UROP Project Report 2021: Providing Automatic Feedback on Argumentation Quality to Learners of English

Suchir Salhan
Computer Laboratory &
Gonville & Caius College
University of Cambridge, U.K.
sas245@cam.ac.uk

Abstract

We are developing a learning tool that aims to help students learn the skill of argumentation by providing high-level, automatic adaptive feedback on the quality of their argumentation in essays. The tool consists of a pretrained model that automatically analyses argumentation structure in written English and a front-end interface as a Google Doc add-on. We will conduct an online study that evaluates how users respond to a software tool that provides automatic feedback on the argumentation structure of an essay they have written.

1 Introduction

Argumentation Mining is the task of automatically identifying and extracting arguments from natural language text. Argumentation analysis aims to turn unstructured text into structured argument data that provides information about (1) individual arguments that are made, and (2) the relationships between arguments, for instance whether an argument support or undermine the overall message.

Educational Motivation: Argumentation is a high-level academic skill that forms a crucial part of essay writing. While the landscape of automated writing assistance has traditionally focused on lexical and syntactic checking and feedback, recent advances in deep learning, and argumentation mining in particular, allow the scope of automated writing assistance applications to broaden to provide high-quality automated feedback at the discourse level (Dale and Viethen, 2021).

Argumentation is often an implicitly learnt task: learners have to deduce the components of a good argument. A teacher will often say to a student that they have spelt a word incorrectly or that they have written an ungrammatical sentence and provide a correction. However, if a student has written an essay that contains an illogical or even unpersuasive argument, a teacher may not necessarily say to the student how they should go about writing a persuasive paragraph, that consists of a main point backed up by evidence. Even if they do, this advice is far less formulaic and harder to consistently apply to different essays.

Automated argumentation writing assistance may have more pedagogical utility in implicitly learnt tasks, like argumentation, compared to traditional syntactic and lexical feedback. Such tools allow students to independently learn the components of effective argumentation, and potentially allow students to improve the clarity and persuasiveness of their essay writing. In the Learning to Argue project, we first develop a learning tool that provides automatic provides automatic adaptive feedback to learners of English. We evaluate whether students do, in fact, improve their argumentation skills through automated adaptive feedback through a pilot study. We will release the learning tool as a Google-Doc add-on so users of Google Docs can receive argumentation feedback on any piece of writing.

2 Background

Essay-writing typically follows a claim-oriented procedure: the main argument consists of a set of **claims**, a controversial statement that an author tries to persuade a reader which the reader will typically not accept without additional support from simple **premises**. *Figure 1* shows the claims (shown in bold), the premises (shown in red), and the conclusion (shown in blue) in an argumentative essay.

2.1 Neural Network Techniques in Argumentation Mining

Neural Network models have achieved competitive performance in argumentation mining tasks. In particular, transformer architectures have achieved

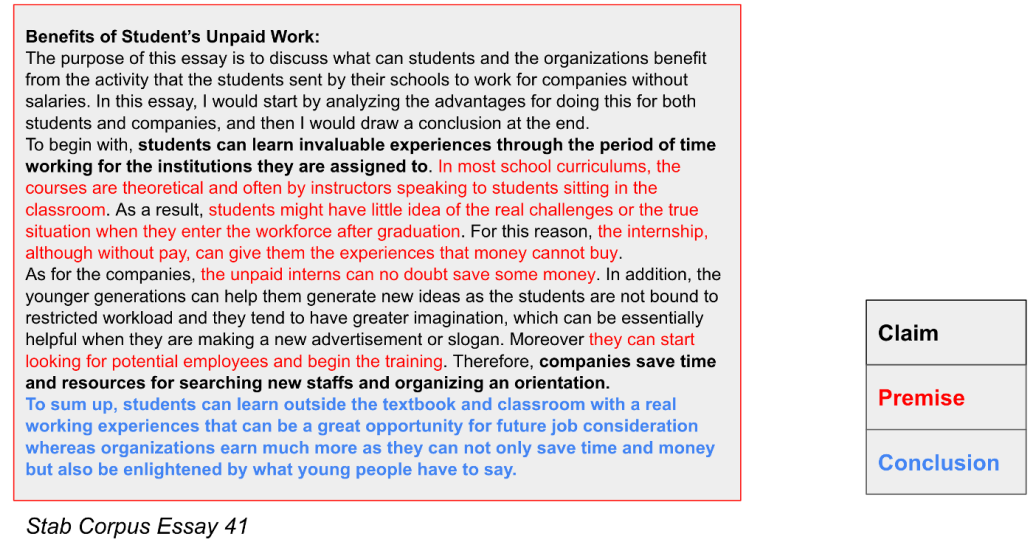


Figure 1: Extraction of Argumentative Structure from Student Essay from *Argument Annotated Essays Corpus* (Stab and Gurevych, 2014)

