Overview of PragTag-2023: Low-Resource Multi-Domain Pragmatic Tagging of Peer Reviews

Nils Dycke, Ilia Kuznetsov, Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt
ukp.informatik.tu-darmstadt.de

Abstract

Peer review is the key quality control mechanism in science. The core component of peer review are the review reports – argumentative texts where the reviewers evaluate the work and make suggestions to the authors. Reviewing is a demanding expert task prone to bias. An active line of research in NLP aims to support peer review via automatic analysis of review reports. This research meets two key challenges. First, NLP to date has focused on peer reviews from machine learning conferences. Yet, NLP models are prone to domain shift and might underperform when applied to reviews from a new research community. Second, while some venues make their reviewing processes public, peer reviewing data is generally hard to obtain and expensive to label. Approaches to low-data NLP processing for peer review remain under-investigated. Enabled by the recent release of open multi-domain corpora of peer reviews, the PragTag-2023 Shared Task explored the ways to increase domain robustness and address data scarcity in pragmatic tagging – a sentence tagging task where review statements are classified by their argumentative function. This paper describes the shared task, outlines the participating systems, and summarizes the results.

1 Introduction

Scholarly communication lies at the heart of scientific discovery (Johnson et al., 2018) and is argumentative by nature. Scientific publications present results, interpret them, justify the experimental setup, and substantiate the claim for new knowledge (Teufel et al., 2009). Peer review reports, in turn, assess the validity, novelty and impact of the underlying publication and argue for or against its acceptance. Peer review is a key component of scientific quality assurance. It is a complex process prone to heuristic behavior (Rogers and Augenstein, 2020) and bias (e.g. Stelmakh et al., 2020; Wang and Shah, 2018). A growing area of NLP for peer review analysis aims to investigate those issues by analyzing argumentation in peer review reports (e.g. Kang et al., 2018; Cheng et al., 2020; Hua et al., 2019; Kuznetsov et al., 2022; Dycke et al., 2023). The resulting systems have numerous potential applications, incl. facilitating meta-scientific analysis of reviewing practices, helping authors and program chairs aggregate information from multiple reviews, and supporting junior reviewers in giving thorough, objective and helpful feedback.

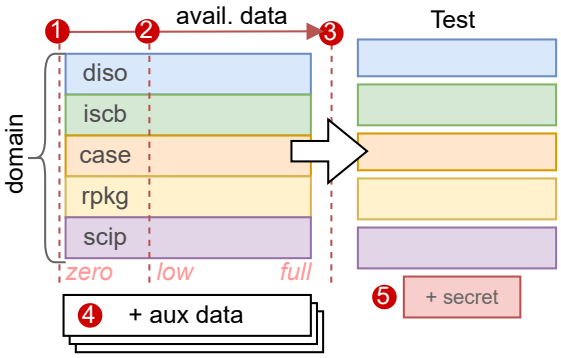

Figure 1: PragTag-2023 Overview. Given a mixed-domain corpus of peer reviews annotated with pragmatic tags, the participants submit systems trained with varying amounts of training data (1-3) with optional use of unlabeled auxiliary data (4). The systems are evaluated in each of the five domains (Section 3.1), as well as on a previously unseen secret domain (5).

Standards and practices of scholarly communication vary across research communities. Yet, to date, NLP for peer review has focused on data from machine learning conferences (Kang et al., 2018; Hua et al., 2019; Cheng et al., 2020; Kennard et al., 2022), and the applications outside of this domain remain under-investigated. This over-focus on one domain can be attributed to data scarcity – while some communities make their reviewing public, peer reviews are generally hard to obtain and legally clear for research use (Dycke et al.,

Proceedings of the 10th Workshop on Argument Mining, pages 187–196

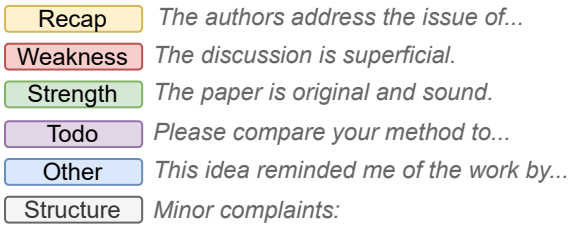

Figure 2: Pragmatic tags. `Recap` neutrally summarizes the paper; `Weakness` and `Strength` outline the negative and positive aspects of the work; `Todo` covers explicit requests to the paper authors; `Other` marks non-argumentative statements; `Structure` denotes structural elements of the review text.

