MILAB at PragTag-2023: Enhancing Cross-Domain Generalization through Data Augmentation with Reduced Uncertainty

Yoonsang Lee[1]*, Dongryeol Lee[2]*, Kyomin Jung[2, 3]†
[1]College of Liberal Studies, Seoul National University
[2]Dept. of Electrical and Computer Engineering, Seoul National University
[3]ASRI, Seoul National University
{lysianthus, drl123, kjung}@snu.ac.kr

Abstract

This paper describes our submission to the PragTag task, which aims to categorize each sentence from peer reviews into one of the six distinct pragmatic tags. The task consists of three conditions: `full`, `low`, and `zero`, each distinguished by the number of training data and further categorized into five distinct domains. The main challenge of this task is the domain shift, which is exacerbated by non-uniform distribution and the limited availability of data across the six pragmatic tags and their respective domains. To address this issue, we predominantly employ two data augmentation techniques designed to mitigate data imbalance and scarcity: *pseudo-labeling and synonym generation*. We experimentally demonstrate the effectiveness of our approaches, achieving the *first* rank under the `zero condition` and the *third* in the `full and low conditions`.[1]

1 Introduction

Peer review is a fundamental procedure for assessing the quality of academic manuscripts (Ware and Mabe, 2015). Most reviews are characterized by concise argumentative feedback, wherein reviewers highlight both strengths and weaknesses while offering suggestions for revision. This observation has led researchers to frame the structures of peer reviews as a subset of argument mining (Lawrence and Reed, 2020; Lauscher et al., 2018; Hua et al., 2019). Parallel to these insights, efforts have been made to automate the peer review process (Yuan et al., 2022; Wang et al., 2020). The automation of this process yields two primary advantages: it facilitates authors by distilling the main feedback from reviews and helps reviewers by aggregating information from multiple reviews.

Recently, Dycke et al. (2023) introduced a novel task, pragmatic tagging for peer review, wherein each sentence of a scientific review is classified into one of six predefined pragmatic categories. The proposed task is tailored for a multi-domain scientific corpus, where certain domains might employ specific terminologies that are not prevalent in others or require a unique evaluative perspective during the review process (Rogers and Augenstein, 2020). Furthermore, the nature of scientific review necessitates profound domain knowledge and careful examination by the reviewer, thereby posing challenges in large-scale data collection. Such challenges, referred to as cross-domain generalization (Caciularu et al., 2021; Du et al., 2020), have been the subject of intensive investigation within natural language processing.

To address these challenges, we propose two approaches to enhancing the generalization of the model over multiple domains: pseudo-labeling and synonym generation. Under `full` and `low conditions`, we finetune BERT (Devlin et al., 2018) based classifiers using the training data and pseudo-label auxiliary data through an ensemble approach to ensure label quality. In the `zero condition`, we exploit the existing sections of the ARR dataset and inject intrinsic characteristics of pragmatic tags without utilizing any large language models. Our method accomplished the *highest* performance in the `zero condition` as well as the *third* in the `full` and `low conditions`.

2 Related Works

Multi-class Classification The task of categorizing input sentences into multiple labels has seen extensive development across various domains (Soleimani and Miller, 2016; Dang et al., 2020). Among the readily available models for text classification, RoBERTa (Liu et al., 2019) stands out, characterized by its incorporation of a classification layer with a transformer encoder. Notably,

* Equal contribution.
† Corresponding authors.
[1]The codes are available at https://github.com/lilys012/pragtag

