneous Sources using Attention-based Neural Networks. *arXiv preprint arXiv:1802.05758*.

Paul Stapleton and Yanming Amy Wu. 2015. Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance. *Journal of English for Academic Purposes*, 17:12–23.

Yi Tay, Minh Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Stephen Edelston Toulmin. 1958. The uses of argument.

Raphael Vallat. 2018. Pingouin: statistics in python. *The Journal of Open Source Software*, 3(31):1026.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers*, pages 1680–1691.

Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in esol learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 33–43.

11. Language Resource References

Scott A Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, and Ulrich Boser. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assessing Writing*, 54:100667.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, Magali Paquot, et al. 2009. *International corpus of learner English*, volume 2. Presses universitaires de Louvain Louvain-la-Neuve.

Andrea Horbach, Dirk Scholten-Akoun, Yuning Ding, and Torsten Zesch. 2017. Fine-grained essay scoring of a complex writing task for native speakers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 357–366, Copenhagen, Denmark. Association for Computational Linguistics.

Jeziel C Marinho, Rafael T Anchiêta, and Raimundo S Moura. 2021. Essay-br: a brazilian corpus of essays. *arXiv preprint arXiv:2105.09081*.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 229–239.

Wei Song, Ziyao Song, Lizhen Liu, and Ruiji Fu. 2020. Hierarchical multi-task learning for organization evaluation of argumentative student essays. In *IJCAI*, pages 3875–3881.

The Learning Agency Lab. 2022. *The ELLIPSE Corpus*. released from Kaggle competitions Feedback Prize: English Language Learning.