MuLMS-AZ: An Argumentative Zoning Dataset for the Materials Science Domain

Timo Pierre Schrader[1,6] **Teresa Bürkle**[2] **Sophie Henning**[1,3] **Sherry Tan**[4] **Matteo Finco**[2]
Stefan Grünewald[1,5] **Maira Indrikova**[2] **Felix Hildebrand**[2] **Annemarie Friedrich**[6]
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]Robert Bosch GmbH, Stuttgart, Germany [3]LMU Munich, Germany [4]TU Darmstadt, Germany
[5]University of Stuttgart, Germany [6]University of Augsburg, Germany
timo.schrader|teresa.buerkle|sophie.henning@de.bosch.com
annemarie.friedrich@informatik.uni-augsburg.de

Abstract

Scientific publications follow conventionalized rhetorical structures. Classifying the *Argumentative Zone* (AZ), e.g., identifying whether a sentence states a MOTIVATION, a RESULT or BACKGROUND information, has been proposed to improve processing of scholarly documents. In this work, we adapt and extend this idea to the domain of materials science research. We present and release a new dataset of 50 manually annotated research articles. The dataset spans seven sub-topics and is annotated with a materials-science focused multi-label annotation scheme for AZ. We detail corpus statistics and demonstrate high inter-annotator agreement. Our computational experiments show that using domain-specific pre-trained transformer-based text encoders is key to high classification performance. We also find that AZ categories from existing datasets in other domains are transferable to varying degrees.

1 Introduction

In academic writing, it is custom to adhere to a rhetorical argumentation structure to convince readers of the relevance of the work to the field (Swales, 1990). For example, authors typically first indicate a gap in prior work before stating the goal of their own research. *Argumentative Zoning* (AZ) is a natural language processing (NLP) task in which sentences are classified according to their argumentative roles with varying granularity (Teufel et al., 1999, 2009). AZ information can then be used for summarization (Teufel and Moens, 2002; El-Ebshihy et al., 2020), improved citation indexing (Teufel, 2006), or writing assistance (Feltrim et al., 2006).

Manually annotated AZ datasets (Teufel et al., 1999; Fisas et al., 2016; Soldatova and Liakata, 2007) only exist for few domains and employ differing annotation schemes. The resulting models are not directly applicable to our domain of interest, materials science research, which presents a challenging domain for current NLP methods (e.g., Mysore et al., 2019; Friedrich et al., 2020; O'Gorman et al., 2021). In this paper, we present MuLMS-AZ, the first dataset annotated for AZ in this domain. Working together with domain experts, we derive a hierarchical multi-label **annotation scheme** (see Table 1). Our scheme includes domain-specific labels such as descriptions of the materials' PREPARATION and CHARACTERIZATION, which are crucial distinctions also for NLP applications from the domain experts' view.

Label	Count	Label	Count
MOTIVATION	363	EXPLANATION	603
BACKGROUND	3155	RESULTS	2953
- PRIORWORK	1824	CONCLUSION	680
EXPERIMENT	2579	HEADING	702
- PREP.	962	CAPTION	485
- CHARACT.	1347	METADATA	210

Table 1: MuLMS-AZ label counts (multi-label).

This **resource paper** makes the following contributions:

- We present a **dataset** of 50 scientific articles (more than 10,000 sentences) in the domain of materials science manually annotated by domain experts with a hierarchical fine-grained **annotation scheme** for AZ with high agreement. The corpus will be publicly released.[1]
- We apply several neural models to our dataset that will serve as strong baselines for future work using our new benchmark. We find (a) that using domain-specific pre-trained transformers is key to a successful model, (b) that multi-task learning with existing AZ datasets leads to small benefits, and (c) that the effectiveness of transfer learning of materials science AZ labels from other corpora differs by label.

[1] https://github.com/boschresearch/mulms-az-codi2023

Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023), pages 1–15

2 Related Work

In this section, we describe related work on AZ.