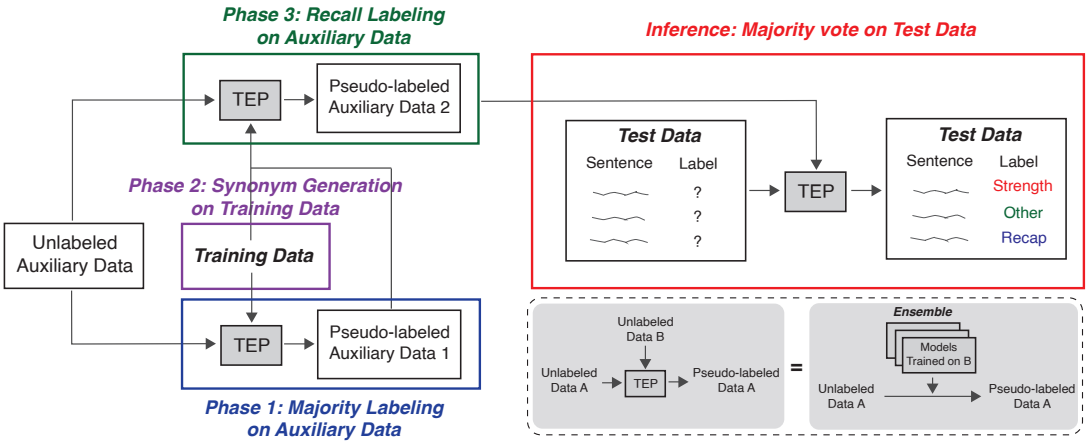

Figure 1: Overview of our proposed approach to pragmatic tagging in the full condition. Phase 1: pseudo-labeler models are trained using provided training data and subsequently utilized to label unlabeled auxiliary data. Phase 2: Training data are augmented by a synonym generator. Phase 3: Augmented data from Phase 1 and 2 are used to finetune the labeler. Models reapply tagging to the auxiliary data with increased certainty. Phase 4: Classifier trained with the labeled data from Phase 3 are ensembled to predict the labels of the test data.

this model is acclaimed for its capability to generalize across diverse domains. However, for datasets tailored to specific domains, models such as SciBERT (Beltagy et al., 2019) and BioBERT (Lee et al., 2020) have been proposed. Additionally, existing research illustrates that the performance of these models can be further enhanced through the employment of ensemble techniques (Saha and Srihari, 2023).

Data augmentation Data augmentation is widely exploited to enrich and generalize datasets (Chen et al., 2023). A sentence can be expanded through the utilization of rule-based techniques and interpolation (Feng et al., 2021). Furthermore, in the case of unlabeled datasets, a trained model can assign pseudo-labels to the unlabeled data, thereby facilitating supplementary training (Lee et al., 2013).

3 Dataset

Task Data The data for the task is sourced from F1000RD (Kuznetsov et al., 2022), which is a comprehensive multi-domain collection of both reviews and their pragmatic labels. Under the `low condition`, only 20% of the full task dataset is employed. Detailed statistics of the six tags across five distinct domains are described in Table 1.

Auxiliary Data The auxiliary data is comprised of two datasets: F1000raw and ARR-22 (Dycke et al., 2022). The former, F1000raw, is an extensive, unlabeled corpus originating from the same source as F1000RD. Conversely, ARR-22 represents a collection of peer reviews from the ACL community. Each review within ARR-22 is segmented into sections designated as Paper Summary, Comments / Suggestions / Typos, Summary of Strengths, and Summary of Weaknesses. It is important to note that the utilization of any external datasets beyond these is strictly prohibited for our task.

	Full						
Domain	**Strg.**	**Weak.**	**Strc.**	**Rec.**	**Td.**	**Oth.**	**Total**
scip	46	73	70	52	115	105	461
iscb	30	93	53	77	173	70	496
rpkg	67	85	64	69	132	89	506
diso	43	81	61	76	135	79	475
case	34	45	53	72	126	58	388
Total	220	377	301	346	681	401	2326

Table 1: Task data statistics based on full conditions and five domains: science policy research (scip), bioinformatics (iscb), R package (rpkg), disease outbreak (diso), and medical case reports (case). Within each domain, the count of sentences is categorized by six labels: Strength (Strg.), Weakness (Weak.), Structure (Strc.), Recap (Rec.), Todo (Td.), and Other (Oth.).

4 Methodology

The efficacy of an individual model can be influenced by various hyperparameters throughout the training process, which could potentially lead to inaccurate predictions. Therefore, we opt for an ensemble approach for our task, as depicted in Figure 1. From the entire training data, we set aside 18 reviews to constitute a validation subset. This subset excludes reviews that belong to the `low condition` dataset. The validation subset is consistently applied across all scenarios for the selection of hyperparameters and models.

	Majority	Consensus
F1000raw	**0.8454**	0.8333
F1000raw+ARR	0.8263	0.8251

Table 2: F1-mean score for auxiliary data labeling. Models are trained using the F1000raw dataset or in conjunction with the ARR dataset. Validation data is labeled by majority and consensus methods.

seed	model	learning rate	score
42	RoBERTa-base	1e-5	**0.7498**
142	RoBERTa-base	2e-5	**0.7667**
242	SciBERT	3e-5	0.7260
342	BioBERT	1e-5	**0.7534**
442	RoBERTa-base	3e-5	0.7306

Table 3: Classifier performance under the low condition. Bold score indicates the selection for majority labeling.