state-of-the-art performance in (1) **argument classification** tasks, where natural language text has to be classified as being argumentative or non-argumentative, and in (2) **argument clustering** tasks which aim to identify similar arguments.

Reimers et al. (2019) found that **Bidirectional Encoder Representations from Transformers (BERT)** improve the state-of-the-art for the UKP Sentential Argument Mining Corpus by **20.8 %** and for the IBM Debater - Evidence Sentences dataset by **7.4 %**. BERT is a contextualised transformer-based language model that is pre-trained on left and right context in all layers. Vaswani et al. (2017) develop the architecture of a **transformer** that consists of an encoder and decoder. For an input sequence of words in a sentence $w_1, w_2, \ldots, w_n \in S$, the **encoder** is a stack composed of n identical layers that each consists of a **multihead self-attention mechanism** and a feed-forward neural network, ultimately producing an encoder representation. The self-attention mechanism relates word w_i to all preceding words $w_1, w_2, \ldots, w_{i-1}$ producing an embedding x_i. Mathematically, the the self-attention mechanism is scaled dot product attention, producing an attention matrix, Z, given by:

$$Z = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

where Q is the query matrix, K is the key matrix and V is a value matrix.

Devlin et al. (2019) develop BERT's model architecture, drawing on the transformer architecture presented in Vaswani et al. (2017), as a multi-layer *bidirectional* transformer encoder. The bidirectional architecture of BERT is designed to pre-train deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers. BERT is pre-trained for masked language model and next sentence prediction, and can be fine-tuned for various tasks including sentence classification.

3 The *Learn to Argue* Model

We are developing a learning tool that consists of a pre-trained model that automatically analyses argumentation structure in written English and a front-end interface as a Google Doc add-on. The tool provides argumentation highlighting and feedback to the user: it identifies the argumentation claims and premises in the learner text, and provides a general comment about the argumentation. The tool may also provide an assessment of argumentation quality. *Figure 2* depicts the architecture of our argumentation tool.

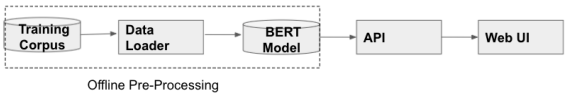


Figure 2: Architecture of *Learn to Argue* Tool

4 Preprocessing

4.1 Dataset

We use the Argument Annotated Essays Corpus (AAEC), developed by Stab and Gurevych (2014). The corpus consists of 402 argument-annotated persuasive essays, and features topic and stance identification, annotation of argument components, and argumentative relations.

As our learning model will be used in educational applications, the corpus provides topic-specific argumentation data for our model to be trained on.

4.2 Data Extraction

As a first step, we transform the corpus text from the original brat annotated file format to a csv format that includes the **argumentative discourse units (ADU)** from each essay and assign an associated ADU Value. Claims are assigned an ADU Value of 2, premises have ADU value of 1, while non-argumentative statements have an ADU value of 0. The pre-processing stage allows us to train a model to distinguish between (1) argumentative and non-argumentative statements and (2) claims and premises. *Figure 3* contains a sample of my spreadsheet.

