Token-level Classification for Argument Mining with Multitask Learning

Haoyang Ma haoyangma@umass.edu
Rukai Cai rukaicai@umass.edu
Sora Ryu sryu@umass.edu
Yunfei Luo yunfeiluo@umass.edu

Abstract

Writing skill is one of the most significant must-have tenets for people to live in their successful lives. Many recent studies have shown the robustness of the deep learning techniques in Natural Language Processing tasks including transformer based models. Even though diverse pre-trained NLP models are available, overfitting problem is still an inevitable issue for the downstream tasks. In this project, we focus on modeling for the task of argumentative and rhetorical elements segmentation. We adopt multitask learning approach with the hypothesis that related tasks could not only benefit but also regularize each others. The experimental results shows that our proposed multitask-based model outperform the baseline, especially when related dataset and tasks are introduced.

1 Introduction

The continuing growth in the volume of data refers to a high potential for a number of meaningful information extraction, but more efforts would be needed to unlock the wealth of information that the data contains.

Argument mining deals with automatic identification of argument structures and its relationships from natural language text. Besides its basic task of component segmentation and classification, it has been applied to many different tasks such as identifying argumentative discourse types (Stab and Gurevych, 2014), assessing the quality of argumentation (Wachsmuth et al., 2016) and classifying argumentative relations (Jo et al., 2021). Argument mining approaches have been extended to various practical applications with great importance, including the user-generated web discourse (Habernal and Gurevych, 2016), political debates (Haddadan et al., 2019), legal cases (Mochales and Ieven, 2009), and conversational agents which gives feedback of structural wrongness in arguments (Mirzababaei Behzad, 2021).

Nowadays, writing skill is growing in importance for people to be successful. It is significant for students not only for their academic works but also enhancing their communication and thinking abilities. Especially, argumentation is essential in academic writing which can show their logical thinking and reasonable opinions. However, National Assessment of Educational Progress has reported that only less than a third of high school seniors are proficient in their writing. Moreover, existing automated writing feedback tools have limitations in terms of failing its structures and their expensive cost. Therefore, in this paper, we propose our own multitask learning model for argumentative and rhetorical element classification, which could further develop students' writing skills via automatically evaluating the structures of their essays.

2 What you proposed vs. what you accomplished

- ~~Collect and preprocess dataset~~
- ~~Build and train (specific baseline model) on collected dataset and examine its performance~~
- ~~Build and train the proposed multitask learning model and examine its performance~~
- ~~Tuned Hyper-parameters and tried several variations of the proposed multitask learning model~~
- ~~Explored the behaviors of the model when new datasets and tasks are introduced.~~
- ~~Use CoNLL-2003 as a training dataset for transfer learning.~~

- ~~Compared the baseline model and the Multitask Learning with CoNLL-2003 model and do Error Analysis~~

- *Try different backbone model, e.g. Roberta*: Due to limited computational resource.

- *Adversarial training*: Hard to build a enough adversarial training set due to limited time.

3 Related work

(Wambsganss et al., 2020) Previous research studies on argument mining involve several subtasks: (1) argument identification, which identifies whether each part of arguments are argumentative or non-argumentative (Florou et al., 2013; Petasis, 2019), (2) argument component classification, where argumentative text is classified into claims and premises (Rooney et al., 2012), or claim, warrant and evidence (Lugini and Litman, 2019), or claim, backing, premise, rebuttal and refutation (Habernal and Gurevych, 2016) based on the modified Toulmin's model (Toulmin, 2003) (3) argument relation classification, which identifies the relation among the argument components whether it is support or attack (Boltuzic and najder, 2014), or major claim, claim or premise (Niculae et al., 2017). Our task lies on the argument component classification task.