AZ Datasets. Table 2 shows the statistics of several related datasets. Three larger-scale datasets manually annotated with AZ information are the AZ-CL corpus (Teufel et al., 1999; Teufel and Moens, 1999), consisting of computational linguistics papers, the Dr. Inventor Multi-Layer Scientific Corpus (DRI, Fisas et al., 2016, 2015), featuring computer graphics papers, and, closest to our domain, the ART corpus (Soldatova and Liakata, 2007), covering topics in physical chemistry and biochemistry. Appendix E explains these datasets in more detail. Teufel et al. (2009) also apply and adapt the annotation scheme of the AZ-CL corpus to the chemistry domain. Accuosto et al. (2021) label sentences with argumentation-related categories (e.g., *proposal*, *means*, *observation*). Making use of sentence-wise author-provided keywords, a dataset of about 388k abstracts with silver standard rhetorical role annotations has been derived from PubMed/MEDLINE (de Moura and Feltrim, 2018).

Modeling. AZ has been modeled as a sentence classification task using maximum entropy models (Teufel and Kan, 2009), SVMs, and CRFs (Guo et al., 2011) leveraging a variety of word, grammatical, heuristic, and discourse features (Guo et al., 2013). Ensemble-based classifiers have also been shown to be effective (Badie et al., 2018; Asadi et al., 2019). LSTM-based models relying on word embeddings have been applied to AZ and to the fundamentally very similar task of assigning zones to sentences in job ads (Liu, 2017; de Moura and Feltrim, 2018; Gnehm and Clematide, 2020). BERT-style (Devlin et al., 2019) models work well for AZ (Mo et al., 2020; Brack et al., 2022). Multi-task training has been found to be beneficial for these models both in-domain (Lauscher et al., 2018) as well as cross-domain (Brack et al., 2021).

Datasets in the Materials Science Domain. Several datasets targeting the domain of materials science research have recently been released. Mysore et al. (2019) annotate paragraphs describing synthesis procedures with graph structures capturing relations and typed arguments. Friedrich et al. (2020) mark similar graph structures corresponding to experiment information in 45 open-access publications. Several works and datasets address named entity recognition in the domain (Yamaguchi et al., 2020; O'Gorman et al., 2021).

	AZ-CL	ART	DRI	MuLMS-AZ
# docs	80	225	40	50
# sents	12818	34995	10784	10186
# labels	7	11	10	12

Table 2: Manually annotated AZ corpora.

3 Data Sources and Annotated Corpus

In this section, we present our new dataset.

Source of Texts and Preprocessing. We select 50 scientific articles licensed under CC-BY from seven sub-areas of materials science: electrolysis, graphene, polymer electrolyte fuel cell (PEMFC), solid oxide fuel cell (SOFC), polymers, semiconductors, and steel. The four SOFC papers were selected from the SOFC-Exp corpus (Friedrich et al., 2020). 11 papers were selected from the OA-STM corpus[2] and classified into the above subject areas by a domain expert. The majority of the papers were found via PubMed[3] and DOAJ[4] using queries prepared by a domain expert. For the OA-STM data, we use the sentence segmentation provided with the corpus, which has been created using GENIA tools (Tsuruoka and Tsujii, 2005). For the remaining texts, we rely on the sentence segmentation provided by INCEpTION v21.0 (Klie et al., 2018) with some manual fixes.

Annotation Scheme. AZs are functional sentence types, i.e., they capture the rhetorical function of a sentence. Together with several domain experts, we design a hierarchical scheme tailored to the materials science domain as shown in Table 3. In addition, we provide ABSTRACT, HEADING, METADATA, CAPTION, FIGURE/TABLE annotations for structural information. We assume a multi-label setting in which annotators may assign any number of labels to a sentence. Our detailed guidelines are available with our dataset.