	A	B
1	**Sentences**	**Value**
2	which would be hard to obtain otherwise	1
3	living your life at a slower pace increases your happiness	2
4	It is very common to see state of the art laboratories in the campus	1
5	I work and study at the university and I don t have spare time during the day	1
6	I came across a situation once such that one of my subordinates used to suspect his co worker for carrying tales to me	1
7	Although it is real in some aspects	0
8	this great invention in the 22st century causes many problems in our society	0
9	TV may take some of your time away from your family and friends but it can also make family get together and help to get in conversation	0
10	wealthy families can afford better health services and costly medicines	2
11	singers and athletes do not play a helpful part in improving the material life	1
12	it will help one to get a friend much easier	2
13	Computer is a modern device that is useful for human life	0
14	Many people find that students at schools and universities still learn much more from lessons with teachers	0
15	Many people believe that students only need to study academic subjects like mathematics	0
16	Creativity should it be given freedom or restrictions	0
17	Life in Big Cities Vs People s health	0
18	People are no longer restricted in the office	0
19	it is generally felt that the access to these tools of communication is available in every corner of the world	1
20	while I believe that other factors contribute a slight incline level of violence	0
21	It would be unwise to still deny the truth and keep focusing on economic developments and unfortunately	1
22	and this lack of communication may impair their future negotiation and interpersonal skills	1
23	focuses on the link between diamonds and conflict	1
24	learning about life through personal experience seems more appropriate	2
25	there are many difficulties a student might face when studying and living overseas	2
26	the root of violence is criminal mind not the weapon itself	1
27	companies continue to improve their products and service	0
28	like some announcements from the boss	1
29	she rarely went to class but spent most time to understand confusing topic	1
30	There are different reasons which persuade that people should not restore old buildings	0

Figure 3: Spreadsheet containing sentences with associated argumentative value

5 Model Selection and Evaluation

We evaluated several pre-trained transformer-based language models which were fine-tuned on the data extracted from the Argument Annotated Essay corpus.

We finetune pre-trained transformer-based models using the *Hugging Face* transformer library. The library contains high performing transformer models that are relatively easy to use. We first tokenize the extracted text using the appropriate tokenizer from the Hugging Face library, which generates an *input ids* and an *attention mask*. These two outputs are fed into the pretrained trained transformer model. We then generate the last hidden states to obtain the **[CLS] token**. The CLS token appears at the beginning of the tokenized representation of each sentence. Although it is a fixed positional embedding and does not contain any information itself, the token's output is inferred by all other words in the sentence. Therefore, the [CLS] token contains information present in other words, and is a good representation for sentence-level classification.

We first trained a simple model on the extracted [CLS] token, before using the whole sequence output to feed into a bidirectional neural network.

As a first step, we wanted to compare the performance of the transformer language models to non-transformer models. I trained a **long-short term memory (LSTM)** recurrent neural network model. The loss curve and the accuracy vs. epoch graphs in *Figure 4* show how the model is learning argumentative structure from the corpus. The accuracy metric is particularly important in this project, as this allows us to determine whether users can receive feedback that *correctly* identifies claims, premises and non-argumentative text.

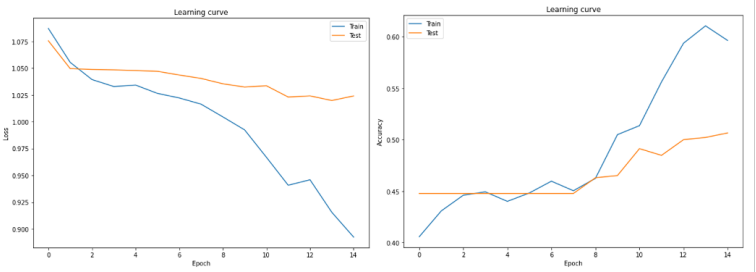
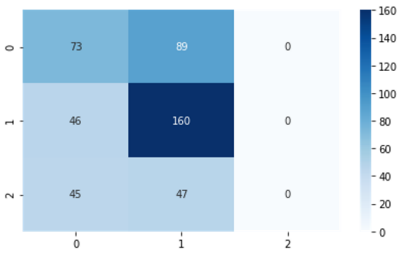


Figure 4: Learning curves for LSTM basline

The confusion matrix shows the model accuracy for claims, premises and non-argumentative units in more detail. The classification report also

contains the precision, recall and F1 score for the model. The F1 score is important in our project, as we are looking for a balance between the precision and recall of our model, and our data has a class imbalance. In any argumentative text, there will always be less claims than premises.

We can see in *Figure 5* that the LSTM is performing particularly poorly for classifying text as claims.



	precision	recall	f1-score	support
0	0.45	0.45	0.45	162
1	0.54	0.78	0.64	206
2	0.00	0.00	0.00	92
accuracy			0.51	460
macro avg	0.33	0.41	0.36	460
weighted avg	0.40	0.51	0.44	460

Figure 5: Classification Report and Confusion Matrix for LSTM Baseline

Next, we finetuned the pretrained BERT model from the Hugging Face transformers library. The learning curves are shown in *Figure 6*. We can see during training that BERT is performing much better than the LSTM baseline, with an accuracy of almost 0.85, compared to the LSTM accuracy of 0.60.