(Alhindi and Ghosh, 2021) recently applied neural network based model to argument component segmentation, by reforming the task as token level classification, where claim and premise tokens are identified with BIO notation ("B-Claim", "I-Claim", "B-Premise", "I-Premise", "O-Arg"). Pre-trained BERT model was used for transfer learning by fine-tuning on this downstream task, and also performed the experiment with multitask learning setting, where binary task of sentence level argument identification is added. Based on the findings of this paper that a multi-task architecture outperformed other models, we also employ multi-task learning architecture on our approach, but add more tasks on it and classify with more argument components.

(Caruana, 2004) demonstrated that multitask learning can improve the performance of each task by leveraging the information in training signals from other related tasks, and therefore, even the generalization performance can be improved. (Tran and Litman, 2021) conduct argument mining under the scenario of online persuasive discussion, and present the potential improvement with multitask learning approach. Inspired by those findings, we explore the multitask learning through adding two coarser classification tasks from original dataset (Georgia-State-University, 2021), but also include two other datasets: DBPedia (Auer, 2007) and CoNLL-2003 (Erik F. Tjong Kim Sang, 2003). Both datasets are expected to be highly related to our main task, as DBPedia provides three level hierarchical multi-label text classification, and CoNLL-2003 provides multitask learning for NER, POS and Chunking.

With much empirical evidence that shown the robustness of attention mechanism (Vaswani et al., 2017) in the field of NLP, the modeling for the downstream tasks becomes effective. For example, (Devlin et al., 2018) proposed the well-known transformer based model, BERT, that have updating the state-of-the-art performance for many classical NLP tasks. However, as BERT only take 512 tokens at once, with punctuation and words splitted into sub-tokens, it is not applicable to take essays as input with hundreds of words. (Beltagy et al., 2020) proposed LongFormer, a variation of the transformer based model, that could take 1024 and even more tokens at once. With the best of our knowledge, this model achieve the state-of-the-art performance on modeling tasks with long document.

4 Method

4.1 Problem Setup

The dataset we are using is from Kaggle, (Georgia-State-University, 2021). We followed (Alhindi and Ghosh, 2021) to form the modeling task as NER. In our case, there are 7 categories representing the discourse element type as "Claim, Counterclaim, Rebuttal, Lead, Position, Evidence and Concluding Statement". In order to differentiate the segments, the categories are further split into "begin" and "inside". For example, "claim" is split into "claim_b" and "claim_i" representing whether or not the token is the beginning of the segment. As a result, there are 15 unique labels with an extra outlier class.

The texts are tokenized with the pre-trained tokenizer of transformer based models like (Devlin et al., 2018) and (Beltagy et al., 2020). Each token will be assigned to one of the 15 component categories. More specifically, our models take input

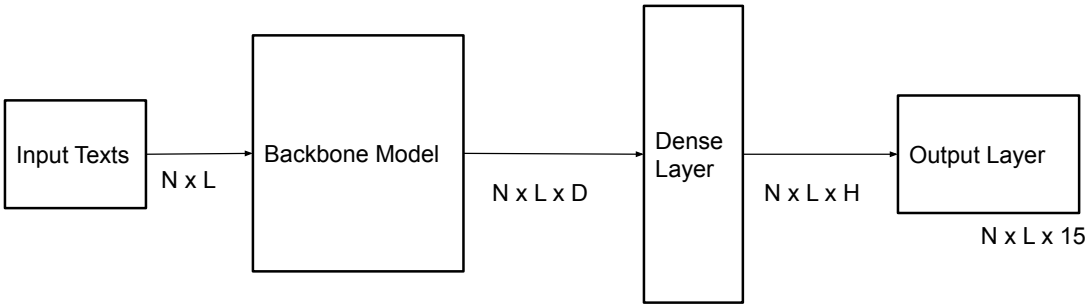

Figure 1: The Baseline model, a transfer learning approach, with LongFormer as backbone

texts

$$X \in R^{N \times L}$$

where N is the batch size and L is the number of tokens, and output

$$f(X, \theta) \in R^{N \times L \times C}$$

where $C = 15$ in our case, and $f(X, \theta)$ represents the model with feed forward function f and parameters θ. We use cross entropy loss to optimize the parameters.

