Opinion diversity through hybrid intelligence

Meer, M.T. van der

3

Will It Blend? Mixing Training Paradigms & Prompting for Argument Quality Prediction

This Chapter describes our contributions to the Shared Task of the 9th Workshop on Argument Mining (2022). Our approach uses Large Language Models for the task of Argument Quality Prediction. We perform prompt engineering using GPT-3 and investigate the training paradigms of multi-task learning, contrastive learning, and intermediate-task training. We find that a mixed prediction setup outperforms single models. Prompting GPT-3 works best for predicting argument validity, and argument novelty is best estimated by a model trained using all three training paradigms.

Michiel van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Báez Santamaría. 2022. Will It Blend? Mixing Training Paradigms & Prompting for Argument Quality Prediction. In *Proceedings of the 9th Workshop on Argument Mining*, pages 95–103, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

3.1 Introduction

As debates are moving increasingly online, automatically processing and moderating arguments becomes essential to further fruitful discussions. The research field of automatic extraction, analysis, and relation detection of argument units is called Argument Mining [AM, 224]. The shared task of the 9th Workshop on Argument Mining (2022) focuses on argument quality [416]. Argument quality can be broken down into multiple dimensions, each with its own purpose, or be extended to *deliberative quality* [414]. In this work, we consider two aspects of the *logical* argument quality dimension: *validity* and *novelty*. Given a premise and a conclusion, a valid relationship indicates that sound logical inferences link the premise and conclusion. A novel relationship indicates that new information was introduced in the conclusion that was not present in the premise.

Prediction of an argument's validity and novelty can be either through binary classification (Task A) or by explicit comparison between two arguments (Task B). We focus on Task A. A system that is able to estimate validity and novelty could be a building block in AM for online deliberation. For instance, in assisting humans to detect arguments in online deliberative discussions [121, 398] or presenting diverse viewpoints to users in a news recommendation system [318]. We address the task of validity and novelty prediction through a variety of approaches ranging from prompting, contrastive learning, intermediate task training, and multi-task learning. Our best-performing approach is a mix of a GPT-3 model (through prompting) and a contrastively trained multi-task model that uses NLI as an intermediate training task. This approach achieves a combined Validity and Novelty F_1-score of 0.45.

3.2 Related Work

Given the two related argumentation tasks (novelty and validity), a Multi-Task Learning (MTL) setup [83] is a natural approach. Multi-task models use training signals across several tasks, and have been applied before in argument-related work with Large Language Models (LLMs) [73, 222, 389]. We use shared encoders followed by task-specific classification heads. The training of these encoders was influenced by the following two lines of work.

First, intermediate task training [309, 430] fine-tunes a pre-trained LLM on an auxiliary task before moving on to the final task. This can aid classification performance, also in AM [351]. Second, contrastive learning is shown to be a promising approach [10, 301] in a previous AM shared task [131]. Contrastive learning is used to improve embeddings by forcing similar data points to be closer in space and dissimilar data points to be further away. Such an approach may cause the encoder to learn dataset-specific features that help in downstream task performance.

In addition to MTL, we look at prompt engineering for LLMs, which has shown remarkable progress in a large variety of tasks in combination with [58] or without few-shot learning [364]. For this task we draw inspiration from ProP [8], an approach that ranked first in the "Knowledge Base Construction from Pre-trained Language Models" challenge at ISWC 2022.[1] ProP reports the highest performance with (1) larger LLMs, (2) shorter prompts, (3) diverse and complete examples in the prompt, (4) task-specific prompts.