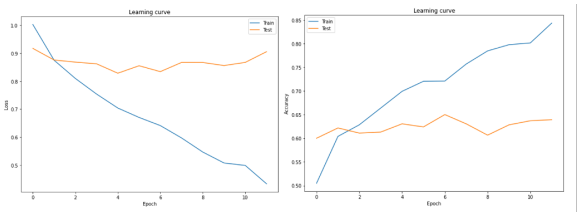
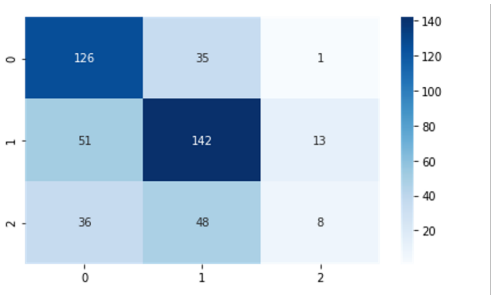


Figure 6: Learning curves for BERT model

The classification report and confusion matrix in *Figure 7* also show that BERT has a higher F1, recall and precision score, notably for claims. This again is an improvement on the LSTM baseline.



	precision	recall	f1-score	support
0	0.59	0.78	0.67	162
1	0.63	0.69	0.66	206
2	0.36	0.09	0.14	92
accuracy			0.60	460
macro avg	0.53	0.52	0.49	460
weighted avg	0.56	0.60	0.56	460

Figure 7: Classification Report and Confusion Matrix for BERT

We then evaluated the performance of RoBERTa. The learning curves in *Figure 8* show that RoBERTa has a slightly lower accuracy than BERT.

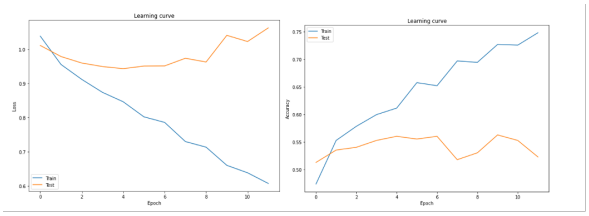
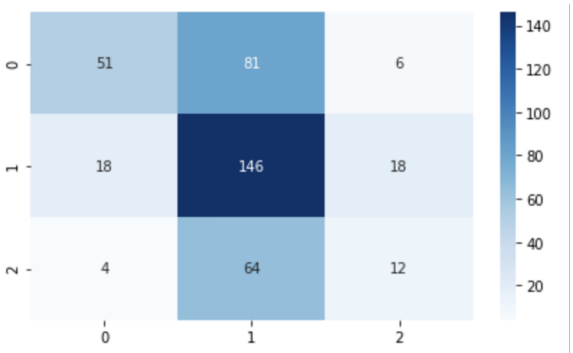


Figure 8: Learning curves for RoBERTa model

However, RoBERTa does have better scores for classifying claims than BERT.



	precision	recall	f1-score	support
0	0.70	0.37	0.48	138
1	0.50	0.80	0.62	182
2	0.33	0.15	0.21	80
accuracy			0.52	400
macro avg	0.51	0.44	0.44	400
weighted avg	0.54	0.52	0.49	400

Figure 9: Classification Report and Confusion Matrix for RoBERTA

The original project scope was to initially try out BERT and RoBERTa. However, I did additionally experiment with more compact transformer models. DistilBERT actually was the best performing model with my original architecture. *Figure 10* shows the model has high accuracy. It is a more compact and efficient model, which makes it easier for deployment. Thus for performance and deployment, the DistilBERT model was particularly convenient.

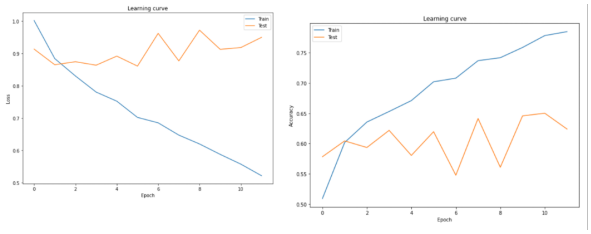
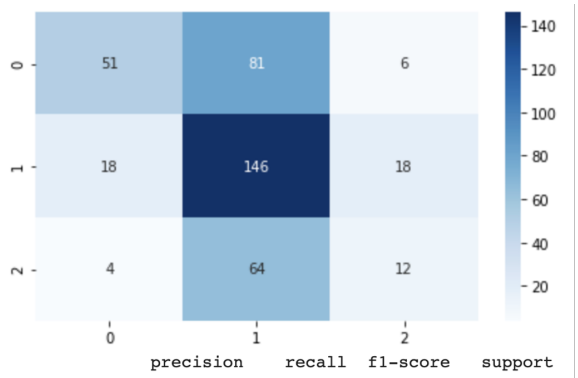