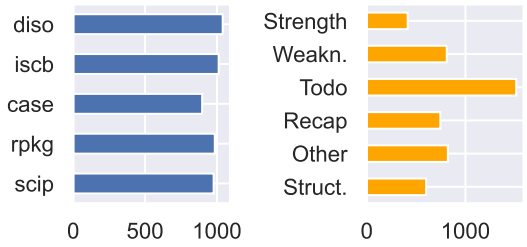

Figure 3: Number of sentences by domain (left) and label (right) in the F1000RD core data (train and test).

2022). In addition, due to the technical nature of peer review texts, they are expensive to annotate. Measuring the effects and mitigating the impact of domain shift and data scarcity are important and under-researched questions in NLP for peer reviews.

The introduction of open multi-domain corpora of peer reviews (Dycke et al., 2023) and domain-neutral review analysis tasks (Kuznetsov et al., 2022) makes it possible to investigate these questions empirically. The PragTag-2023 Shared Task[1] collaboratively explored multi-domain NLP for peer reviews under data scarcity. As an exemplary task we took pragmatic tagging – a sentence-level argumentation labeling task that classifies peer review statements by their communicative purpose (Section 2). PragTag-2023 has received five diverse submissions that provide new insights into multi-domain low-data pragmatic tagging, and propose a wide spectrum of methods to increase model robustness under four increasingly challenging conditions. This paper describes the shared task setup, summarizes the submissions, and aggregates the main insights from the competition. To support further investigation of multi-domain low-data NLP for peer review, we archive the code and data of the shared task and make them publicly available[2].

2 Pragmatic tagging

Task. Pragmatic tagging is a sentence classification problem where given the sequence of sentences $s_1^r, ..., s_n^r$ from a review report r, a model should predict the pragmatic label for each sentence $l_1^r, ..., l_n^r$ from the label set L. We adopt the label set proposed by Kuznetsov et al. (2022), illustrated in Figure 2. The label set was evaluated in an annotation study and shown to be well-applicable across different research fields and communities while yielding good inter-annotator agreement of approx. 0.7 Krippendorff's α. The core sources of disagreement are the coarse granularity of the schema (necessary for generalization), sentence-level analysis (necessary to avoid discrepancies due to differences in sub-sentence splitting), and the natural ambiguity of the classes (e.g. `Weakness` vs `Todo`).

Evaluation. Kuznetsov et al. (2022) provide the data, but do not specify metrics for evaluating NLP systems for pragmatic tagging. In PragTag-2023, we evaluate system performance via the F1 score. Since the label distribution is skewed, we opt for the macro-averaged F1 within each domain for evaluation. We then compute scores for each domain individually and use the mean across all domains as the final leaderboard score (Section 4).

Baselines. To contextualize the submission scores, we implemented two baselines. The supervised **RoBERTa** baseline is a `roberta-base` model (Liu et al., 2019) fine-tuned for 20 epochs on the training data available for a given experimental condition (Section 4.2). The **majority** baseline directly assigns the most frequent pragmatic tag from the training data to the input sentence.

