Transfer Learning of Argument Mining in Student Essays

Yuning Ding[1], Julian Lohmann[2], Nils-Jonathan Schaller[3], Thorben Jansen[3], Andrea Horbach[1,4]
[1]CATALPA, FernUniversität in Hagen, Germany
[2]Institute for Psychology of Learning and Instruction, Kiel University, Germany
[3]Leibniz Institute for Science and Mathematics Education at the University of Kiel, Germany
[4]Hildesheim University, Germany

Abstract

This paper explores the transferability of a cross-prompt argument mining model trained on argumentative essays authored by native English speakers (EN-L1) across educational contexts and languages. Specifically, the adaptability of a multilingual transformer model is assessed through its application to comparable argumentative essays authored by English-as-a-foreign-language learners (EN-L2) for context transfer, and a dataset composed of essays written by native German learners (DE) for both language and task transfer. To separate language effects from educational context effects, we also perform experiments on a machine-translated version of the German dataset (DE-MT). Our findings demonstrate that, even under zero-shot conditions, a model trained on native English speakers exhibits satisfactory performance on the EN-L2/DE datasets. Machine translation does not substantially enhance this performance, suggesting that distinct writing styles across educational contexts impact performance more than language differences.

1 Introduction

Argumentative writing is a central skill to succeed across school subjects (Graham et al., 2020) and automated feedback is an effective way to foster writing skills (Fleckenstein et al., 2023). Figure 1 shows an example of providing students with feedback by highlighting different argumentative elements, such as *lead, position, claim* and *conclusion* in their writing. Such feedback offers guidance to students for enhancing the structure of their essays.

However, training a dedicated feedback model for each new task could incur substantial costs. One approach to mitigate this expense is to transfer a pre-trained model to new datasets. While existing research highlights model transferability across different writing prompts (Ding et al., 2022), no research demonstrates whether employing English argument mining models across languages and different educational contexts yields consistent performance.

Such educational contexts for argumentative writing can be specified according to two dimensions: native vs. foreign language instruction on the one hand and independent vs. integrated writing tasks on the other hand.

We first have a closer look at the differences between L1 and L2 writing. In L1 teaching contexts, the emphasis is primarily on content, whereas in L2, the focus is on language acquisition and structure. These differences are also reflected in distinct cognitive models and therefore writing styles (Devine et al., 1993). Beyond obvious characteristics such as spelling and grammar errors in L2 writing, like the misspelled *advetisments* and the subject-verb disagreement exemplified in Figure 1, prior studies also unveiled that non-native English writers tend to craft shorter sentences and employ fewer hedges (e.g., *probably*, *may*) to moderate the strength of their claims, in contrast to native English speakers (Burrough-Boenisch, 2002). Moreover, L2 writers prefer a more straightforward argumentation structure and often avoid counter-arguments (Sanders and Schilperoord, 2006).

As for the second dimension, in independent tasks, individuals are typically provided with a specific writing prompt and are required to formulate their essays based solely on their thoughts, experiences, and knowledge. For example, the independent writing prompt of EN-L2 in Figure 1 is:

> *Do you agree or disagree with the following statement? Television advertising directed toward young children (aged two to five) should not be allowed. Use specific reasons and examples to support your answer.*

In integrated tasks, writers are presented with one or more texts related to a particular topic and then asked to synthesize information from the provided texts and incorporate it into their writing. For in-

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 439–449

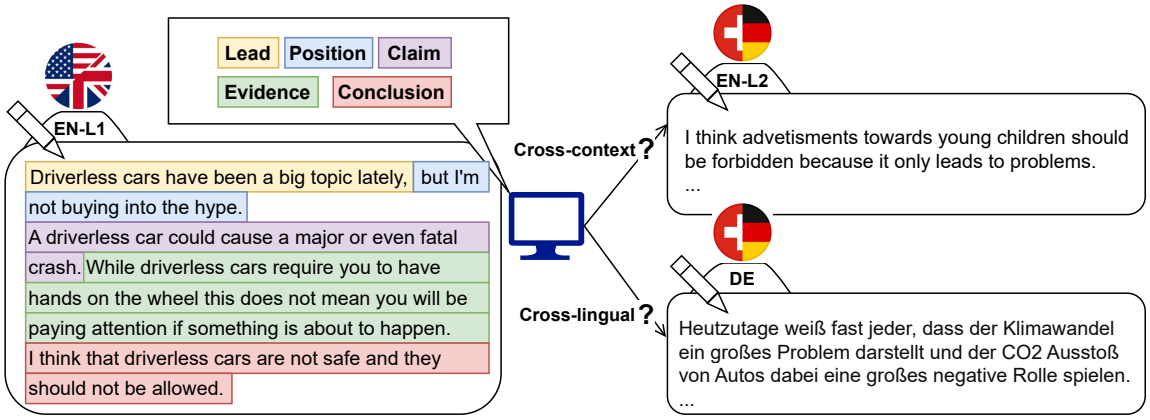

Figure 1: Example of automatic feedback provided by an argument mining model trained on EN-L1 essays and its uncertain transferability to EN-L2 (upper right) and DE (lower right) data.

stance, the DE dataset has an integrated writing prompt, which discusses using renewable energy sources to combat climate change. It presents three options: a wind farm, a solar park, and a hydropower plant. Students are asked to evaluate these options based on specific criteria, taking a stand in favor of one source and providing supporting arguments. Earlier studies have shown that task type can influence lexical complexity and argument structure in essays (Cumming et al., 2005; Guo et al., 2013).