4.2 Transfer Learning from LongFormer

BERT (Devlin et al., 2018), which is one of the well-known transformer based models, takes 512 tokens at once. Such configuration reveal the limitation when applying the model to long text. In order to address this limitation, LongFormer (Beltagy et al., 2020) is set as the backbone of our modeling task in this project. This backbone model could not only take input with longer sequence length, but also address the computational efficiency by adopting dilated sliding window for attention pattern. Figure 1 presents the pipeline of our baseline method.

4.3 Multitask Structure

4.3.1 Coarse Classification

As (Tran and Litman, 2021) have shown that multitask structure could improve the performance of neural networks on argument mining task, we adopt the similar approach. We consider a related task with coarse labels rather than 15 fine labels in our main task. The labels "Claim, Counterclaim, Rebuttal" are considered as "Argument", "Lead, Position, Concluding Statement" are considered as "Declaration", and "Evidence" remain unchanged. As a result, we will have 7 labels in total with the labeling setting described in section 4.1.

More specifically, given the one-hot embedded ground truth of our main task

$$Y_{main} \in R^{N \times L \times 15}$$

and ground truth of the coarse task

$$Y_{coarse} \in R^{N \times L \times 7}$$

with the loss function L, we have designed our integrated loss function

$$\alpha \cdot L(f(X, \theta), Y_{main}) + \beta \cdot L(f(X, \theta), Y_{coarse})$$

where α and β are hyper-parameters representing the weights set to each task. This integrated loss is used for optimizing the model parameters θ, which means that all the tasks are trained together.

4.3.2 Binary Classification

Further breaking down the coarse task, we are also thinking about adding a binary classification task, where each token is either the "begin" or "inside"

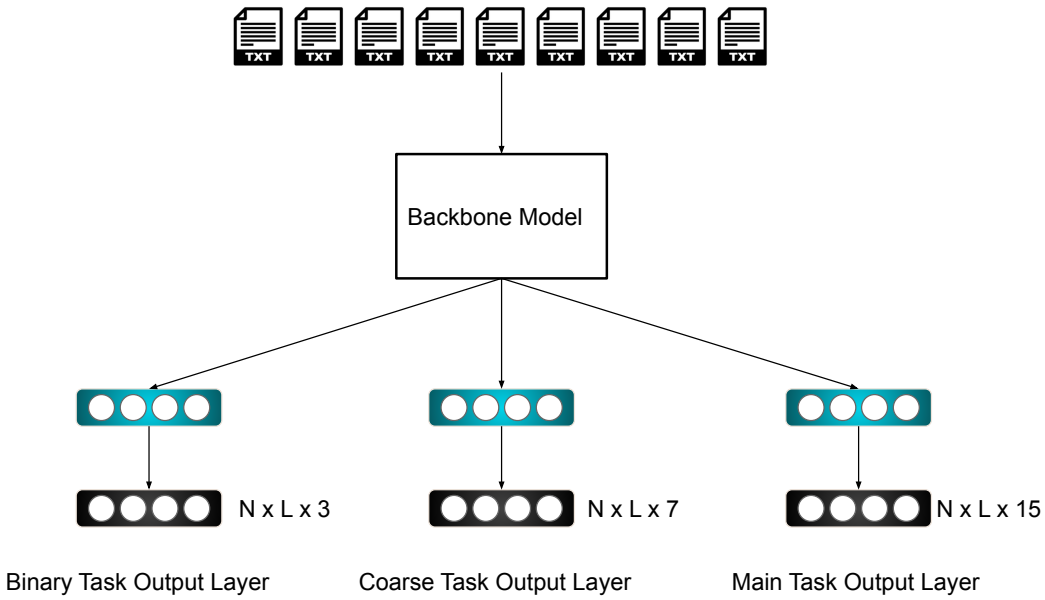

Figure 2: Overview of the multitask learning model. N is the batch size, L is the number of tokens, D as the number of output features from backbone model, and H as the hidden size for the downstream task.

of a segment. The setting of loss function stay consistent with section 4.3.1. Figure 2 provide an overview of our proposed model.

