Argument Novelty and Validity Assessment via Multitask and Transfer Learning

Milad Alshomary
Paderborn University
milad.alshomary@upb.de

Maja Stahl
Paderborn University
maja.stahl@upb.de

Abstract

An argument is a constellation of premises reasoning towards a certain conclusion. The automatic generation of conclusions is becoming a very prominent task, raising the need for automatic measures to assess the quality of these generated conclusions. The SharedTask at the 9th Workshop on Argument Mining proposes a new task to assess the novelty and validity of a conclusion given a set of premises. In this paper, we present a multitask learning approach that transfers the knowledge learned from the natural language inference task to the tasks at hand. Evaluation results indicate the importance of both knowledge transfer and joint learning, placing our approach in the fifth place with strong results compared to baselines.

1 Introduction

Conclusions are essential to understanding the reasoning behind their arguments. In daily life argumentation, argument conclusions are often left implicit (Alshomary et al., 2020) because they are easy to infer or for rhetorical reasons. While it is easy for humans to infer these conclusions, machines struggle with such a task. This phenomenon motivated a line of computational argumentation research to study the task of automatic generation of conclusions (Alshomary et al., 2020; Syed et al., 2021). Evaluating these approaches using traditional text generation measures like BLEU or ROUGE is not enough since multiple conclusions can be considered valid for a given argument. Additionally, one might desire specific criteria in a generated conclusion, like being informative (Syed et al., 2021).

In this regard, the SharedTask at the 9th Workshop on Argument Mining proposed two quality dimensions of argument conclusions to be assessed. The first is *validity*, defined as whether a given conclusion can be logically inferred from its premises. The second is *novelty*, assessing whether the conclusion goes beyond what is mentioned in the premises to provide novel insights. This paper describes our approach to the automatic assessment task of conclusion's validity and novelty.

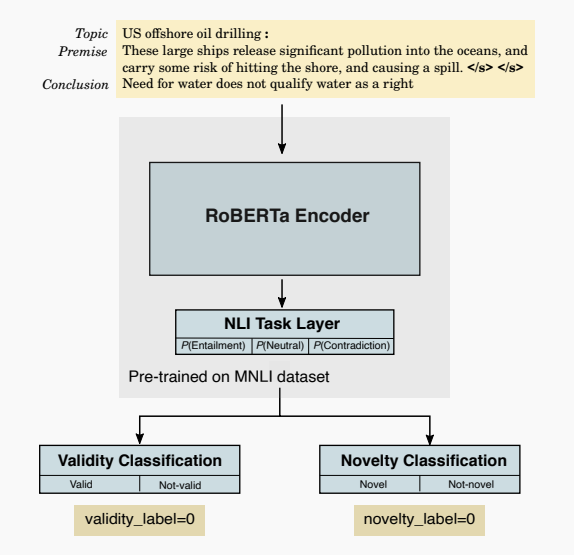

Figure 1: Our proposed model, which jointly models the validity and novelty assessment tasks, starts from a transformer-based model pre-trained on the natural language inference (NLI) task. First, we pass the input through the RoBERTa encoder. Then, the last hidden state of the encoder is projected into a probability distribution representing the NLI labels. Each classification head (novelty and validity) then learns to map this distribution into the corresponding labels.

We address the novelty and validity assessment tasks via a multitask learning approach, employing already acquired knowledge from the natural language inference task. In particular, the two assessed quality dimensions are orthogonal. That is, conclusions that can be easily inferred from their premises and hence valid are less likely to be novel. Similarly, the more novel a conclusion is, the harder to judge its validity. Accordingly, we believe that jointly modeling the two assessment tasks allows the model to exploit such dynamics. Moreover, the natural language inference task (NLI) is very simi-

Proceedings of the 9th Workshop on Argument Mining, co-located with the 29th International Conference on Computational Linguistics, pages 111–114
October 17, 2022.

lar and is widely studied (Wang et al., 2018). One can understand the entailment attribute between two sentences in the NLI task as a validity criterion. Hence, we start from a transformer-based model fine-tuned on the NLI task. As shown in Figure 1, this model consists of a transformer-based encoder and a classification head that predicts one of three labels, *entailment*, *contradiction*, *neutral*. We stack two classification heads on top of the model, one to predict novelty and the other for validity.