Targeting the challenge of transfer learning brought by the differences described above and the languages, the following research questions are investigated in this paper:

RQ. 1 *How do linguistic structures and argumentation styles differ among English L1, L2, and German datasets?*

RQ. 2 *How can the argument mining model, initially trained on the English L1 dataset, be effectively transferred to L2 and German datasets? How much data is needed to achieve the best transfer performance?*

RQ. 3 *In the context of cross-lingual transfer, what roles are played by language differences and task disparities in influencing the model's performance?*

Through a comparative analysis of English L1, L2, and German datasets, we answer **RQ 1** by showing the statistical and structural distinctions inherent in argumentative essays across different educational contexts and languages. While the English L1 and L2 data are written for independent tasks, the German dataset is collected from integrated writing tasks, this completes our study with a focus not only on cross-lingual transfer learning but also on cross-educational contexts. We then conduct two experimental studies to transfer argument mining models trained on a large English L1 dataset to the L2 and German datasets for **RQ 2**. In addressing the challenge of cross-language transfer in **RQ 3**, our research extends to experiments involving the machine-translated version of the German dataset. This expanded scope enables a more profound examination of the variances in model performance arising from linguistic disparities and diverse writing tasks.

The answer to these questions could be invaluable in developing educational applications: with the appropriate adjustments, models trained on English L1 data can effectively be transferred to an L2 dataset. This would greatly benefit the development of educational applications, particularly in contexts where resources are limited, by providing students with access to high-quality learning tools and feedback systems. Additionally, the impact of linguistic differences on the model's effectiveness is essential for the development of educational applications aimed at student populations from different linguistic backgrounds, ensuring they receive the support they need to improve their argumentative writing skills.

2 Related Work

Transfer learning has been extensively studied for many years. Surveys such as Pan and Yang (2009), Weiss et al. (2016), and Zhuang et al. (2020) provide a comprehensive overview of the developments in this area over the years. Similarly, numerous studies have explored the topic of argument mining through literature reviews, evidenced by works like Peldszus and Stede (2013)

and Lawrence and Reed (2020). In this paper, we focus our review of related work specifically on transfer learning within the educational domain and argument mining in student essays.

2.1 Transfer Learning in Education

In many educational scoring tasks, transfer learning is important in avoiding retraining a model for every new task. Especially in the area of automated essay scoring, cross-prompt and prompt-independent models are widely researched (e.g. Jin et al. (2018), Ridley et al. (2021), Xue et al. (2021))

Fewer approaches have focused on a transfer between languages in educational scoring, for example for content scoring (Horbach et al., 2018, 2023) or language proficiency classification (Vajjala and Rama, 2018).

Approaches for cross-lingual argument mining in the educational domain are even scarcer. Eger et al. (2018) automatically translated an educational argument mining dataset into various languages showing the feasibility of a cross-lingual transfer. To the best of our knowledge, we are the first to attempt such a transfer on authentic ecologically valid cross-lingual data, extending the research body on cross-lingual argument mining approaches in other domains such as medicine (Yeginbergenova and Agerri, 2023) or general controversial topics (Toledo-Ronen et al., 2020).

Differences in the educational and cultural context of argumentative essay scoring have been studied by Chen et al. (2022) finding that, for the ICLE corpus containing essays by English learners with 16 native languages, culture influenced learners' argumentation patterns substantially.

2.2 Argument Mining in Student Essays

Various approaches for argument mining in student essays exist with many of them adopting the persuasive essay scheme introduced by Stab and Gurevych (2014), such as Wambsganss et al. (2020); Putra et al. (2021) and Alhindi and Ghosh (2021). This model comprises four key categories: *major claim, claim, premise,* and *non-argumentative elements*.

In this study, we have five different argumentative elements, namely *lead, position, claim, evidence,* and *conclusion*, which is a simplified version of the task definition set by the Kaggle Feedback Prize competition [1] on the PERSUADE dataset (Crossley et al., 2022). This dataset adopts a variant of the Toulmin argument mining model (Toulmin, 1958), the same as the German dataset we used for the transfer learning task (Schaller et al., 2024). Ding et al. (2022) trained a sequence tagging model using the pre-trained Longformer (Beltagy et al., 2020) on PERSUADE, achieving an F1 score of .55. We leverage their framework in our experiments.

3 Data

In our experiments, we work with three different datasets: PERSUADE, MEWS, and DARIUS. In the following, we go into details for each dataset, describe our label mapping as the basis for the transfer learning, and compare the sequencing of argumentative elements in each dataset.