4.4 Introducing External Datasets and Tasks

(Vu et al., 2020) have shown that a NLP task can benefit from transfer training from related tasks. In this project, we explore the behavior of adopting multitask learning structure based on the tasks clustering shown in (Vu et al., 2020).

4.4.1 Multitask Learning with Wiki-Pedia

Since the essays are written by students on some specific topics, we made a hypothesis that by training with topic classification tasks, our main task could get benefit.

Figure 3 shows the workflow of our proposed method, where all the tasks are trained simultaneously.

4.4.2 Multitask Learning with Highly Related Tasks

According to (Vu et al., 2020), our main task lie in the domain of sequence-labeling, where we are doing token-level classification. Thus, by following the empirical conclusion reached by (Vu et al., 2020) after clustering the tasks based on task-embedding, as shown in Figure 4, we decide to let our model multitasking on the tasks like Name-Entity-Recognition (NER) and Part-of-Speech (POS).

The dataset we have chosen for this approach is CoNLL-2003 (Erik F. Tjong Kim Sang, 2003), where for each sample sentence, each word have labels corresponding to three tasks: NER, POS, and Chunking. The entire workflow is slightly different with the one for multitasking with WikiPedia dataset, where all the sub-modules for each task are designed for token-level classification, as shown in Figure 5.

4.4.3 Transfer Learning with Highly Related Tasks

Since we are doing token-level classification, this means we can also implement a transfer learning model compared to multitusk learning model. Instead of learning from the CoNLL-2003 and our dataset simultaneously, We first build a classification model based on LongFormer and trained it on CoNLL-2003. After training, the weight of the LongFormer in our model is saved and reused to fine-tune a new classification model on our dataset.

4.5 Datasets

(Auer, 2007) DBPedia is a dataset project aiming to extract structured content from the information created in Wikipedia. This is an extract of the