Corpus Statistics. Documents are rather long (on average 203.7 sentences per document with a standard deviation of $\pm$73.2). There is a tendency towards long sentences (28.7 tokens per sentence on average), but with high variation of $\pm$17.9 due to, e.g., short headings. Table 1 shows how often each AZ label occurs. When ignoring tags for structural information 8133 sentences have exactly one AZ label (or the AZ label and its supertype), 1056 sentences have two labels, and 11 sentences have 3

[2] https://github.com/elsevierlabs/OA-STM-Corpus

[3] https://pubmed.ncbi.nlm.nih.gov/

[4] https://doaj.org/

Label	Description	Example
MOTIVATION	aims/motivation of the study	*In this study, we perform a systematic analysis of ...*
BACKGROUND	textbook-like technical background	*The method is based on the Kelvin equation.*
- PRIORWORK	specific prior work relevant to current study	*Irmawati et al. has concluded that ...*
EXPERIMENT	description of the experiment	*We evaluate PtCo nanoparticle catalyst ...*
- PREPARATION	steps describing the preparation of samples	*The mixture was subjected to stirring for 60 minutes.*
- CHARACT.	characterizations and characterization techniques of the involved materials	*Ni foam surface coverage of the WO3 thin film and its homogeneity were analyzed by energy–dispersive X-ray spectroscopy (EDS).*
EXPLANATION	statements (hypotheses or assumptions) relevant to results or experimental settings	*In our calculation, all Pt loadings were considered to be electrochemically active.*
RESULTS	details on experimental results	*The hydrogen adsorption/desorption peak is at about 0.2V.*
CONCLUSION	conclusions and take-aways	*This result indicated that ...*

Table 3: Content-based MuLMS-AZ Argumentative Zoning sentence labels.

labels. Labels are similarly distributed across data splits (see Appendix D).

Inter-Annotator Agreement. Our entire dataset has been annotated by a single annotator, a graduate student of materials science, who was also involved in the design of the annotation scheme. We compare the annotations of this main annotator to those of another annotator who holds a Master's degree in materials science and a PhD in engineering. The agreement study is performed on 5 documents (357 sentences). Due to the multi-label scenario, following Krippendorff (1980) we measure κ (Cohen, 1960) for each label separately, comparing whether each annotator used a particular label on an instance or not (see Table 4). Our annotators achieve "substantial" agreement (Landis and Koch, 1977) on most labels, "perfect" agreement on identifying HEADINGS (see also Appendix D). Lower, though still "moderate", agreement on MOTIVATION, EXPLANATION and CONCLUSION can in part be explained by their lower frequency which makes it generally harder to obtain high κ-values. Intuitively, they also have a more difficult nature compared to the other tags, e.g., we observe disagreements regarding what constitutes a MOTIVATION or an EXPLANATION versus what is purely reporting BACKGROUND. The full confusion matrix and a discussion of agreement on subtags are given in Appendix D; a discussion of multi-label examples can be found in Appendix F.

Our scores are in the same ballpark as those reported by Teufel et al. (1999) on a similar annotation task. For their 7-way task, they report κ scores around 0.71-0.75, with differences between categories in one-vs-all measurements ranging from about 0.49 to 0.78. In sum, we conclude that agreement on AZ is satisfactory in our dataset.

AZ Label	κ	AZ Label	κ
HEADING	0.89	METADATA	0.76
MOTIVATION	0.44	BACKGROUND	0.75
CONCLUSION	0.55	EXPERIMENT	0.78
EXPLANATION	0.39	RESULTS	0.70

Table 4: IAA for AZ on 357 sentences.

4 Modeling

We model AZ as a multi-label classification problem, using BERT (Devlin et al., 2019) as the underlying text encoder. We also test domain-specific pre-trained variants of BERT. SciBERT (Beltagy et al., 2019) has been pre-trained on articles in the scientific domain. MatSciBERT (Gupta et al., 2022) is a version of SciBERT further pre-trained on materials science articles. We use the CLS embedding as input to a linear layer, transform logits using a sigmoid function and choose labels if their respective score exceeds 0.5. For multi-task experiments with other datasets, we use a single shared language model and one linear output layer per dataset. For hyperparameters, see Appendix A.