EN-L1 The PERSUADE corpus (Crossley et al., 2022) encompasses a collection of 26,000 argumentative essays authored by students in grades 6-12 within the United States, mostly English native speakers. Expertly annotated, these essays feature seven categories of argumentative elements: *lead, position, claim, counterclaim, rebuttal, evidence,* and *concluding statement*. The quality of annotations is evaluated using F1 score reaching an inter-rater agreement (IAA)[2] of 0.73.

EN-L2 The MEWS corpus (Rupp et al., 2019) comprises 9,628 essays written by English-as-a-foreign-language learners in Switzerland and Germany. For this study on transfer learning across L1 and L2 context, we randomly drew and annotated a subset of 110 essays responding to the *Television Advertising* (AD) prompt and 100 essays addressing the *Teachers Ability* (TE) prompt [3]. In terms of writing tasks, these two prompts are close to those in EN-L1 because they are independent writing tasks. These essays were annotated following the same schema as EN-L1, achieving an IAA of F1 = 0.52.

DE DARIUS (Schaller et al., 2024) is a corpus comprising 2,521 texts from the "Energy" prompt

[1] https://www.kaggle.com/c/feedback-prize-2021

[2] The calculation of IAA takes an annotation as a true positive when it is identified by two annotators with over 50% overlap in both directions. Elements identified exclusively by the first annotator are considered false negatives, whereas those only recognized by the second annotator are deemed false positives.

[3] Detailed writing instructions are available on Page 13 and Page 34 at https://www.ets.org/pdfs/toefl/toefl-ibt-writing-practice-sets-large-print.pdf

and 2,517 from the “Automotive" prompt, which are written by German high school students. Similar to the datasets above, this dataset also has an annotation of argumentative elements (with different names, see details in Section 3.2). The IAA among different layers of annotation ranges between 0.57 and 0.98.

The performances of transfer learning on **DE** dataset can be influenced by both language and educational contexts. To keep them apart, we translate it into English as the dataset **DE-MT**, using DeepL Pro[4]. Applying experiments on this dataset would help us distinguish between the impact of the writing task migration and the language transition during transfer learning.

3.1 Dataset Comparison

Table 1 shows the descriptive statistics of the three datasets. We see that EN-L1 and EN-L2 have a comparable length in terms of the average number of sentences (21.25 and 20.56 respectively), whereas the German texts are significantly shorter (9.53). However, this difference does not originate from language, since it is almost the same as the translated data DE-MT (9.81). Instead, this large difference may be attributed to the nature of the writing tasks. As emphasized above, the writing prompts of EN-L1 and L2 are similar, requiring students to produce independent argumentative essays. In contrast, the DE dataset employs integrated writing prompts, potentially leading to shorter, more concise responses.

This point can be also observed in the average number of words per essay, where the DE dataset has the smallest amount (149.89). The EN-L1 dataset leads with 402.31, followed by EN-L2 (349.68). The EN-L2 dataset, despite having a larger average number of sentences, exhibits fewer average words per essay. This observation suggests that L2 writers tend to compose shorter sentences, aligning with findings from the prior study (Burrough-Boenisch, 2002).

Dataset	#Essays	ϕ#Sentences	ϕ#Words
EN-L1	26,000	20.56	402.31
EN-L2	210	21.25	349.68
DE	5038	9.53	149.89
DE-MT	5038	9.81	163.13

Table 1: Descriptive statistics of datasets.

[4]https://www.deepl.com/pro?cta=header-pro

3.2 Label Mapping

For a consistent annotation mapping across diverse datasets, we adopt a streamlined label-set inspired by Ding et al. (2024) for the EN-L1 and EN-L2 datasets. Specifically, we employ the labels *lead, position, claim, evidence*, and *conclusion*, by merging the labels *counterclaim* and *rebuttal* into a single label *claim*. The labels are defined as follows.

- *Lead*: an introduction to grab the reader's attention and point toward the position.
- *Position*: an opinion on the main question.
- *Claim*: a claim that supports the position, refutes another claim or gives an opposing reason to the position.
- *Evidence*: ideas or examples that support claims.
- *Conclusion*: a concluding statement that restates the claims

The DE dataset has a four-layer annotation schema. On the *Content Zone* layer, *introduction, main part* and *conclusion* are labeled to delineate the text's framing and structure. On the *Major Claim* layer, sentences referring to the author's final position on the given topic are labeled as *major claims*. While the *Argument* layer, focused on argument quality, is less relevant to our argument mining study, the layer of *Toulmin's Argumentation Pattern (TAP)* is directly pertinent. This layer aligns with the argument schema in EN-L1 and EN-L2, encompassing the annotated elements:

- *Claim*: an assertion that characterizes the position taken.
- *Data*: fact that provides the basis for a claim.
- *Warrant*: an aspect that explains to what extent data supports a claim.
- *Rebuttal*: an objection to a presented data and/or warrant.