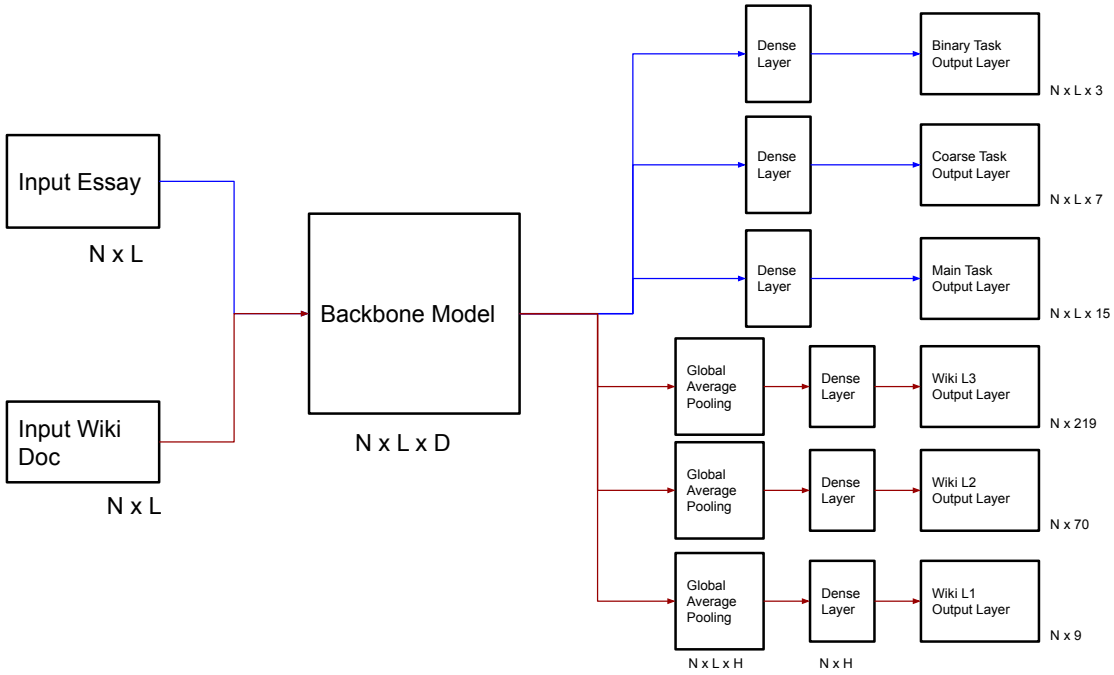

Figure 3: Multitasking with multi-label text classification

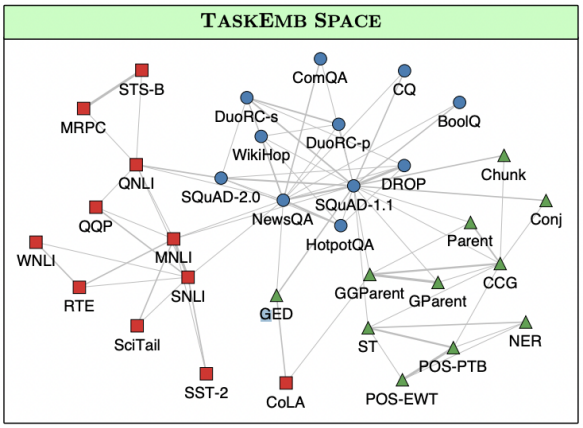

Figure 4: Tasks clustering, figure from (Vu et al., 2020)

data (after cleaning, kernel included) that provides taxonomic, hierarchical categories (”classes”) for 342,782 wikipedia articles. There are 3 levels, with 9, 70 and 219 classes which is an excellent benchmark for hierarchical multiclass/multilabel text classification respectively.

(Erik F. Tjong Kim Sang, 2003) CoNLL-2003 is a dataset of language-independent named entity recognition. The data consists of eight files covering two languages: English and German. For each of the languages there is a training file, a development file, a test file and a large file with unannotated data. It concentrates on four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups.

(Georgia-State-University, 2021) is a dataset of essays written by students from grade of 6 to 12. The dataset is collected by Georgia State University and The Learning Agency Lab, which focused on developing science of learning-based tools and programs for social good. This dataset contains 15k training samples, and 10k held-out test samples. Each sample is a complete essay, and the labeling of the argumentative components are provided.

For research purpose, we will split 20% essays out from the training set as a validation set for evaluating our approaches from the baseline. And for all the dataset, we clip the number of samples to be same with the (Georgia-State-University, 2021) dataset, for multitasking pipeline.

5 Results

5.1 Preliminary Results

Before we conduct our main experiments, we wanted to check how the models perform differently after some few hyperparameter tuning stage. The preliminary result, depicted in the Table 1, shows that LongFormer Multi task learning model with hidden layers for each task outperformed other configurations. All the experiments we have conducted are done with the setting of train-validation split with fraction of 80%