3 Data

The participants of the shared task were given two types of data (Figure 4). The smaller-scale *core data* contains peer review texts labeled with pragmatic tags on the sentence level. Core data is used for training and evaluating the systems. The large-scale *auxiliary data* consists of two unlabeled text collections. It can be used to enhance the systems' robustness to low-data conditions in the

[1] https://codalab.lisn.upsaclay.fr/competitions/13334

[2] https://github.com/UKPLab/pragtag2023

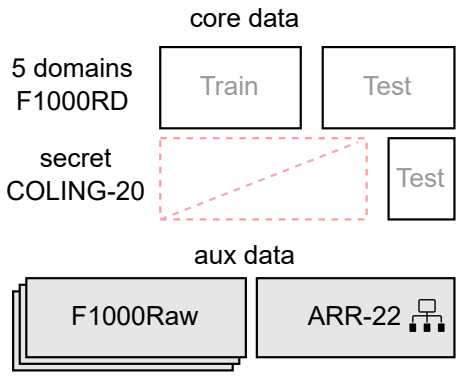

Figure 4: PragTag-2023 data overview. In addition to labeled core data from F1000RD and COLING-20, the parcitipants are provided with two unlabeled collections: a large multi-domain corpus of unstructured peer reviews (F1000Raw), and smaller collection of *semi-structured* peer reviews in the NLP domain (ARR-22).

multi-domain setting.

3.1 Core Data

The core data originates from the F1000RD corpus (Kuznetsov et al., 2022), and contains review reports with manually annotated pragmatic tag labels for each sentence. Each review report belongs to one of the five domains:

- Disease outbreaks (`diso`)
- Computational biology (`iscb`)
- Medical case studies (`case`)
- R Packages (`rpkg`)
- Scientific policy research (`scip`).

The core data from F1000RD covers 4911 sentences from 224 peer review reports. Figure 3 shows the label and domain distribution in the F1000RD data. The instances are unequally distributed both across domains (slightly) and across pragmatic tags (substantially). The skewed pragmatic tag distribution reflects a natural distribution in peer review texts, with most sentences dedicated to critically assessing the work and suggesting improvements. The differences in the number of instances across domains stem from the per-review data sampling procedure in the F1000RD corpus and the review length variation across domains. To account for the uneven distribution, PragTag-2023 employed macro-averaging by label and by domain during evaluation (Section 2).

We split the core data into training set (2326 sentences) and test set (2585 sentences), at random, on review basis, per domain. We did not provide a fixed development set – instead, the participants were free to derive it from the training set by themselves. We note that the training data is a *mixed* collection with instances from all domains; per-instance domain identifier is provided. The test data, on the other hand, is split *by domain*, and evaluation is performed on *each* of the domains separately. The rather uncommon $50/50$ training-test split is thus necessary to ensure sufficient amount of test data in each domain.

In addition to the five F1000RD domains listed above, the final phase of the competition evaluated the systems on a previously unpublished *secret* test set. This collection includes 255 sentences from 10 peer reviews in computational linguistics taken from the COLING-20 portion of the NLPeer corpus and annotated with pragmatic tags following the F1000RD tagset. Labeling was performed by two annotators proficient in the NLP domain, reaching an agreement of 0.65 Krippendorff's α – slightly lower than in the original study. The labels were adjudicated by an expert annotator closely familiar with the F1000RD labeling schema. The domain and composition of this new data were unknown to the participants until the start of the final evaluation.

3.2 Auxiliary data

Using unlabeled or partially-labeled auxiliary data is a common way to mitigate domain shift and to address the lack of labeled data. To enable application of such techniques, the shared task provided the participants with two additional auxiliary datasets.

F1000Raw is a large multi-domain collection of papers and peer reviews from a wide range of domains. The data originates from the F1000Research platform – same source as the non-secret core data. F1000Raw corresponds to the F1000-22 subsection of the NLPeer corpus (Dycke et al., 2023), excluding the instances that appear in the core shared task data, and covers approx. 10k reviews for 4.8k papers, 3.8M review words in total (Dycke et al., 2023). Like the core data, F1000Raw contains full-text, unstructured peer reviews. Unlike the core data, F1000Raw does not contain explicit domain identifiers or pragmatic tag labels.