Based on the above definitions, the mapping detailed in Table 2 is established to facilitate our transfer learning approach. Firstly, these five types of argumentative elements in three datasets can be compared in the following analysis. Secondly, we can train an argument mining model detecting these elements on the EN-L1 dataset and test its transferability on the other datasets (zero-shot transfer in Section 4). With the label mapping, the essays in EN-L2 and DE can also be added gradually to fine-tune this model for potentially better performance (learning curve study in Section 5). In the

following, we refer to the mapped labels by their names in the English datasets, i.e. *lead, position, claim, evidence* and *conclusion*.

EN-L1 and EN-L2	Annotation Layer in DE	Label in DE
Lead	Content Zone	introduction
Position	Major Claim	major claim
Claim	TAP	claim or rebuttal
Evidence	TAP	data or warrant
Conclusion	Content Zone	conclusion

Table 2: Label mapping of three datasets.

3.3 Analysis - Label Distribution and Length

Figure 2 visualizes the distribution and the average length (in the number of words) of five types of argumentative elements in the respective dataset.

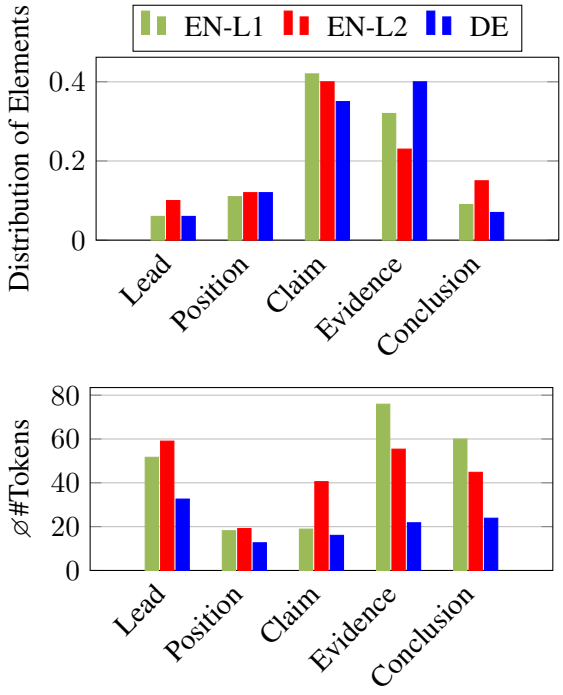

Figure 2: Distribution (upper) and average number of tokens (down) of argumentative elements in three datasets.

The distribution of various argumentative elements is generally comparable among three datasets, with both *claim* and *evidence* emerging as the dominant major classes across all datasets. For the average length in general, DE has the least number of words in all the elements. This again corresponds to our previous analysis of text length, that German argumentative essays tend to be briefer and might have originated from its integrated writing prompts.

The *claim* is most frequent in the EN-L1 dataset, followed closely by EN-L2, while DE exhibits a slightly lower frequency. This suggests a consistent emphasis on presenting central arguments across both English datasets. However, EN-2 stands out with the longest average length for *claim*, suggesting that argumentative essays written by second-language learners may provide more detailed or elaborate claims for central positions compared to native speakers and German writers.

DE exhibits the highest frequency of *evidence* labels among the three datasets, indicating a relatively higher occurrence of supporting details in argumentative essays compared to EN-L1 and EN-L2. However, DE also has the shortest average length for this label. It indicates that although EN-L2 has a higher frequency of evidence, the individual instances are shorter. It could also suggest the possibility of multiple spans or fragmented annotations for longer evidence segments.

EN-L2 stands out with the highest frequency of *lead* and *conclusion* labels, implying emphasis at the beginning and end of essays by non-native English writers. In contrast, native writers (EN-L1 and DE) display lower percentages for these labels. Especially for DE, it exhibits both the lowest frequency and the shortest average length of *conclusion*, suggesting a brief concluding style in German essays.

3.4 Analysis - Label Transition

To examine the structure of essays in datasets, we visualize the argumentation flow as transition graphs where argumentative elements correspond to states and arrows mark the transitions from one element to another annotated with the transition probability (Figure 3). We add two states 'START' and 'END', indicating the beginning and end of an essay. For a clearer illustration, all transition arrows with probabilities below 0.2 are omitted.

EN-L1 essays (left subfigure) predominantly start with a *lead* and follow with a *position*. Subsequently, the transition to *claim* is most likely, and from there, essays often transition to another *claim* or an *evidence*. Finally, almost all the essays end with *conclusion*. This style is influenced by the five-paragraph essay model, which is the most frequently taught form of writing in classrooms in the US (Campbell, 2014). It usually consists of one introductory paragraph, three body paragraphs for support, and one concluding paragraph.