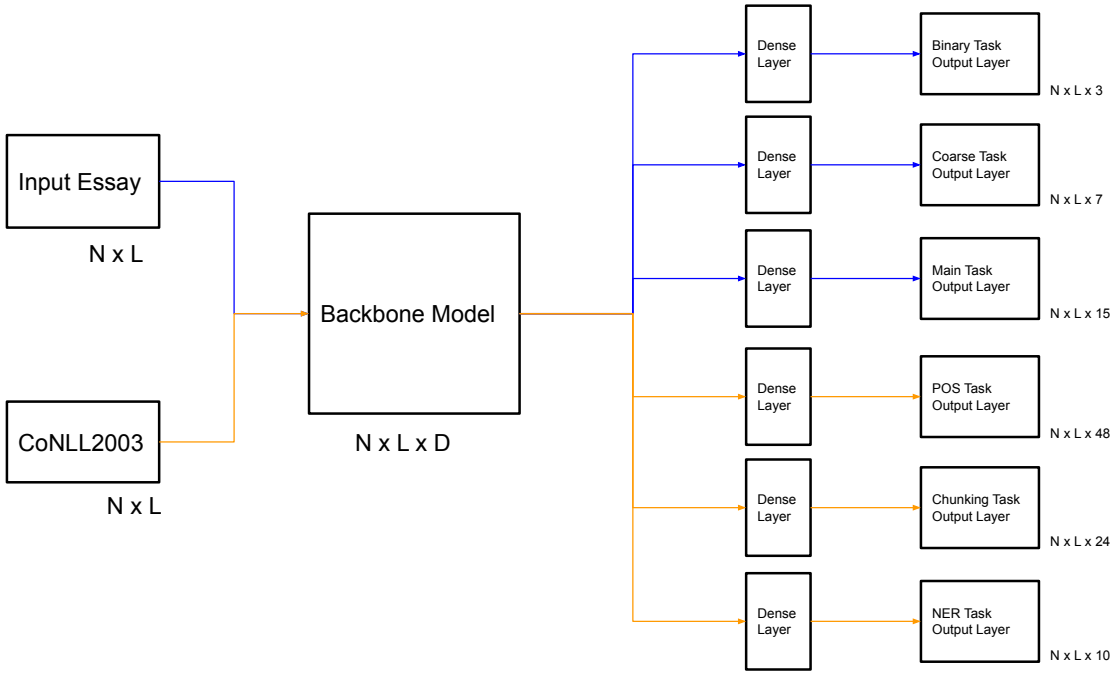

Figure 5: Workflow for multitasking with CoNLL-2003 on the tasks of NER, POS, and Chunking

Model	Val. Acc.(%)
LongFormer (baseline)	70.48
LongFormer-Multi(Single Hidden)	70.67
LongFormer-Multi-Hidden	**70.74**
Multitask (DBPedia classification)	70.57
CoNLL 1st, then Multitask with topics	70.49

Table 1: The preliminary result shows that the multi-hidden LongFormer multitasking model achieves the highest validation accuracy among 5 candidates.

and 20%.

5.2 Empirical Results

With this insight, we conducted our regular experiments which include single task learning, which was our baseline model, multi task learning, multi task learning with WikiPedia dataset, multi task learning with CoNLL-2003 dataset and transfer learning from CoNLL-2003 dataset. The experiment result is shown in the Table 2.

From our main experiment, we have found out that multitask learning is effective. As we compare the last two models, where the former one did multitask learning with CoNLL-2003 dataset and the latter one did transfer learning from the same dataset, the multitask learning was much more effective for most of the tasks, except for POS tagging. Furthermore, when we compare 5 models' performance for the main labeling task, multitask learning with only our own dataset was already effective.

In conclusion, our main task benefits more from the proposed multitask learning structure than the transfer learning approach. And the more related the sub-tasks are, the better the performance of the main task.

Tasks	Single	Multi	Multi + wiki	Multi + CoNLL	Transfer from CoNLL
Main Labeling(15 classes)	70.51	70.58	70.53	70.57	70.39
Coarse Labeling(7 classes)	71.53	71.51	71.44	71.51	71.33
Binary Labeling(3 classes)	78.17	78.18	78.16	78.17	78.16
Wiki L3(219 classes)	-	-	91.38	-	-
Wiki L2(70 classes)	-	-	93.70	-	-
Wiki L1(9 classes)	-	-	98.32	-	-
NER	-	-	-	99.84	99.20
POS	-	-	-	99.23	99.28
Chunking	-	-	-	99.36	99.29

Table 2: The result of our experiments. The numbers show each of the validation categorical accuracy after 5 epoch. For reliable result, we have conducted each experiment three times, and these are the average performance of them.