ARR-22 is a corpus of papers and peer reviews in the NLP domain from the data collection campaigns at ACL Rolling Review (Dycke et al., 2022). It covers 684 reviews for 476 papers, approx. 266k review words in total (Dycke et al., 2023). The reviews are semi-structured, and each review is

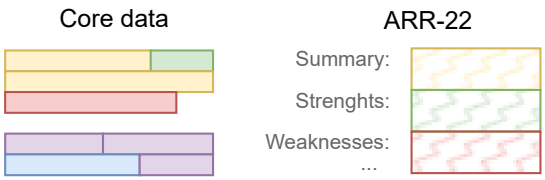

Figure 5: Difference between sentence-level unstructured core data and section-level semi-structured ARR-22 data from peer reviews that use review forms. Colors correspond to different pragmatic tags (see Figure 2).

split into free-text fields: "*Summary*", "*Strengths*", "*Weaknesses*", "*Suggestions*" and "*Ethical concerns*". The similarity between the review form fields and the pragmatic tags is not coincidental: both reflect review pragmatics, implicitly (pragmatic tags) or explicitly (form fields). Yet, unlike the core data, ARR-22 does not contain sentence-level pragmatic tags, and not every sentence in a review section corresponds to its overall pragmatics (Figure 5). Finding a solution to bridge this gap is left to the participants.

We envisioned F1000Raw as a valuable source of data for increasing cross-domain robustness of the participating systems. We envisioned ARR-22 as a potential distant supervision source for low-data scenarios explored in PragTag-2023.

4 Setup

4.1 Implementation

The shared task was run via CodaLab (Pavao et al., 2023). The competition website provided necessary information about the task, the core and auxiliary data, as well as a starting kit including an evaluation script and a baseline implementation. The participants would apply their system to the test set inputs and submit the predictions via CodaLab, where they would be compared to the gold outputs. The score would be stored in the participants' dashboard and could be submitted to the publicly available leaderboard.

4.2 Conditions and Rules

The participants submitted systems to one of the following conditions, simulating different training data availability scenarios:

- **No-data**: The system observed no instances of the core data neither at training time nor at inference time.
- **Low-data**: The system is trained on 20% of the core training data (33 reviews, 739 sentences). The exact 20% split is provided by the shared task organizers and is identical among all participants.
- **Full-data**: the system has access to 100% of the core training data (117 reviews, 2326 sentences).

The test data was identical across these three conditions. At the end of the competition, the participants could submit *any* of their systems to a special **Final** condition, which included the core test data as well as the secret test data, as detailed in Section 3.1.

To promote reproducibility of the results and fair competition, we imposed a few restrictive **rules** on the submissions. The teams were allowed to use PragTag-2023 auxiliary data without restrictions. However, pre-training or fine-tuning the submissions on *any* other data was not allowed. We imposed no requirements upon the system architecture. However, in case of large language models, the participants were requested to only use non-commercial models with publicly available weights, e.g. Llama (Touvron et al., 2023). Submissions built on top of commercial models like ChatGPT and GPT-4 (OpenAI, 2023), etc. were not considered for the evaluation. To prevent optimization on the hidden test data, each team was allowed up to five submissions to each of the conditions. A special **Sandbox** condition with no submission limit was provided for troubleshooting purposes.

5 Submissions

Out of over 20 teams that signed up for the competition, five teams have made it to the final submission. The submitted systems explore a wide range of techniques and architectures for multi-domain pragmatic tagging in low-resource scenarios. We summarize the main ideas behind each submission below and refer to the system papers for details.