Figure 10: Learning curves for the original DistilBERT model

DistilBERT has higher precision than BERT for 1 and 2, and a higher weighted F1. It is also a smaller model, which overall was a very convenient outcome.



	precision	recall	f1-score	support
0	0.70	0.37	0.48	138
1	0.50	0.80	0.62	182
2	0.33	0.15	0.21	80
accuracy			0.52	400
macro avg	0.51	0.44	0.44	400
weighted avg	0.54	0.52	0.49	400

Figure 11: Classification Report and Confusion Matrix for the original DistilBERT model

When we were deploying the DistilBERT model, we found it more convenient to modify the original architecture. This allowed us to use the Hugging Face `save_pretrained` function to upload the model to Google Cloud. We passed the input encoding from the tokenizer into a Tensorflow dataset object, using the `from_tensor_slices` constructor method.

The modified model achieves a higher accuracy than the original architecture.

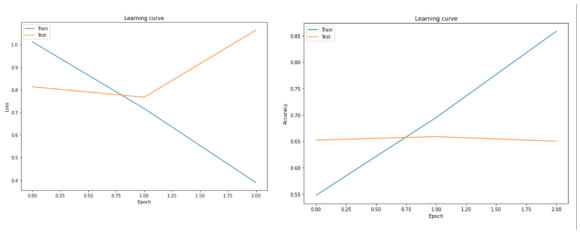


Figure 12: Learning curves for the modified DistilBERT model

I did also try other more compact transformer models Electra and ALBERT, however their performance was not as good as the other models.

We decided to use DistilBERT as the language model for the Learning to Argue model, given its performance and strong evaluation metrics.

6 Model Deployment

The DistilBERT model can generate predictions, classifying the `test_text` as a premise in the code snippet below:

```
test_text = """All of these skills help them to get on well with other people
and will benefit them for the whole life"""
predict_input = loaded_tokenizer.encode(test_text,
                                        truncation=True,
                                        padding=True,
                                        return_tensors="tf")

output = loaded_model(predict_input)[0]

prediction_value = tf.argmax(output, axis=1).numpy()[0]
prediction_value
```

```
1
```

Figure 13: DistilBERT Model Classification of test text as premise

We saved and loaded the distilBERT model to Google Cloud, and are integrating the model into a Google Docs add-on. Users can use the instructions of the add-on to receive argumentative feedback, from which they can improve their essay.

A current prototype of the Learning to Argue add-on is shown below in *Figure 14*.