Similar to EN-L1, EN-L2 essays (sub-figure in the middle) also start with a *lead* predominantly. However, the *lead* is no longer followed directly by the *position*, but by *claims*. Instead, the *position*

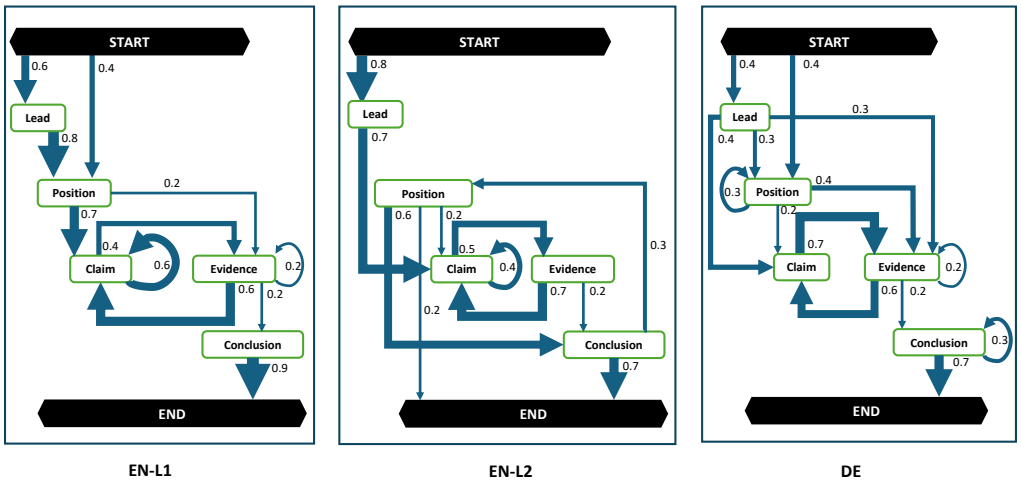

Figure 3: Transitions of different elements in the essays from three datasets.

can be mostly found at the end of the essays, which is illustrated by the 0.6 probability of transitioning from *position* to *conclusion*, as well as the 0.3 probability of transitioning backward from *conclusion* to *position*. By delving deeply into the teaching guidance of argumentative writing in Germany, we found a possible reason for this phenomenon: when it comes to stating a position in argumentative writing, German students are encouraged to state their own opinion at the end for a balanced discussion (Becker-Mrotzek et al., 2010).

The structuring style in German essays (right sub-figure) is more diverse. Firstly, almost the same amount of essays start with a *lead* or a *position*, which aligns with the suggestion in the earlier study that arguments in German have a higher level of directness (Tannen, 1998). In other words, German writers tend to jump straight into the position instead of introducing the topic first with a lead. Besides *claim* and *position*, 40% *lead* was directly followed by the *evidence*. Unlike the English datasets, the *claims* in DE are rarely followed by another claim but dominantly followed by an *evidence*. This discrepancy can be attributed to the integrated writing task in DE, which imposes a greater demand on students to integrate evidence from sources into their writing (Cumming et al., 2005). At last, we notice that more self-transitions in DE (30% *conclusions*, 20% *evidence* and 30% *positions*), which may not be an inherent property of the essays but rather an annotation artifact based on a high granularity.

4 Study 1: Zero Shot Transfer

For our first study, we adopt the sequence tagging architecture developed by Ding et al. (2022), which pre-processes the annotated training data into tokens with Inside-Outside-Beginning (IOB) tags and uses them as the input to the pretrained Longformer model (Beltagy et al., 2020) for token classification. We trained two models on 90% of EN-L1 data with the *Longformer*[5] to transfer on EN-L2 data and its multi-lingual variation *XLM-R Longformer*[6] for the DE data. After 10 epochs of training with a maximal length of 1024 tokens, the IOB tags of tokens are post-processed into predictions for different argumentative elements.

Following the same schema as for the IAA evaluation, we evaluate our results also through the F1 score: all gold standards and predictions for a given argumentative element are compared. If the overlap between the gold standard and prediction in both directions is higher or equal to 0.5, the prediction is considered a true positive. If multiple matches exist, the match with the highest is taken. Any unmatched ground standards are false negatives and any unmatched predictions are false positives.

4.1 EN-L1 to EN-L2

Table 3 shows the performance of the argument mining model tested on EN-L1 and two different prompts on EN-L2. Overall, the transfer performance of the model achieves an F1 score of 0.56 and 0.42 on EN-L2 dataset, which is only a slight drop from the performance on EN-L1, indicating its effectiveness in extracting argumentative elements from essays written by both native and non-native English speakers. The model demonstrates the best

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

[6] https://huggingface.co/markussagen/xlm-roberta-longformer-base-4096

proficiency in detecting *lead* elements across all datasets. A possible explanation is that this element is often found at the beginning of essays and therefore easy to find. The transferability of the model tested on the prompt AD is better than TE, implying the prompt similarity between AD and EN-L1 is higher than TE and EN-L1.

	Test Data		
	EN-L1	**EN-L2 AD**	**EN-L2 TE**
Lead	.76	.79	.63
Position	.61	.57	.28
Claim	.44	.41	.28
Evidence	.69	.49	.39
Conclusion	.78	.52	.50
Overall	.66	.56	.42

Table 3: Zero shot transfer from EN-L1 to EN-L2 with English pretrained Transformer

We examine the typical misclassification in the two prompts together. The confusion matrix in Table 4 illustrates that most of the confusion arises between a label and no assigned span, suggesting challenges in accurately delineating argumentation unit boundaries. More specifically, a gold argument is often divided into multiple predicted spans or vice versa. This issue results in numerous spans lacking a counterpart with significant overlap. Among the instances of actual confusion between two labels, we noted a common misclassification of *evidence* being incorrectly labeled as *claims*.