CATALPA_EduNLP (Ding et al., 2023) investigated a wide array of approaches. For the full- and low-data setting, this includes supervised sentence labeling via RoBERTa (Liu et al., 2019) augmented with additional features (domain, position, context, word normalization), as well as IOB-style sequence tagging using long-document Transformers and nearest-neighbor-based labeling using SBERT (Reimers and Gurevych, 2019). In the zero-shot setting, the team experimented with labeling test instances based on their similarity to class defini-

	mean	case	diso	iscb	rpkg	scip	secret
DeepBlueAI	**84.1**	82.9	84.1	82.8	**86.0**	**89.0**	**80.1**
NUS-IDS	83.2	83.8	**85.4**	**83.3**	84.8	87.8	74.1
MILAB	82.4	**84.0**	83.7	80.1	85.4	86.5	74.9
SuryaKiran	82.3	82.0	82.8	81.8	82.8	86.5	77.9
CATALPA	81.3	80.8	82.0	81.1	82.5	82.5	78.8
Ensemble	84.4	84.0	85.2	83.3	87.3	88.7	78.0
RoBERTa	80.3	80.3	80.8	79.9	83.1	83.8	73.7
Majority	8.0	9.3	7.3	7.5	8.6	7.9	7.3

Table 1: Final evaluation leaderboard, mean F1-macro score across domains and scores per domain, converted to percentage points for readability. Top: submissions, Middle: majority-vote ensemble of system predictions (Section 7), Bottom: baselines. Bold: best score per column (w/o ensembling).

tions from the shared task description, as well as with prompting via GPT3.5. The participants used the ARR-22 auxiliary data, addressing the gap in label distribution between ARR and the core data via subsampling, and explored data augmentation based on F1000raw auxiliary data. By ensembling best per-domain configurations selected on the validation set, they found that a BERT-based model with additional features outperforms sequence tagging and nearest-neighbor labeling on the full data, while a BERT-based model augmented with additional data performs best in the low-data setting. Prompting GPT3.5 in the zero-shot setting was shown vastly superior to SBERT-based classification based on task definitions – yet, following the PragTag rules, GPT3.5 result was not used for the leaderboard submission.

DeepBlueAI (Luo et al., 2023) focused their approach on increasing the robustness of pre-trained models in the sentence labeling setting. The experiments were conducted using three models – RoBERTa, DeBERTa (He et al., 2023) and XLM-RoBERTa (Conneau et al., 2020). The participants augmented the model via max pooling and attention pooling, introduced adversarial training via fast gradient method, and reported comparative performance of the models trained under different settings via cross-fold validation, showing that the modifications lead to variable performance gains. The authors report that the DeBERTa model consistently outperforms the other two models on the task. To tackle the secret test set in the final phase of the competition, the authors used a voting approach combining a range of models trained in different configurations and selecting the label with the maximum vote, stressing the benefits of fusing different types of models for prediction.

NUS-IDS (Gollapalli et al., 2023) explored multiple approaches to the task for each experimental condition. In the zero-shot no-data condition, the participants proposed two methods: a question-answering model that selects passages from the peer review based on a set of questions derived from peer reviewing guidelines of NLP conferences, and a prompting-based approach based on the Flan-T5 (Chung et al., 2022) model. For the low- and full-data setting, the participants experimented with fine-tuning pre-trained language models, additionally exploring ensembling and data augmentation techniques by tentatively labeling the auxiliary shared task data. The results indicate that prompting via Flan-T5 outperforms question-answering based approach in the no-data setting; in low- and full-data data, fine-tuning a T5 model (Raffel et al., 2019) on tentatively labeled auxiliary data followed by fine-tuning on the core task data performs best.

MILAB (Lee et al., 2023) approached the problem of data scarcity and domain shift via data augmentation. In particular, to compensate for the lack of data, the team applied an ensemble of RoBERTa-based classifiers to label auxiliary data from F1000raw and ARR-22. Apart from majority labeling, the authors explored a novel recall labeling technique: the models assign tentative labels to the unlabeled instances in the decreasing order of recall on a validation set, while labeling the residual instances as Other. Additionally, the authors experimented with diversifying the data by applying off-the-shelf synonym generation followed by BERTScore filtering (Zhang et al., 2020). The results indicate that the proposed data augmentation

techniques combined with ensembling improve the model performance on the task, especially in the no-data condition.