As shown in Table 1, the dataset suffers from strong class imbalance. Classifiers tend to underperform on minority labels (Johnson and Khoshgoftaar, 2019). To address this problem, we apply the **multi-label random oversampling** (ML-ROS, Charte et al., 2015) algorithm during training. The main idea behind ML-ROS is to dynamically duplicate instances of minority classes while taking the multi-label nature of the problem into account. In a nutshell, the algorithm performs several oversampling iterations, keeping track of the imbalance ratios associated with each label and choosing instances that carry minority labels until a predefined number of additional samples have been chosen. Details are given in Appendix B.

Method	LM	mic.-F1	mac.-F1
No Oversampling	BERT	$72.6_{\pm 1.0}$	$65.5_{\pm 0.7}$
	MatSciBERT	$76.3_{\pm 0.7}$	$70.1_{\pm 0.7}$
	SciBERT	$76.2_{\pm 0.9}$	$70.2_{\pm 0.6}$
ML-ROS	SciBERT	$76.7_{\pm 0.7}$	$70.6_{\pm 0.9}$
+ MultiTask ART	SciBERT	$75.0_{\pm 0.9}$	$68.9_{\pm 1.1}$
+ MultiTask AZ-CL	SciBERT	$\mathbf{77.2}_{\pm 0.3}$	$\mathbf{71.1}_{\pm 0.5}$
*human agreement**		*78.7*	*74.9*

Table 5: AZ classification results on MuLMS-AZ test set. *Not directly comparable: computed on documents from agreement study (see Appendix D).

Training data	PM Label	P	R
PM, AZ-CL, ART, DRI	OBJECTIVE	36.1	28.3
PM, AZ-CL, ART, DRI	BACKGROUND	84.2	40.0
PM, ART, DRI	METHOD	58.1	74.7
PM, ART, DRI	RESULT	82.4	30.9
PM, ART, DRI	CONCLUSION	43.5	29.9
MuLMS-AZ	OBJECTIVE	56.8	54.3
MuLMS-AZ	BACKGROUND	82.1	78.8
MuLMS-AZ	METHOD	79.9	78.2
MuLMS-AZ	RESULT	82.1	83.2
MuLMS-AZ	CONCLUSION	43.5*	29.9*

Table 6: Results for transfer learning experiment. Precision and recall on MuLMS-AZ test set. *not a typo.

5 Experimental Results

We here detail our experimental results.

Settings. We split our corpus into train, dev, and test sets of 36, 7, and 7 documents. For all experiments and for hyperparameter tuning, we always train five models. The training data is split into five folds. Similar to cross-validation, we train on four folds and use the fifth fold for model selection (cf. van der Goot, 2021), repeating this process five times (also for hyperparameter tuning). The dev set is only used for tuning, and we report scores for the five models on test. In this setting, deviations are naturally higher than when reporting results for the same training data. For hyperparameters and implementation details, see Appendix A. To evaluate our experiments, we use hierarchical precision, recall, and F1 (Silla and Freitas, 2011). These scores operate on the sets of labels per instance, always including the respective supertypes of gold or predicted labels.

Results. Table 5 shows the performance of our neural models on MuLMS-AZ. Overall, the categories can be learned well, approaching our estimate of human agreement. SciBERT clearly outperforms BERT, i.e., using domain-specific embeddings is a clear advantage. However, MatSciBERT does not add upon SciBERT. We hence conduct the remaining experiments using SciBERT. Using ML-ROS results in minor increases for most labels (see also Appendix G). When multi-task learning with the AZ-CL dataset (using 40% of its samples), further increases are observed. It is worth noting that multi-task training with ART does not result in increases although the chemistry domain should be much closer to our domain. This might indicate that despite the domain discrepancy, AZ annotations in AZ-CL are more compatible with ours.