	Lead	**Position**	**Claim**	**Evidence**	**Conclusion**	**None**
Lead	100	4	6	14	0	86
Position	4	61	4	4	6	142
Claim	7	8	147	120	11	447
Evidence	5	2	14	157	14	205
Conclusion	0	11	6	15	64	158
None	72	79	277	284	89	N.A.

Table 4: Confusion matrix between gold standards (columns) and predictions (rows) of EN-L2.

4.2 EN-L1 to DE

Table 5 shows the result of zero-shot transfer learning from EN-L1 to DE and DE-MT. We first notice that on the same test dataset of EN-L1 the performance decreased by changing the pretrained Longformer into its multi-lingual version XLM-R Longformer. Especially for the label *position*, the F1 score dropped from .61 to .29. These results align with earlier studies, showing multilingual models have worse performance than their monolingual counterparts on certain downstream tasks (Conneau et al., 2020).

The transfer performance to DE is not as good as EN-L2, as evidenced by the lower F1 scores across all labels. However, the model's performance decline is not solely attributable to language differences between English and German, as even the machine-translated German dataset (DE-MT) exhibits similar performance. The F1 scores for *claim* and *evidence* are particularly low across both the DE and DE-MT datasets. This poor performance is likely influenced by the differences observed in the distribution and length of these elements in the integrated tasks, as discussed in Section 3.3.

	Test Data				
		Energy		Automotive	
	EN-L1	**DE**	**DE-MT**	**DE**	**DE-MT**
Lead	.73	.61	.63	.59	.62
Position	.29	.28	.35	.32	.32
Claim	.38	.15	.17	.14	.16
Evidence	.65	.28	.29	.27	.30
Conclusion	.74	.48	.50	.48	.48
Overall	.61	.36	.38	.36	.38

Table 5: Zero shot transfer from EN-L1 to DE and DE-MT with multi-lingual Transformer

The confusion matrix in Table 6 shows the same pattern as Table 4. Besides the majority of confusion occurring between a label and no assigned span, *claim* and *evidence* are often wrongly switched. When comparing the number of unmatched gold standard labels (7036) with that of unmatched predicted labels (25779), we see the model tends to assign a label rather than not assign anything.

	Lead	**Position**	**Claim**	**Evidence**	**Conclusion**	**None**
Lead	1102	357	76	58	0	630
Position	18	1147	285	133	41	3079
Claim	44	217	1319	1114	135	10276
Evidence	29	61	1028	3301	66	10702
Conclusion	0	445	74	128	1041	1092
None	261	614	2073	3813	275	N.A.

Table 6: Confusion matrix between gold standards (columns) and predictions (rows) of DE.

In summary, while the performance of our argument mining model does not match that achieved on the source dataset (EN-L1), considering it does not see any data from the target domain during training, it performs reasonably well in different educational contexts and languages (EN-L2 and DE). This highlights the potential of the generalization capability of this model.

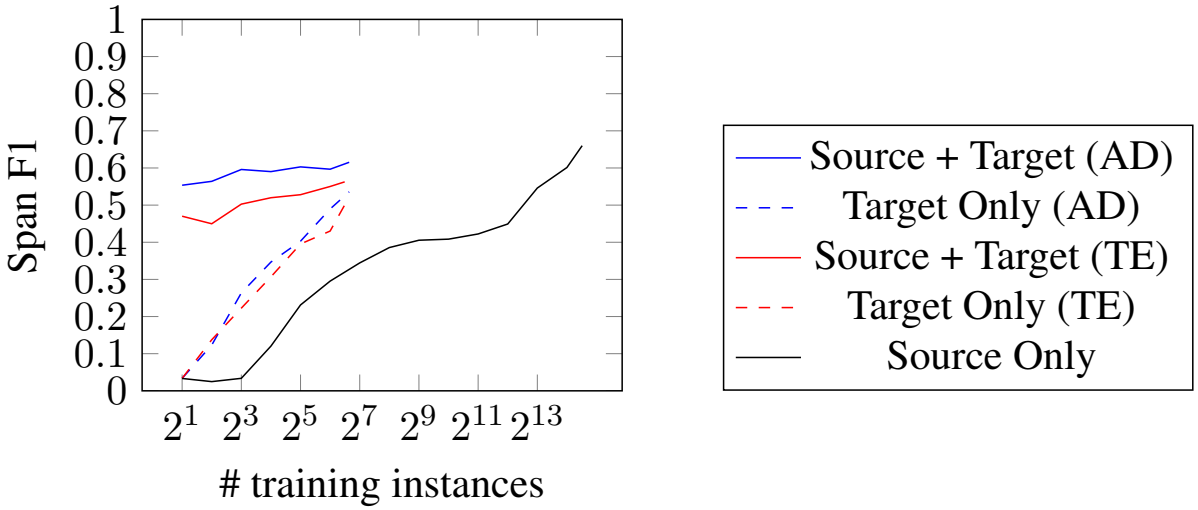

Figure 4: Learning curve EN-L2 and EN-L1

5 Study 2: Learning Curve

After having established zero-shot transfer performance, we investigate the potential of using a small amount of target domain training data to improve the performance of our argument mining model on the target test dataset. This process involves fine-tuning the model trained on EN-L1 data using a portion of data from the target domain (EN-L2 and DE), allowing it to adapt its representations of the features in L2 and German argumentative essays.