6 Error analysis

We conducted error analysis on baseline model and multitask learning model which was trained with CoNLL-2003 dataset. For each model, 5 different essays were evaluated. In overall, both models could easily predict the "Lead_b", "Lead_i", "Concluding statement_b" and "Concluding statement_i" labels, as they are comparably easier to recognize and has some positional pattern in the paragraph such as beginning and end respectively.

On average, the baseline model showed some unstable performance on token-level classification, such as interrupting by "Evidence_i" in the sentence consisted of "Claim_b" and "Claim_i". Also, multitask learning model was able to better understand the contextual structure rather than the baseline model, as demonstrated in Figure 6. In the middle of the essay for this example, the essay supports the claim by explaining that letting students to know when they are allowed to use cell phones can help them to focus on their studies. However, the baseline model didn't understand the contextual meaning of them, but simply relied on the word 'However' and end up making the mis-classification. In this sense, we could realize that multitask learning model is much more able to grasp the semantic meaning of each sentence and understand the relationship between consecutive sentences, resulting in a higher performance in overall.

The Table 3 shows the representative example that both baseline model and multitask learning model failed at. While the essay is simply consisted of "claim" and "evidence", both models predicted that it has some more diverse structure of "claim", "counter claim" and "rebuttal". We think that this is a challenging example for model to predict, because it needs to understand the overall contextual meaning as well as grasp the overall flow of the structure. The part of this essay is explaining that texting and driving is also dangerous as much as driving under the influence which we normally think as the riskiest driving habit by showing a statistical research study as an evidence to support its claim. The next two sentences are only for the support for the claim, and trying to give a new insight that texting and driving is also very dangerous habit. Therefore, we should not exaggerate in a way that texting and driving is way more dangerous than driving under the influence. Both are hazardous driving habits, and we should not conclude that driving under the influence is less dangerous based on the less number of accidents it causes. In other words, it would be too much overdoing to recognize the following sentences as counterclaim and rebuttal, considering that the second sentence is not completely opposing the claim. However, the performance of model especially multitask learning was satisfying, considering that this was actually difficult example for us as well.

There were some poor quality of annotation in the dataset, such as one token is having duplicated labels. Also, there were some cases that the label ended with '_b' which means the beginning of the argument was annotated in the verb, not in the beginning of the sentence. For our next future work, we would like to annotate each token more precisely so that even higher improvement of the performance can be expected.

Figure 6: An example of wrong prediction made by the baseline model. We visualized text with each label to offer much easier interpretation.

Original Text	Ground truth	Baseline	Multitask
Texting	claim_b	claim_b	claim_b
and driving is just dangerous as	claim_i	claim_i	claim_i
driving	claim_i	evidence_i	claim_i
under the influence.	claim_i	claim_i	claim_i
Some	other	other	other
people	evidence_b	counterclaim_b	counterclaim_b
would agree that driving while drunk is most common and dangerous thing you can do.	evidence_i	counterclaim_i	counterclaim_i
However,	evidence_i	rebuttal_b	rebuttal_b
The National Safety Council states that someone texting while driving is six times more likely to cause an accident than someone driving drunk (”On the Road”).	evidence_i	rebuttal_i	rebuttal_i

Table 3: The error analysis for another example that both baseline model and multitask learning model with CoNLL-2003 dataset failed at. Note that this is just the part of the essay.

7 Contributions of group members

We discussed ideas together, and every one was working on reading papers and coding documentations. We also integrated the report together.