SuryaKiran (Suri et al., 2023) explored the use of unsupervised pre-training on F1000raw auxiliary data to increase domain robustness of the pragmatic tag classifier. In particular, the participants pre-trained the DeBERTa model on F1000raw using masked language modeling objective (Devlin et al., 2019), and later used an ensemble of five models further fine-tuned on different training data splits to make the test set prediction. Their results demonstrate that pre-training via masked language modeling leads to improved performance only in some cases; the authors attribute this to the vocabulary discrepancies between the domains. The team submitted their system only to the final evaluation.

6 Main results

The final leaderboard of PragTag-2023 is shown in Table 1. The participants were invited to submit their best system trained under *any* condition to the leaderboard – expectedly, the best-performing systems trained on full data were submitted. As we can see, on average, all systems outperform the RoBERTa baseline fine-tuned on full training data, and the majority baseline scores poorly due to the macro-averaging of F1 across labels. The submission by DeepBlueAI achieves the highest F1-score both on average, and on the secret test domain. However, this superior performance is not absolute, and on per-domain basis we observe variation in the system rankings: the CATALPA system performs second-best on the `secret` test set, NUS-IDS achieves best performance in the `diso` and `iscb` domains, and the best score in the `case` domain is taken by MILAB. We note consistent and substantial performance degradation on the secret domain across all submissions and baselines. We attribute this to domain shift: while the systems could observe *some* data from *each* of the other domains during training, the secret data is truly out-of-distribution, originating from an entirely different research community and reviewing platform. This gap in performance highlights the importance of cross-domain study of NLP for peer reviews.

Turning to the data scarcity, Table 2 summarizes mean submission scores for various data conditions, from no-data zero-shot learning to full-data fine-tuning. Here, too, all submissions have outperformed the RoBERTa baselines, albeit by a smaller margin in the low-data setting. The no-data and low-data results show great variation both in terms of absolute scores and in terms of leaderboard rankings. Especially in the no-data setting, the highest- and lowest-scoring submission differ by almost 30 percent F1-measure, compared to the 3 percent gap on full data. The submission by MILAB scores best in the no-data scenario, while the system by NUS-IDF performed best on low data. Secret test set not taken into account, DeepBlueAI and NUS-IDS share the first place in the full-data condition. These observations demonstrate the value of evaluating NLP systems for pragmatic tagging in varying data availability conditions.

	no-data	low-data	full-data
MILAB	**51.6**	77.1	83.9
NUS-IDS	40.2	**81.3**	**85.0**
CATALPA	22.2	74.5	81.8
DeepBlueAI	-	80.8	**85.0**
RoBERTa	-	74.4	80.3

Table 2: Mean F1-macro score across domains for different data scarcity conditions, without secret domain.

7 Analysis

Access to all the participating system's predictions at once allows additional insights into the task. Given the broad range of approaches proposed by the PragTag-2023 participants, a natural question arises if these approaches are complementary. We investigate this by combining the predictions of the best-performing submissions via majority vote. The results show that a majority ensemble indeed outperforms every individual system on average (Table 1, middle). Considering per-domain results reveals more nuance: the ensemble maintains the best systems' performance for the domains `case` and `iscb`, slightly lagging behind on `diso` and `scip`, substantially improving the best result in `rpkg`, and showing average performance on the `secret` test set. This variation demonstrates the importance of fine-graned evaluation of pragmatic tagging in multi-domain setting, and we deem the use of alternative, e.g. weighted, ensembling methods for the task promising.