[1] LM-KBC, `https://lm-kbc.github.io/`

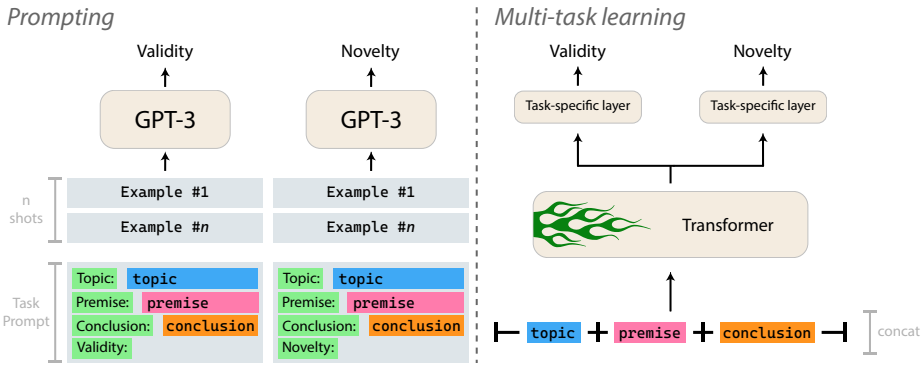


Figure 3.1: The two argument quality prediction setups used in our approach. At inference time, predictions from different setups may be mixed.

Split	Size	Distribution	Topics	Topic Overlap	
				w. train	w. dev
train	750	331/18/296/105	22	–	0
dev	202	33/44/87/38	8	0	–
test	520	110/96/184/130	15	0	8

Table 3.1: Shared task data overview. **Distribution** indicates the class distribution of {non-valid, non-novel}/{non-valid, novel}/{valid, non-novel}/{valid,novel} counts. The red count indicates a severe data imbalance in the training set.

3.3 Data and Training Paradigms

3.3.1 Data

The task data is in American English and consists of Premise, Conclusion, Topic, and a Novel and Validity label. As highlighted in Table 3.1, arguments that are both non-valid and novel are underrepresented in the data. We use the original training and validation distribution as provided and do not use any over- or undersampling strategies. Instead, we opt to resolve the data imbalance by adopting different training paradigms (see Section 3.3.2).

The content included in the dataset concerns common controversial issues popular on debate portals [144], with topics varying from "TV Viewing is Harmful to Children" to "Turkey EU Membership." The training data also contains classes labeled "defeasibly" valid and "somewhat" novel, which are not in the development or test set. We map these to negative labels (i.e. not novel or not valid) to refrain from discarding data. However, we do not measure the effect of this decision on performance.

3.3.2 Training Paradigms

In our work, we mix different training paradigms to obtain our final approach. A schematic overview is given in Figure 3.1. Below, we outline each of the paradigms individually.

Multi-task Learning Since both validity and novelty are related, a shared encoder is used to process the text input into an embedding, which is fed to task-specific layers. We do not use any parameter freezing, allowing gradients from either task to pass through the entire encoder. During training, a single task is sampled uniformly at random, and a batch is sampled containing instances for that task.

Intermediate task training In our case, we use two related tasks for intermediate task training: Natural Language Inference (NLI) and argument relation prediction. For NLI, we use a released RoBERTa model [248] trained on the MNLI corpus [433], predicting whether two sentences show logical entailment. This is related because making sound logical inferences plays a role in validity. The released argument relation RoBERTa model [327] was trained on the relationship (inference, contradiction, or unrelated) between two sentences in a debate [415]. This is related to novelty and validity. For instance, unrelated arguments may be novel but not valid, and vice versa.

Contrastive Learning We use SimCSE's [134] supervised setting to further fine-tune the previously mentioned RoBERTa MNLI model in a contrastive manner. To train the model we take triples of premises and conclusions in the form of premise, conclusion with a positive novelty rating, and conclusion with a negative novelty rating.

3.4 Approach

Approach 1: GPT-3 Prompting In our prompt-engineering approach, we use OpenAI's GPT-3[2] [58] for few-shot classification of novelty and validity labels. We construct a prompt by concatenating the topic, premise, and conclusion in a structured format, and request either a validity or novelty label in separate prompts. In addition, we show four static examples before asking for a label from the model, selected from short, difficult examples (i.e. those with the lowest annotation agreement) in the training dataset.