As a first step to explaining what part of rhetorical information can be induced based only on data from other corpora, we perform a transfer learning experiment. We carefully manually map the AZ labels of the various datasets (see Appendix E) to the coarse-grained categories used by PubMed (PM). Using these mapped labels, we train binary classifiers that aim to detect the presence of a particular PM label. As training data, we use ART, DRI, and a selection of documents from the PM dataset by de Moura and Feltrim (2018) that were published in materials science journals (see Appendix C). We add AZ-CL to the training data only if an unambiguous mapping of its categories to the PM label in question is possible. Here, we use the dev set of MuLMS-AZ for model selection and hyperparameter tuning. Results for running the resulting classifiers on MuLMS-AZ are reported in Table 6. For BACKGROUND and RESULTS, we observe a high precision, which indicates that similar rhetorical elements may be used. OBJECTIVE and METHOD seem to differ most across datasets as they are likely very domain- and problem-specific. When training with mapped labels on the entire MuLMS-AZ, we observe much higher recall scores across all label groups, again indicating the usefulness of our in-domain training data.

6 Conclusion and Outlook

We have presented a new AZ corpus in the field of materials science annotated by domain experts with high agreement. Our experimental results demonstrate that strong classifiers can be learned on the data and that AZ labels can be transferred from related datasets only to a limited extent.

Our new dataset opens up new research opportunities on cross-domain AZ, class imbalance scenarios, and integrating AZ information in information extraction tasks in materials science.

Limitations

This resource paper describes the dataset in detail, providing strong baselines and first initial cross-domain experiments. It does not aim to provide an extensive set of experiments on cross-domain argumentative zoning yet.

The entire dataset is only singly-annotated. The agreement study was performed on complete documents and hence has only limited data for several labels. Due to the limited funding of the project, we could double-annotate the entire dataset.

Finally, we only test one model class (BERT-based transformers). A potential next step is to test a bigger variety of models and embeddings. Because AZ labels are interdependent within a document, especially document-level models or CRF-based models are promising methods to try. We have also tested only one method (multi-label random oversampling) to deal with the strong class imbalance in the dataset. We have not yet tested further such methods (Henning et al., 2023) or data augmentation methods.

Ethical Considerations

We took care of potential license issue of the data underlying our dataset by exclusively selecting open-access articles published under CC BY.

The main annotator was paid above the minimum wage of our country in the context of a full-time internship. The annotator was aware of the goal of the study and consents to the public release of the data. The remaining domain experts participated on a voluntary basis due to their interest in the topic.

References

Pablo Accuosto, Mariana Neves, and Horacio Saggion. 2021. Argumentation mining in scientific literature: From computational linguistics to biomedicine. In *Frommholz I, Mayr P, Cabanac G, Verberne S, editors. BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval; 2021 Apr 1; Lucca, Italy. Aachen: CEUR; 2021. p. 20-36.* CEUR Workshop Proceedings.

Nasrin Asadi, Kambiz Badie, and Maryam Tayefeh Mahmoudi. 2019. Automatic zone identification in scientific papers via fusion techniques. *Scientometrics*, 119(2):845–862.

Kambiz Badie, Nasrin Asadi, and Maryam Tayefeh Mahmoudi. 2018. Zone identification based on features with high semantic richness and combining results of separate classifiers. *Journal of Information and Telecommunication*, 2(4):411–427.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Arthur Brack, Anett Hoppe, Pascal Buschermöhle, and Ralph Ewerth. 2021. Sequential sentence classification in research papers using cross-domain multi-task learning. *CoRR*, abs/2102.06008.

Arthur Brack, Anett Hoppe, Pascal Buschermöhle, and Ralph Ewerth. 2022. Cross-domain multi-task learning for sequential sentence classification in research papers. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–13.

Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. 2015. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3–16. Recent Advancements in Hybrid Artificial Intelligence Systems and its Application to Real-World Problems Progress in Intelligent Systems Mining Humanistic Data.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

G. Bennemann de Moura and V. Delisandra Feltrim. 2018. Using lstm encoder-decoder for rhetorical structure prediction. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 278–283, Los Alamitos, CA, USA. IEEE Computer Society.

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.

Alaa El-Ebshihy, Annisa Maulida Ningtyas, Linda Andersson, Florina Piroi, and Andreas Rauber. 2020. ARTU / TU Wien and artificial researcher@ LongSumm 20. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 310–317, Online. Association for Computational Linguistics.