Analysis of the confusion matrix between the true labels and the majority ensemble predictions allows us to see which labels are particularly hard for the systems to handle. Figure 6 presents the

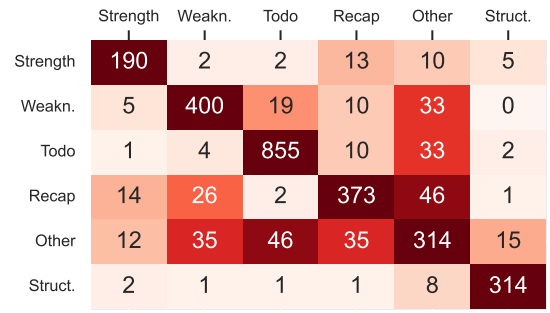

Figure 6: Confusion matrix of PragTag-2023 submission majority ensemble on the final test data: true label (rows) vs predicted label (columns).

results. We observe that, in aggregate, the systems are successfully able to distinguish between *Strengths*, *Weaknesses*, *Todo* and *Structure*, while the *Recap* and especially the open *Other* class constitute frequent sources of confusion, in line with the annotation study observations by Kuznetsov et al. (2022). This result suggests that future labeling schemata for pragmatic tagging might consider refining the *Recap* and *Other* class definitions, or, alternatively, merging these classes into a general *Other* class, eliminating the hard distinction and resulting in more robust systems, at the loss of granularity. We leave this exploration to the future.

8 Discussion

A high-level picture of the submissions to the PragTag competition reveals several trends. Despite the advances in LLM development, fine-tuning of BERT-family LMs was still used by most participants, although some have experimented with prompting. While our rules prohibited the use of commercial LLMs, new open LLMs like Llama (Touvron et al., 2023) have been released. Investigating the performance of these models for our task is a promising avenue for future studies.

While some submissions focused on modifying the model architecture and pre-training regime, others explored data augmentation and creative adaptations of the task, e.g. by casting it as a question-answering task or labeling the instances based on the similarity to guideline class definitions. Most participants used auxiliary data as an unlabeled substrate for pseudo-labeling or language model pre-training. We note the wide use of model ensembling across the submissions, and believe that such techniques will remain relevant in the age of LLMs. PragTag-2023 was designed to accommodate various approaches to the task: pragmatic tagging can be cast as sentence labeling and as sequence labeling, and can can be approached via prompting. While the participants have experimented with many of these options, in-context learning (Dong et al., 2023) remained under-explored. We deem such exploration promising.

The ongoing adaptation of the field to the last-generation LLMs presents new challenges to the benchmarking and shared task methodology. The technical requirements of pre-training and fine-tuning LLMs put the teams without access to massive data and compute at disadvantage. The opaqueness of the LLM pre-training for commercial models introduces the risk of model exposure to the test data or related datasets. PragTag-2023 attempted to mitigate these issues by explicitly limiting the competition to the models for which open weights are available and pre-training procedure is known, and by prohibiting the use of any additional pre-training sources apart from the core and auxiliary data provided with the task. An alternative solution could be to limit the competition to several open LLM instances, inference-only. This, however, would limit the scope of methods the participants can explore to prompting-based approaches. We leave the search for flexible, fair and reproducible benchmarking methodology in the age of LLMs to future work.

9 Conclusion

This paper has introduced PragTag-2023: the shared task in low-resource multi-domain pragmatic tagging of peer reviews. We have described the rationale behind the task, introduced the data and outlined a range of experimental conditions under which the competition took place. The shared task participants proposed a wide range of techniques for increasing the robustness of pragmatic tagging across domains and data availability scenarios. The results of the competition underline the importance of evaluating pragmatic tagging systems across different domains and in different data availability conditions. The arguably most important gain from an organized competition is not finding the best-performing system for the task, but the accompanying exploration of approaches to solving the problem at hand. To this end, we hope that the ideas and observations from the PragTag-2023 submissions foster future progress in pragmatic tagging, and in cross-domain and low-data processing of peer reviews in general.