However, it is important to note that fine-tuning requires access to labeled data from the target domain. In a practical application scenario, when a teacher wants to fine-tune such a model for a new educational context or language, it is important to know how much data needs to be labeled, since human annotation effort is often a crucial factor.

Therefore, we perform a series of learning curve experiments, in which we systematically vary the amount of training data from target datasets.

5.1 EN-L1 to EN-L2

Since EN-L2 only has 210 labeled data, we use the ten-fold cross-validation data splitting and report the average performance. On each training data set, we fine-tune the model from zero-shot transfer for 10 epochs. In comparison to the fine-tuning (**Source+Target**), we also trained the Longformer from the beginning only using these training data from EN-L2 (**Target Only**).

Figure 4 plots the amount of training data from the target domain on the x-axis and the model performance (F1) on the y-axis. Both Source+Target curves start with relatively high F1 scores but exhibit slow growth as the number of training instances increases. In contrast, the "Target Only" curves demonstrate faster growth with increasing training instances. However, despite this rapid improvement, these lines do not achieve the same level of performance as the "Source + Target" scenarios. This indicates that the current amount of labeled data in EN-L2 is insufficient to match the performance achieved by incorporating knowledge from EN-L1. Therefore, the transfer learning strategy is necessary for the limited labeled data in the target domain.

To estimate the amount of data needed for labeling, the "Source Only" curve provides a reference. This curve represents the scenario where the model is trained solely on data from EN-L1. As the number of training instances from the target domain increases, the model performance on the target task is expected to approach the upper bound at F1=.66 with 23,400 labeled training data instances.

5.2 EN-L1 to DE

Figure 5 shows the learning curves of DE and DE-MT datasets. Same to Figure 4, all the Source + Target curves start at a relatively high-performance level but exhibit a slower rate of improvement. Unfortunately, the gap between them and Target Only lines can be quickly narrowed. This implies that in educational context transfer, such as transitioning from independent to integrated tasks, better performance can be attained by training the model from scratch using an adequate amount of labeled data from the target domain.

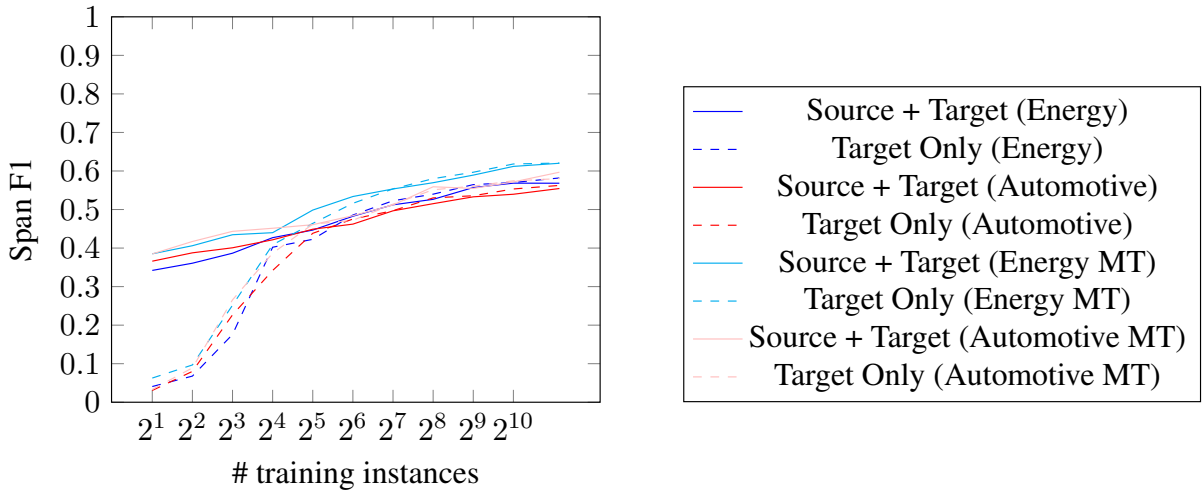

Figure 5: Learning curve DE and DE-MT

Regarding language transfer, the performance achieved through machine translation (MT) is found to be very close to that of the original source. As a result, there appears to be no significant benefit in using machine-translated data for training purposes.