Approach 2: NLI as Intermediate-task, Contrastive learning and Multi-Task Learning This model consists of a shared encoder with task-specific classification heads. We initialize the shared encoder using a pretrained RoBERTa model on the MNLI corpus. We then perform contrastive learning with a triplet loss. Afterward, the model is fine-tuned using MTL on the shared task training data. During training, we switch uniformly at random during training between the novelty and validity tasks.

Approach 3: Mixing Approach 1 (GPT-3) & Approach 2 (NLI+contrastive+MTL) Our Mixed Approach uses Approach 1 (prompt engineering) for validity labels, and Approach 2 (fine-tuned model) for novelty labels.

Approach 4: ArgRel as Intermediate-task and Multi-Task Learning This model uses intermediate-task training on the argument relation prediction task followed by Multi-Task Learning in the same set-up as in Approach 1, but without contrastive learning.

[2] https://beta.openai.com/playground

Model	F1		
	Validity	Novelty	Combined
SVM (TF-IDF + stemming)	0.60	0.08	0.21
GPT-3 (CLTeamL-1)	0.75	0.46	0.35
NLI+contrastive+MTL (CLTeamL-2)	0.65	0.62	0.39
GPT-3 & NLI+contrastive+MTL (CLTeamL-3)*	**0.75**	**0.62**	**0.45**
ArgRel+MTL (CLTeamL-4)	0.57	0.59	0.33
GPT-3 & ArgRel+MTL (CLTeamL-5)	0.75	0.59	0.43

Table 3.2: Test set performance. CLTeamL-*n* indicates an official submission to the Shared Task with *n* corresponding to the Approach number also in Section 3.4. Bold scores indicate the best-performing approach in the shared task. "Combined" indicates the Shared Task organizer's scoring metric for both tasks.

Approach 5: Mixing Approach 1 (GPT-3) & Approach 4 (ArgRel+MTL) This approach uses Approach 1 (prompt engineering) for validity and Approach 4 (ArgRel+MTL) for novelty labels.

Baseline: SVM Support Vector Machines (SVMs) are strong baselines for argument mining tasks with relatively small multi-topic datasets [319]. We train an SVM separately for validity and novelty as a competitive baseline.

3.4.1 Implementation details

We use Python3 and the HuggingFace `transformers` [436] framework for training our models. The SVM baseline instead uses sklearn [299]. Our code is publicly available.[3] All models trained use RoBERTa (large) [248] as the base model, and the intermediate task trained models are obtained directly from the HuggingFace Hub.[4] We provide hyperparameters for fine-tuned trained models in Appendix B.1. Model selection was done based on the combined (validity and novelty) F_1 performance on the development set. All experiments were run for 10 epochs, after which the best-performing checkpoint was selected for use in creating predictions on the test set. The training was performed on machines including either two GTX2080 Ti GPUs, or four GTX3090 GPUs.

3.5 Experiments and Results

We compare our approaches' performance on the test set with the shared task's metric: Combined F_1 of Validity and Novelty [165]. This combined score is the macro F_1 for predicting validity and novelty in four combinations (valid and novel, valid and not novel, not valid and novel, not valid and not novel). Additionally, we analyze our approaches' errors and their connection to labels, annotator confidence, and topic. See Table 3.2 for performance on the test set. We also present an SVM-based approach as a baseline.

[3] `https://github.com/m0re4u/argmining2022`
[4] `https://huggingface.co/`

Model	F1 Validity		F1 Novelty	
	valid	non-valid	novel	non-novel
GPT-3	0.81	0.68	0.26	0.66
MTL	0.80	0.50	0.48	0.75

Table 3.3: Per-label performance on the test set.

		Predicted	
		-	+
True	-	237	57
	+	184	42

(a) GPT-3

		Predicted	
		-	+
True	-	265	29
	+	145	81

(b) MTL

Table 3.4: Confusion matrices for the novelty labels.