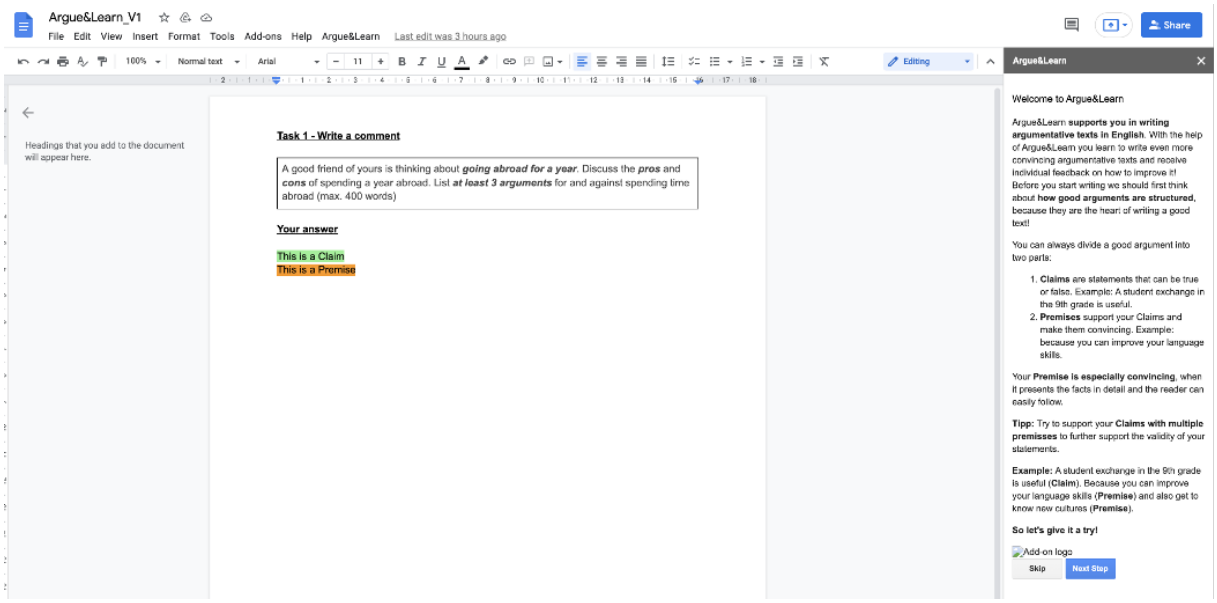


Figure 14: Prototype of Learn to Argue Google Docs add-on

7 Pilot Study

After my model has been deployed, we will run a short pilot study to evaluate our tool. We want to assess whether learners do improve their argumentation skills through the automated adaptive feedback that the learning tool provides. We also want to evaluate how users respond to the learning tool, and whether the tool provides an improved learning experience for English language learners. The pilot study has undergone the review procedure of the Department of Computer Science and Technology Research Ethics Committee at the University of Cambridge

Participants will complete a Google Form, which contains information about the study, pre-study questions, instructions on how to use our learning tool and write an essay, and tool evaluation. All participants are given information about the study, and about how we will use their data.

The study will consist of a 20-25 minute experiment, consisting of a 5 minute pre-study, a 15-20 minute writing phase. In the experiment, participants will write a short essay (around 300 words) on a prescribed title, and edit the essay using the learning tool.

We then ask some preliminary questions about age, language background, education and some self-evaluation questions about technology usage and previous experience with automated writing assistance.

We then include some writing instructions, about how to use our Google Doc add-on. We ask participants to submit a final copy of their essay, however I have suggested that we could incorporate some tool logging so we can see how users actually interact with our tool, and how they use it to improve their essay. Finally, we ask users to evaluate our tool.

Once the tool is usable and performing well, Caines & Wambsganss will continue the data collection project in a German high school, with a group of 25-30 English learners from 3-4 classes between grade 7 to 11. We require parental approval for minors to participate in the study. If students are under 16 years old, they will be asked to conduct the exercise by logging into an institute google account, thus ensuring that the minimum age of managing a Google account in Germany will be fulfilled. We are in contact with a teacher in the school, who has already informed the head teacher and all other relevant stakeholders about the study to ensure the treatments are aligned with the German High School Law and data protection requirements of the state.

8 Discussion

One of the main findings of our research is that neural argumentation techniques can help students learn a skill like argumentation that is implicitly taught in a traditional classroom setting.

However, there is a lot of scope in the argumentative mining field to fully exploit the potential of

machine learning. Lippi and Torroni (2015) highlight the possible application of transfer learning, statistical relational learning and active learning in the argumentative mining pipeline. All of these are productive lines of inquiry, particularly active learning. **Active Learning** is a technique that can help reduce the amount of annotation required to train a model, by focusing the attention of the expert annotator on only cases where their contribution is most informative. The paradigm of Active Learning is particularly important in argumentative mining, where much effort has been placed on developing domain-specific annotated corpuses, like the Argument Annotated Essay corpus for educational argumentation. Active Learning not only presents a viable alternative to the arduous annotation process in argumentation mining, it also allows researchers to leverage annotator disagreement about the nature of argumentative units or their relations to improve the model performance. Fornaciari et al. (2021) found that active learning in models that utilise deep pre-training is particularly effective. As the Learning to Argue model has made use of deep pre-trained transformer-based models, it would be very interesting to see how active learning can be integrated in our project.

A possible area of development in the Learning to Argue project is to allow the model to select a set of Q data points that it is unsure how to classify from the essays collected in the pilot study (i.e a pool of potential unlabelled training elements), so they can be labelled by a human annotator, which could be an advanced English learner who has participated in our study or a teacher in the main school. Grießhaber et al. (2020) use a BERT model in an active learning scenario for low-resource text classification, and found active learning improved model stability and training performance. It would be interesting to see whether we see similar results in argumentation mining.