4.1 Pseudo-labeling

To overcome the scarcity of training data, we devise a strategy involving pseudo-labeling (Lee et al., 2013) for the auxiliary data. We train five RoBERTa-base classifiers (Liu et al., 2019) with the training data, each instantiated with varying random seeds. Subsequently, the F1000raw and ARR datasets (Dycke et al., 2022) are partitioned [2] and labeled via each of the aforementioned classifiers. We now introduce two distinct ensemble methodologies as shown in Figure 1: 1) Majority labeling for Phase 1 and Phase 4. 2) Recall labeling for Phase 3.

Majority labeling Majority labeling selects the tag that receives the majority vote among the classifiers. We also compare it with consensus labeling, which retains only the reviews labeled identically. Table 2 indicates that the combination of majority labeling and only utilizing the F1000raw dataset outperforms other combinations. In scenarios of `low condition`, different random seeds, pretrained models, and learning rates are employed for training initial classifiers. F1000raw dataset is then majority labeled across four distinct models: three distinguished by their performance on the validation set (bold in Table 3), and an additional model trained on synonym-augmented data.

Recall labeling We propose a novel approach named Recall labeling to minimize the uncertainty of each label. For each pragmatic tag, we select the model with the highest recall. In descending order of their recall scores in Table 4, models label the sentences. Notably, *Other* tag consistently registered the lowest recall across all experiments. After labeling the distinct tags, any residual sentences are designated as "*Other*." To further avoid the noise from arbitrary segmentation, we intentionally omit the sentences consisting of a singular word.

[2]Using NLTK, https://www.nltk.org

Strength	Weakness	Structure
0.936	0.892	1.0
Recap	**Todo**	**Other**
0.928	0.990	0.685

Table 4: Recall scores of the best model selected for each pragmatic tag.

4.2 Synonym generation

The disparities in data quantities across domains and classes are evident in Table 1. Such class imbalances have been documented to foster biases towards the majority class, subsequently leading to diminished classification performance (Ali et al., 2013; Johnson and Khoshgoftaar, 2019). To address this prevalent issue of class imbalance, we employ data augmentation techniques to harmonize the distribution of labels in each domain. Specifically, we utilize the NLPaug[3] package to substitute nouns in each sentence with their synonymous counterparts. To ensure the quality of augmented sentences, we compute BERTSCORE (Zhang et al., 2019) between augmented and original sentences, and only add top-k augmented sentences into the training dataset.[4]

5 Results

Experiment results over different conditions and domains are presented in Table 5.

5.1 Full-data

Test data is labeled in a majority-vote manner using the best-performing models from Phase 3. The F1-score for each specific model is depicted in Figure 2. Through this methodology, the classifier achieved an F1-score of 0.838. We trained an extra model using the entire task data, including the validation set. The performance in Table 5 is derived from the inclusion of this auxiliary model within the majority labeling paradigm.

[3]https://github.com/makcedward/nlpaug

[4]The selection of k varied across domains.

	f1_mean	f1_case	f1_diso	f1_iscb	f1_rpkg	f1_scip	f1_secret
full	0.839	0.840	0.837	0.801	0.854	0.865	-
low	0.771	0.778	0.746	0.754	0.777	0.800	-
zero	0.516	0.502	0.518	0.551	0.492	0.516	-
final (full)	0.824	0.844	0.840	0.798	0.843	0.864	0.755
final (zero)	0.517	0.502	0.520	0.557	0.508	0.489	0.528

Table 5: Best model performances across the following conditions: full, low, zero, and final phases of both full and zero settings. F1 scores are computed across six distinct domains in a macro average.

5.2 Low-data

As expounded in Section 4.1, a classifier is trained utilizing the F1000raw dataset, subject to majority labeling encompassing four distinct models. We train over 25 epochs with a batch size of 8 and a learning rate of 2e-5.