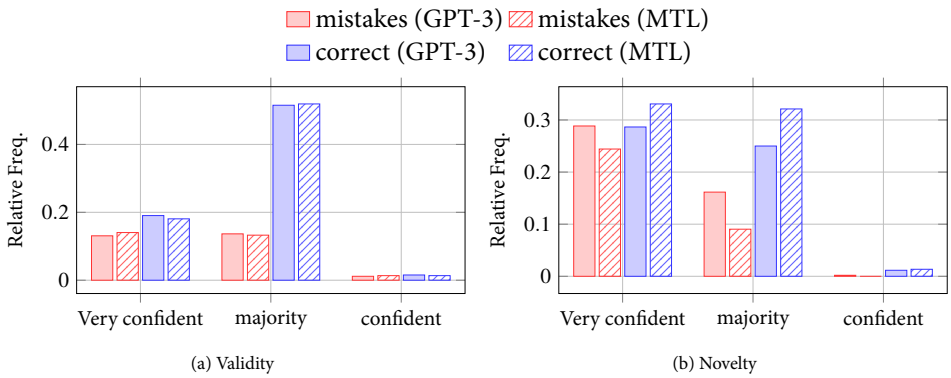
3.5.1 Error Analysis

We perform additional error analysis on three approaches (Approach 1, 2, and 3). We analyze errors in terms of (1) label-specific performance, (2) annotator confidence, and (3) topics. Additional results are in Appendix B.2.

Per-label performance We observe complementary strengths for the GPT-3 model and our MTL approach in Tables 3.3. The MTL model is remarkably stronger than GPT-3 at identifying *novel* arguments, even when considering this is a low-frequency class. We see a similar trend in terms of misclassifications (Table 3.4), as the MTL model has a 40% lower error rate for the novelty label.

Annotator confidence See Figure 3.2 for the relationship between annotator confidence and classification error. Surprisingly, examples labeled as very confident (easy for human annotators) are not consistently correctly classified by any approach. For novelty, GPT-3 gets about half of these examples wrong.

Topics The 3 topics with the highest error rates differ between approaches and tasks. For validity, GPT-3 struggles with "Was the Iraq War Worth it?" (44.8%), while MTL with "Vegetarianism" (40%). For novelty, GPT-3 also struggles with "Vegetarianism" (60%), and MTL with "Withdrawing from Iraq" (44.7%) and "Vegetarianism" (44%).



(a) Validity

(b) Novelty

Figure 3.2: Relative accuracy rates divided over label confidence scores.

3.6 Conclusion

We highlight two main conclusions. First, different models have different strengths relating to the two tasks. A prompting approach with a generative model worked best for validity, while contrastive supervised learning worked best for novelty. The two tasks are related enough to be able to effectively use one multi-task learning model, but merging predictions from multiple heterogeneous models leads to the best score. Second, specific intermediate tasks before fine-tuning work well for low-resource argument mining tasks. NLI seems clearly related to validity prediction. For the novelty tasks, other tasks related to argument similarity [315] might be equally informative.

3.7 Access and Responsible Research

A core consideration in NLP research when sharing results is the accessibility and reproducibility of the solution. While our code is openly available, the approaches including GPT-3 require access to commercially trained models. We used free trial OpenAI accounts (allowing $18 of free GPT-3 credit), but larger datasets and additional tasks can quickly make this approach infeasible. We also considered the freely accessible BLOOM model.[5] BLOOM does not require payment but does require more GPU memory than what was available to us – making it inaccessible. Ultimately, GPT-3 and related LLMs have several biases and risks of use, including the generation of false information [379] and the fact that their training on internet language leads to a very limited set of language, ideas, and perspectives represented [46], with even racist, sexist, and hateful views [137]. This is especially important to mention, as the task description mentions a future use case of generating new arguments.

[5] https://huggingface.co/bigscience/bloom