A consequence of our research on the potentially useful educational applications of argumentation mining is that we can gain a better understanding of how students actually learn implicit skills like argumentation. Through our pilot study and main study, we have been looking at the benefits of automated adaptive feedback in the learning process. I think it would be interesting to further evaluate how the benefits of argumentative writing assistance differ for L2 English learners compared to their native German L1. Wambsganss et al. (2020) have developed a German corpus containing student annotated essays, in the same format as the Argument Annotated Essay corpus. This means that we can develop a German Learning to Argue model in the same way as we have done for English, and compare how participants in our main study interact with both tools.

Our research also motivates the need for the development of argument-specific machine learning techniques. Argumentation as a learning skill has particular nuances that may not be captured by transformer-based models. For example, Becker et al. (2020) note that often important parts of arguments can be omitted by the writer, or are only implicitly implied. Ideally, an argumentative feedback tool should highlight to the author that the clarity of the argument would improve if the a point is made explicitly.

Problems like argument omission and implicit points likely occur due to either over-familiarity with the topic, so the author assumes that a more basic point is also known by a wider audience, or due to lack of understanding. Ideally, we would want a writing support system to distinguish between the two scenarios. In the former case, argumentation models have to determine whether learners are experienced with the topic, or if there are factual errors or other signs of misunderstanding.

Argument omission highlights the situational and context-dependent nature of argumentation. The use of context-dependent embeddings in transformer-based models of argumentation is therefore very relevant. While we do not know exactly *why* transformer-based language models outperform recurrent neural networks (a question that is key theme of contemporary research in NLP), context-dependence is almost certainly a contributing factor to their improved performance. Manning et al. (2020) show that the large improvements brought about by transformer language models in language understanding tasks is due to deep contextual language models implicitly learning emergent linguistic structure. Thus, the transformer-based language models used in the Learning to Argue tool are likely more effective due to their contextualised representations.

However, what our current models lack are *situational embeddings*. Argumentation skills are situation-dependent, and this is one of the major challenges of developing cross-domain and cross-topic argumentation models. Learners may strug-

gle to write persuasive and logical arguments about one topic more than another. A constructivist view of learning would explain this: this view assumes that learning is determined by the experience of the learner. Although the core skills of argumentation may be field-invariant, learners experiences in writing arguments about different topics can differ. Automated adaptive feedback should provide more guidance in cases where learners are struggling to write arguments in certain domains, but do not necessarily need to provide the same level of support in situations where learners accidentally omit claims and premises.

More generally, this issue motivates the need for top-down performance goals in argumentation mining, and in particular we need argumentation mining systems to learn argumentation in a situation dependent sense. To this end, I think argumentation mining could benefit in the long term from the development of constructivist machine learning techniques.

9 Conclusion

During the Learning to Argue project, I have developed a learning tool that provides automatic feedback on argumentation to learners of English, and will be running a pilot study to evaluate the learning tool. We will be running a main study in a German school. For educational technology to be beneficial to learners, it must be integrated within the the learning process of students. Our main finding is that automatic adaptive feedback on argumentation can benefit learners of English. This research on the educational applications of argumentative mining also motivate the creation of argument-specific machine learning techniques, which draw inspiration from human learning.

Acknowledgments

This paper reports on research supported by Cambridge Assessment, University of Cambridge. I would like to thank Paula Buttery and Andrew Caines for giving me the opportunity to work on this UROP Project, and for their support throughout the process. I would also like to thank Thiemo Wambsganss (University of St. Gallen) for helping me throughout the project, and Russell Moore for his Machine Learning classes.

References

Maria Becker, Katharina Korfhage, and Anette Frank. 2020. Implicit knowledge in argumentative texts: An annotated corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2316–2324, Marseille, France. European Language Resources Association.

Robert Dale and Jette Viethen. 2021. The automated writing assistance landscape in 2021. *Natural Language Engineering*, 27(4):511–518.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. 2020. Fine-tuning BERT for low-resource natural language understanding via active learning. *CoRR*, abs/2012.02462.

Marco Lippi and Paolo Torroni. 2015. Argument mining: A machine learning perspective. volume 9524, pages 163–176.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117:30046 – 30054.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

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.