5.3 Zero-data

We segment the ARR dataset into sentences and label them into 4 categories following Dycke et al. (2022): *Strength*, *Weakness*, *Recap*, and *Todo*. *Structure* tends to encompass short instructions that end with ":", in following the examples such as "Typos:" and "However a few queries:". Hence, we label all sentences that end with ":", as well as sentences of five or fewer words as *Structrue*. Lastly, *Weakness* and *Recap* are commonly mislabeled as *Other*, thus we randomly transform 15% of them into *Other*. Surprisingly, synonym generation seems to have introduced perturbations that have led to a disruption in the intended context of the original sentences, thereby slightly decreasing the performance. This could potentially be attributed to the notably lower volume of the ARR dataset compared to F1000raw.

5.4 Secret-data

We further evaluate our best models in the secret domain. In the full data setting, the exclusion of the auxiliary model mentioned in section 5.1 results in a minor decrease of 0.0003 in the F1-mean score, while the F1-secret score increases by 0.006. Notably, there exists a subtle variation in the F1-scores within the same domain under the `zero condition`, as detailed in Table 5. This variance arises due to the random allocation of *Other* tag.

5.5 Discussion

Models tend to exhibit proficiency in classifying examples that are apparent, yet encounter challenges when confronted with ambiguous reviews. Recall labeling assists the classifier, as each model specializes in distinguishing different tags. The cumulative effect of this approach is a reduction in uncertainty during the pragmatic labeling process.

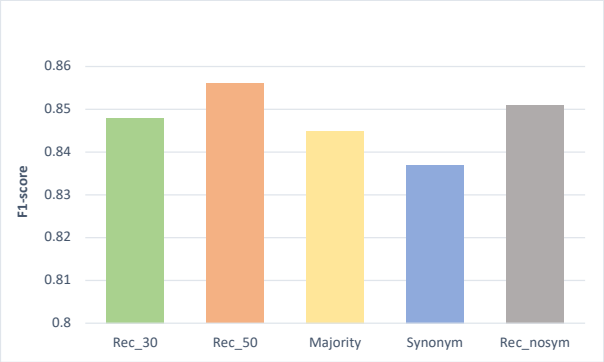

Figure 2: F1-scores of models employed for majority labeling under the full condition. Classifiers are trained using the following methods in the order from left to right: recall labeling over 30 and 50 epochs, majority labeling, synonym generation, and recall labeling among models trained without synonym generation.

6 Conclusion

In this study, we have empirically demonstrated the effectiveness of data augmentation methodologies, particularly in scenarios characterized by limited data availability. Our findings pinpoint that strategies such as pseudo-labeling and synonym generation are instrumental in leveraging unlabeled auxiliary data, therefore amplifying the generalization capacity of the classifier. Furthermore, our exploration of an ensemble approach for pseudo-labeling, aimed at maximizing certainty, suggests promising avenues for enhancing the efficacy of pragmatic tagging processes.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2021R1A2C2008855). This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea

government(MSIT) [NO.2021-0- 02068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University) & NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)]. K. Jung is with ASRI, Seoul National University, Korea.

References

Aida Ali, Siti Mariyam Shamsuddin, and Anca L Ralescu. 2013. Classification with class imbalance problem. *Int. J. Advance Soft Compu. Appl*, 5(3):176–204.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E Peters, Arie Cattan, and Ido Dagan. 2021. Cdlm: Cross-document language modeling. *arXiv preprint arXiv:2101.00406*.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Nhan Cach Dang, María N Moreno-García, and Fernando De la Prieta. 2020. Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3):483.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware bert for cross-domain sentiment analysis. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 4019–4028.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Nlpeer: A unified resource for the computational study of peer review. *arXiv preprint arXiv:2211.06651*.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. Overview of PragTag-2023: Low-resource multi-domain pragmatic tagging of peer reviews. In *Proceedings of the 10th Workshop on Argument Mining*, Singapore. Association for Computational Linguistics.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. *arXiv preprint arXiv:1903.10104*.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.

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.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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 preprint arXiv:1907.11692*.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in nlp? *arXiv preprint arXiv:2010.03863*.

Sougata Saha and Rohini Srihari. 2023. Rudolf christoph eucken at semeval-2023 task 4: An ensemble approach for identifying human values from arguments. *arXiv preprint arXiv:2305.05335*.

Hossein Soleimani and David J Miller. 2016. Semi-supervised multi-label topic models for document classification and sentence labeling. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 105–114.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. Reviewrobot: Explainable paper review generation based on knowledge synthesis. *arXiv preprint arXiv:2010.06119*.

Mark Ware and Michael Mabe. 2015. The stm report: An overview of scientific and scholarly journal publishing.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.