Limitations

Few limitations of our setup can be addressed by future work. As common in scholarly NLP, our study is limited to English. Once available, the future multilingual datasets of research papers and peer reviews would enable the study of NLP for peer review across languages *and* domains. A coarse-grained pragmatic tagging schema could eliminate the hard `Recap` vs `Other` distinction (Section 7) and increase the robustness of the evaluation. Obtaining more labeled data per domain would enable the study of data scarcity on *per-domain basis* as well as *across individual training-test domain pairs*, e.g. training on `case` and evaluating on `rpkg`. Alternatively, shifting the focus to zero-shot learning with instruction-following LLMs would allow using all available data for evaluation – yet it would be methodologically limiting (Section 8). Incorporating other peer review analysis tasks into the setup would provide additional insights into the low-data and cross-domain NLP for peer reviews.

Ethics Statement

Increasing the domain robustness and sample efficiency of NLP systems are key steps towards sustainabile and widely applicable NLP. Pragmatic tagging is a basic argumentation analysis task with many potential applications that would increase the transparency, fairness and efficiency of scholarly peer review. We believe that the potential for misuse of this technology is low. The data used in the shared task was obtained according to strict licensing and data management procedures, and is open and freely available for research use.

Acknowledgements

PragTag-2023 is part of the InterText initiative at UKP Lab.[3]. We thank the shared task participants and the organizers of the 10th Workshop on Argument Mining[4] for making this shared task possible. The work was funded by the German Research Foundation (DFG) as part of the PEER project (grant GU 798/28-1), and by the European Union as part of the InterText ERC project (101054961). Views and opinions expressed here are, however, those of the author(s) only, and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

[3] https://intertext.ukp-lab.de

[4] https://argmining-org.github.io/2023/index.html

References

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv:2210.11416*.

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

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.

Yuning Ding, Marie Bexte, and Andrea Horbach. 2023. CATALPA_EduNLP at PragTag-2023. In *Proceedings of the 10th Workshop on Argument Mining*, Singapore. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning. *arXiv:2301.00234*.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Yes-yes-yes: Proactive data collection for acl rolling review and beyond. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 300–318.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational

study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Sujatha Das Gollapalli, Yixin Huang, and See-Kiong Ng. 2023. NUS-IDS at PragTag-2023: Improving pragmatic tagging of peer reviews through unlabeled data. In *Proceedings of the 10th Workshop on Argument Mining*, Singapore. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv:2111.09543*.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. 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 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

Rob Johnson, Anthony Watkinson, and Michael Mabe. 2018. *The STM Report: An overview of scientific and scholarly publishing*. International Association of Scientific, Technical and Medical Publishers, The Hague, Netherlands.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. 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 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Neha Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1234–1249, Seattle, United States. Association for Computational Linguistics.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and resubmit: an intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986.

Yoonsang Lee, Dongryeol Lee, and Kyomin Jung. 2023. MILAB at PragTag-2023: Enhancing cross-domain generalization through data augmentation with reduced uncertainty. In *Proceedings of the 10th Workshop on Argument Mining*, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.

Zhipeng Luo, Jiahui Wang, and Yihao Guo. 2023. DeepBlueAI at PragTag-2023: Ensemble-based text classification approaches under limited data resources. In *Proceedings of the 10th Workshop on Argument Mining*, Singapore. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *arXiv:2303.08774*.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2023. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683v1*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online. ACL.

Ivan Stelmakh, Nihar B Shah, and Aarti Singh. 2020. Catch me if i can: Detecting strategic behaviour in peer assessment. In *ICML Workshop on Incentives in Machine Learning*.

Kunal Suri, Prakhar Mishra, and Albert Nanda. 2023. SuryaKiran at PragTag 2023 - benchmarking domain adaptation using masked language modeling in natural language processing for specialized data. In *Proceedings of the 10th Workshop on Argument Mining*, Singapore. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1493–1502.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal

Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv:2302.13971*.

Jingyan Wang and Nihar B Shah. 2018. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. *arXiv:1806.05085*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *arXiv:1904.09675*.