We evaluate our approach against the basic RoBERTa model trained on each task independently in our experiments. Results show the gain achieved from both the knowledge acquired from the NLI task as well as the joint learning of the two tasks. First, utilizing knowledge from the NLI task boosts the average F1-score from 0.09 to 0.15. Secondly, the joint learning of the two tasks further raises the average F1-score up to 0.42, placing our approach in the fifth place with strong competitive performance.[1]

2 Related Work

Conclusion inference is the task of generating a natural language conclusion given a set of premises. The generation of these conclusions is important for AI algorithms to understand the reasoning behind arguments. Hence, several works in computational argumentation addressed this task. Alshomary et al. (2020) reconstructed implicit conclusion targets from premises using triplet neural networks. Syed et al. (2021) studied the effectiveness of several transformer-based models on the conclusion generation across various corpora and evaluated the informativeness criteria of conclusions. Gurcke et al. (2021) automatically generated conclusions to then use them for argument quality assessment. Liu et al. (2021) worked on generating perspectives (conclusion) for news articles. (Becker et al., 2021) fine-tuned language models to generate implicit knowledge in sentences. In this work, the proposed task and approaches aim to study the quality of automatically generated conclusions along the validity and novelty dimensions.

Recent advances in natural language processing (NLP) have been driven by transfer learning, where knowledge on one task is used to learn another potentially relevant task. Indeed, it has been shown that language models trained on big corpora can excel in transferring such knowledge into downstream tasks in a zero-shot setting (Radford et al., 2019). Our proposed method uses knowledge learned the natural language inference (NLI) task (Liu et al., 2019) to solve the novelty and validity assessment tasks. Another promising learning paradigm is multitask learning (Zhang et al., 2022), in which two or more relevant tasks are learned in the same neural model, either in a soft or hard parameter sharing setting. Our approach models the validity and novelty tasks jointly in one model with hard parameter sharing.

3 Task and Data

In the SharedTask at the 9th Workshop on Argument Mining, the organizers defined the validity and novelty criteria of argument conclusions as follows:[2]

- *Validity*: The conclusion can be logically inferred from the premise.
- *Novelty*: The conclusion provides novel premise-related content and/or combines the content of the premises in a way that goes beyond what is stated in the premises.

According to these definitions, the organizers proposed two settings for this SharedTask. The first is, given a set of premises in natural language text and a corresponding conclusion, predict two scores that reflect the conclusion's novelty and validity (Subtask A). In the second Subtask, two conclusions are provided, and the task is to rank them according to their novelty and validity (Subtask B). This paper tackles Subtask A, which is the binary classification of validity and novelty dimensions.

The dataset provided by the organizers consists of premises and conclusions, which they manually annotated for validity and novelty dimensions. Additionally, the organizers include the topic of the debate and the confidence scores for the two labels. The data also contained borderline cases for both target dimensions, which are considered to be *somewhat* novel or valid. We excluded those examples from the training and validation sets, as suggested by the organizers. We ended up with 721 training and 199 development examples for validity and 718 training and 200 development examples

[1] Our model and experiments are publicly available under https://github.com/MiladAlshomary/ArgsValidNovelTask

[2] https://phhei.github.io/ArgsValidNovel/, last accessed: 2022-08-25.

Split	Novel?			Valid?		
	Yes	No	All	Yes	No	All
Train	123	595	718	401	320	721
Validation	82	118	200	125	74	199
Test	226	294	520	314	206	520

Table 1: Class distribution for both Novelty and Validity classes in each of the data split.

Model	Validity	Novelty	ValNov
RoBERTa	0.28	0.36	0.09
NLI-based RoBERTa	0.52	0.35	0.15
NLI-based Multitask	**0.71**	**0.60**	**0.42**

Table 2: Macro F1-scores for the validity and novelty tasks, as well as the combined one (ValNov) computed for our approach (NLI-based Multitask) and its baselines on the test set.

for novelty. The test set has a size of 520 instances. Table 1 shows the distribution of each label for all the data splits. We notice that the data is imbalanced in terms of novelty class. In our experiments, we report our approach's effectiveness also when trained on a training split that is balanced through oversampling.

4 Approach

As mentioned, our approach to the Subtask A is to jointly learn the two assessment tasks (*novelty* and *validity*), starting from knowledge acquired by a model trained on the NLI task (Wang et al., 2018). The motivation for our choice is two folds. On the one hand, we argue that the novelty and validity dimensions correlate such that conclusions that are easily inferred to be valid are likely not that novel. Similarly, the more novel a conclusion is, the harder to judge its validity. On the other hand, we see similarities to the natural language inference task. If an NLI model deemed the conclusion to be entailed from its premises, then the conclusion is likely valid but probably not novel.