- Haoyang Ma: Conducting experiments of transfer learning, and help with error analysis.
- Rukai Cai: Preparing datasets with preprocessing, and running repeated experiments
- Sora Ryu: Error analysis and lots of writing
- Yunfei Luo: Constructing pipeline and conducting experiments of Multitask Learning.

8 Conclusion and Future Works

In this project, we explore the behavior of the multitasking on varies datasets and tasks. We aim to find better model for the task of argumentative segmentation on essays written by students from grade 6 to 12. With solid empirical evidence, our proposed deep learning workflow outperform the baseline, and achieve considerable token-level accuracy of 70.58%.

For future work, we expect to explore more technique to improve the performance of our proposed platform, such as trying different regularization techniques to avoid quick overfitting, and adversarial training to enhance the robustness of the model.

References

Alhindi, T. and Ghosh, D. (2021). "sharks are not the threat humans are": Argument component segmentation in school student essays. *CoRR*, abs/2103.04518.

Auer, S. (2007). Dbpedia: a nucleus for a web of open data. https://dl.acm.org/doi/10.5555/1785162.1785216sec-ref.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Boltuzic, F. and najder, J. (2014). Back up your stance: Recognizing arguments in online discussions. In *ArgMining@ACL*.

Caruana, R. (2004). Multitask learning. *Machine Learning*, 28:41–75.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Erik F. Tjong Kim Sang, F. D. M. (2003). Introduction to the conll-2003 shared task: language-independent named entity recognition. https://dl.acm.org/doi/10.3115/1119176.1119195.

Florou, E., Konstantopoulos, S., Koukourikos, A., and Karampiperis, P. (2013). Argument extraction for supporting public policy formulation. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 49–54, Sofia, Bulgaria. Association for Computational Linguistics.

Georgia-State-University (2021). Feedback prize - evaluating student writing. https://www.kaggle.com/c/feedback-prize-2021/overview.

Habernal, I. and Gurevych, I. (2016). Argumentation mining in user-generated web discourse. *CoRR*, abs/1601.02403.

Haddadan, S., Cabrio, E., and Villata, S. (2019). Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Jo, Y., Bang, S., Reed, C., and Hovy, E. (2021). Classifying argumentative relations using logical mechanisms and argumentation schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739.

Lugini, L. and Litman, D. J. (2019). Argument component classification for classroom discussions. *CoRR*, abs/1909.03022.

Mirzababaei Behzad, P.-S. V. (2021). Developing a conversational agent's capability to identify structural wrongness in arguments based on toulmin's model of arguments. volume 4.

Mochales, R. and Ieven, A. (2009). Creating an argumentation corpus: Do theories apply to real arguments? a case study on the legal argumentation of the echr. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 21–30, New York, NY, USA. Association for Computing Machinery.

Niculae, V., Park, J., and Cardie, C. (2017). Argument mining with structured svms and rnns. *CoRR*, abs/1704.06869.

Petasis, G. (2019). Segmentation of argumentative texts with contextualised word representations. In *Proceedings of the 6th Workshop on Argument Mining*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Rooney, N., Wang, H., and Browne, F. (2012). Applying kernel methods to argumentation mining. In *FLAIRS Conference*.

Stab, C. and Gurevych, I. (2014). Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge University Press, 2 edition.

Tran, N. and Litman, D. (2021). Multi-task learning in argument mining for persuasive online discussions. In *Proceedings of the 8th Workshop on Argument Mining*, pages 148–153, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Vu, T., Wang, T., Munkhdalai, T., Sordoni, A., Trischler, A., Mattarella-Micke, A., Maji, S., and Iyyer, M. (2020). Exploring and predicting transferability across nlp tasks.

Wachsmuth, H., Al-Khatib, K., and Stein, B. (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, Osaka, Japan. The COLING 2016 Organizing Committee.

Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Handschuh, S., and Leimeister, J. M. (2020). *AL: An Adaptive Learning Support System for Argumentation Skills*, page 1–14. Association for Computing Machinery, New York, NY, USA.