6 Conclusion

This paper explores the transferability of argument-mining models across different educational contexts and languages. Through comprehensive analyses of various datasets, including those authored by native English speakers (EN-L1), English as a foreign language learners (EN-L2), and native German writers (DE), as well as machine-translated German essays (DE-MT), we answer RQ 1 and show their differences in linguistic structures and argumentation styles.

Our experimental studies designed for RQ 2 reveal that, under zero-shot conditions, models trained on EN-L1 demonstrate satisfactory performance when directly applied to EN-L2/DE datasets. However, fine-tuning the model on target domain data does not increase the performance significantly, highlighting the challenges of transfer learning across different educational contexts and languages. Notably, as the answer for RQ 3, machine translation does not significantly enhance performance, indicating that differences in dataset characteristics stem less from language disparities, but more from distinct educational contexts.

7 Limitations

This study showed the transferability of argument mining models for the English-German language pair on three specific corpora. Whether a transfer works equally well for languages phylogenetically further from the source language and potentially less well-covered in pretrained multilingual transformer models remains an open question.

Acknowledgements

This work was partially conducted at "CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics" of the FernUniversität in Hagen, Germany.

References

Tariq Alhindi and Debanjan Ghosh. 2021. "sharks are not the threat humans are": Argument component segmentation in school student essays. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 210–222.

Michael Becker-Mrotzek, Frank Schneider, and Klaus Tetling. 2010. Argumentierendes schreiben–lehren und lernen. 17:2012.

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

J Burrough-Boenisch. 2002. *Culture and conventions: writing and reading Dutch scientific English*. Ph.D. thesis, Utrecht: LOT.

Kimberly Hill Campbell. 2014. Beyond the five-paragraph essay. *Educational Leadership*, 71(7):60–65.

Wei-Fan Chen, Mei-Hua Chen, Garima Mudgal, and Henning Wachsmuth. 2022. Analyzing culture-specific argument structures in learner essays. In *Proceedings of the 9th Workshop on Argument Mining*, pages 51–61.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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.

Alister Cumming, Robert Kantor, Kyoko Baba, Usman Erdosy, Keanre Eouanzoui, and Mark James. 2005. Differences in written discourse in independent and integrated prototype tasks for next generation toefl. *Assessing Writing*, 10(1):5–43.

Joanne Devine, Kevin Railey, and Philip Boshoff. 1993. The implications of cognitive models in l1 and l2 writing. *Journal of second language writing*, 2(3):203–225.

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, Omid Kashefi, Swapna Somasundaran, and Andrea Horbach. 2024. When Argumentation Meets Cohesion: Enhancing Automatic Essay Scoring. Accepted for LREC-COLING 2024.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844.

Johanna Fleckenstein, Lucas W Liebenow, and Jennifer Meyer. 2023. Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, 6.

Steve Graham, Sharlene A Kiuhara, and Meade MacKay. 2020. The effects of writing on learning in science, social studies, and mathematics: A meta-analysis. *Review of Educational Research*, 90(2):179–226.

Liang Guo, Scott A Crossley, and Danielle S McNamara. 2013. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, 18(3):218–238.

Andrea Horbach, Joey Pehlke, Ronja Laarmann-Quante, and Yuning Ding. 2023. Crosslingual content scoring in five languages using machine-translation and multilingual transformer models. *International Journal of Artificial Intelligence in Education*, pages 1–27.

Andrea Horbach, Sebastian Stennmanns, and Torsten Zesch. 2018. Cross-lingual content scoring. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 410–419.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. Tdnn: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga. 2021. Parsing argumentative structure in English-as-foreign-language essays. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 97–109, Online. Association for Computational Linguistics.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13745–13753.

André A Rupp, Jodi M Casabianca, Maleika Krüger, Stefan Keller, and Olaf Köller. 2019. Automated essay scoring at scale: a case study in switzerland and germany. *ETS Research Report Series*, 2019(1):1–23.

T Sanders and Joost Schilperoord. 2006. Text structure as a window on the cognition of writing. *Handbook of writing research*, pages 386–402.

Nils-Jonathan Schaller, Andrea Horbach, Lars Höft, Yuning Ding, Jan L Bahr, Jennifer Meyer, and Thorben Jansen. 2024. Darius: A comprehensive learner corpus for argument mining in german-language essays. Accepted for LREC-COLING 2024.

Christian Stab and Iryna Gurevych. 2014. 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, Dublin,

Ireland. Dublin City University and Association for Computational Linguistics.

Deborah Tannen. 1998. The argument culture: Moving from debate to dialogue. *New York: Random House Trade.*

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of the association for computational linguistics: Emnlp 2020*, pages 303–317.

Stephen E Toulmin. 1958. *The uses of argument*. Cambridge university press.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. A corpus for argumentative writing support in German. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 856–869, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3:1–40.

Jin Xue, Xiaoyi Tang, and Liyan Zheng. 2021. A hierarchical bert-based transfer learning approach for multidimensional essay scoring. *Ieee Access*, 9:125403–125415.

Anar Yeginbergenova and Rodrigo Agerri. 2023. Cross-lingual argument mining in the medical domain.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.