In particular, we start from a transformer-based model fine-tuned on the NLI task. As shown in Figure 1, this model consists of a transformer-based encoder and a classification head that predicts one of three labels, *entailment*, *contradiction*, *neutral*. The input to our model is a concatenation of the topic, premise, and conclusion. We pass the input through the RoBERTa encoder to obtain the final hidden state, which is passed through the classification layer to obtain a probability distribution over the three NLI labels. We stack two classification heads on top of the model to project this distribution into the corresponding novelty and validity labels. During training, one can compute an average error with respect to the two tasks at each optimization step or consider the error subject to one task at a time. For simplicity, we chose the second option since the framework we build upon supports only this option (details in Sections 5). Although the weights are not updated according to an average loss of the two tasks, the overall training will drive the weights into an area optimal for both tasks.

5 Experiments

In our experiments, we use the RoBERTa model (Liu et al., 2019) fine-tuned on the Multi-Genre Natural Language Inference dataset (MNLI) made public by Williams et al. (2018). For each task, we train a model considering it the main task and the other as an auxiliary one with a loss discounted by a factor of α. We explored a range of α values and chose the ones that lead to the best F1-score on the validation set, that is, 0.9 and 0.7 for novelty and validity, respectively. Additionally, we explored different learning rates for each of the models independently and chose a learning rate of $2e^{-5}$ and $5e^{-6}$ for the novelty and validity models, respectively. We train both models for ten epochs with a batch size of 8. We compare our approach against the RoBERTa model without NLI fine-tuning and once with NLI fine-tuning. Both trained independently on each task. As mentioned in Section 3, the training data is imbalanced along the novelty label. To address this problem, we perform oversampling in which we randomly replicate instances of the class novel to reach a balanced situation. We then train our model and the baselines on it. Our model is built on top of the multitask learning framework made publicly available under `https://multi-task-nlp.readthedocs.io/en/latest/`.

Table 2 shows the F1-score achieved by our approach (NLI-based Multitask) and its baselines computed for the novelty and validity tasks, and the combined task on the test set [3]. We can see that using knowledge from the NLI task boosts the effectiveness of RoBERTa on the validity task

[3]Reported results were computed after the SharedTask deadline when the test set was made publicly available. Wrong prediction file was originally submitted before the deadline, however the approach and training procedure are the same.

Model	Validity	Novelty	ValNov
RoBERTa	0.28	0.52	0.13
NLI-based RoBERTa	0.52	**0.64**	**0.33**
NLI-based Multitask	**0.71**	0.46	0.32

Table 3: Macro F1-scores for the validity and novelty tasks, as well as the combined one (ValNov) computed for our approach (NLI-based Multitask) and its baselines on the test set when trained on the over sampled training split.

from 0.28 to 0.52. Moreover, modeling novelty and validity tasks jointly boost the performance to reach 0.71 and 0.60 F1-score on the validity and novelty tasks, respectively. The combined F1-score recognizes instances as correctly predicted only if validity and novelty are both correctly predicted. Among evaluated baselines in this paper, our model achieves the best combined F1-score of 0.42.

From Table 3, we can see that when training the models on the oversampled training set, we observe a boost in performance for novelty for both the normal and NLI-based RoBERTa models. On the contrary, the effectiveness of our multitask learning approach got worse when we performed the oversampling. However, the overall performance of our approach still improves over the baselines when oversampling. Overall, the best performing model in terms of the combined F1-score is achieved by our model trained on the original data.

6 Conclusion

In this paper we described our approach proposed for the SharedTask at the 9th Workshop on Argument Mining for assessing the validity and novelty of argument conclusions. Our approach jointly models the two binary tasks of novelty and validity making use of knowledge acquired from the natural language inference task. Experimental results, shows the gain achieved from both transferring knowledge from the NLI, as well as the joint modeling of the two tasks.

References

Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. 2020. Target inference in argument conclusion generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4345, Online. Association for Computational Linguistics.

Maria Becker, Siting Liang, and Anette Frank. 2021. Reconstructing implicit knowledge with language models. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 11–24.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siyi Liu, Sihao Chen, Xander Uyttendaele, and Dan Roth. 2021. Multioped: A corpus of multi-perspective news editorials. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4345–4361.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.

Shahbaz Syed, Khalid Al Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. 2021. Generating informative conclusions for argumentative texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3482–3493.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2022. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. *arXiv preprint arXiv:2204.03508*.