Valéria D Feltrim, Simone Teufel, Maria Graças V das Nunes, and Sandra M Aluísio. 2006. Argumentative zoning applied to critiquing novices' scientific abstracts. In *Computing Attitude and Affect in Text: Theory and Applications*, pages 233–246. Springer.

Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3081–3088, Portorož, Slovenia. European Language Resources Association (ELRA).

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 42–51, Denver, Colorado, USA. Association for Computational Linguistics.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-exp corpus and neural approaches to information extraction in the materials science domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268, Online. Association for Computational Linguistics.

Ann-Sophie Gnehm and Simon Clematide. 2020. Text zoning and classification for job advertisements in German, French and English. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 83–93, Online. Association for Computational Linguistics.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 273–283, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Yufan Guo, Roi Reichart, and Anna Korhonen. 2013. Improved information structure analysis of scientific documents through discourse and lexical constraints. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 928–937, Atlanta, Georgia. Association for Computational Linguistics.

Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. 2022. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Klaus Krippendorff. 1980. *Krippendorff, Klaus, Content Analysis: An Introduction to its Methodology. Beverly Hills, CA: Sage, 1980.* Sage Publications, Inc.

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. 2018. Investigating the role of argumentation in the rhetorical analysis of scientific publications with neural multi-task learning models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3326–3338, Brussels, Belgium. Association for Computational Linguistics.

Maria Liakata and Larisa Soldatova. 2008. Guidelines for the annotation of general scientific concepts. *Aberystwyth University, JISC Project Report http://ie-repository. jisc. ac. uk/88*.

Haixia Liu. 2017. Automatic argumentative-zoning using word2vec. *CoRR*, abs/1703.10152.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Wang Mo, Cui Yunpeng, Chen Li, and Li Huan. 2020. A deep learning-based method of argumentative zoning for research articles. *Data Analysis and Knowledge Discovery*, 4(6):60–68.

Sheshera Mysore, Zachary Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. 2019. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 56–64, Florence, Italy. Association for Computational Linguistics.

Tim O'Gorman, Zach Jensen, Sheshera Mysore, Kevin Huang, Rubayyat Mahbub, Elsa Olivetti, and Andrew McCallum. 2021. MS-mentions: Consistently annotating entity mentions in materials science procedural text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1337–1352, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Carlos N Silla and Alex A Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1):31–72.

Larisa Soldatova and Maria Liakata. 2007. An ontology methodology and cisp-the proposed core information about scientific papers. *JISC Project Report*.

John M. Swales. 1990. Discourse analysis in professional contexts. *Annual Review of Applied Linguistics*, 11:103–114.

Simone Teufel. 2006. Argumentative zoning for improved citation indexing. In *Computing attitude and affect in text: Theory and Applications*, pages 159–169. Springer.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Simone Teufel and Min-Yen Kan. 2009. Robust argumentative zoning for sensemaking in scholarly documents. In *Advanced language technologies for digital libraries*, pages 154–170. Springer.

Simone Teufel and Marc Moens. 1999. Discourse-level argumentation in scientific articles: human and automatic annotation. In *Towards Standards and Tools for Discourse Tagging*.

Simone Teufel and Marc Moens. 2002. Articles summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

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, Singapore. Association for Computational Linguistics.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 467–474, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Rob van der Goot. 2021. We need to talk about train-dev-test splits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4485–4494, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kyosuke Yamaguchi, Ryoji Asahi, and Yutaka Sasaki. 2020. SC-CoMIcs: A superconductivity corpus for materials informatics. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6753–6760, Marseille, France. European Language Resources Association.

Appendix

A Hyperparameters

We implement all our models using PyTorch. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer for all our models and set the batch size to $16/32$ depending on what works best and GPU restrictions. The learning rate stays constant after a linear warmup phase. We set a dropout rate to 0.1 for the linear layer that takes the contextualized embeddings that are produced by BERT as input. Early stopping is applied if the micro-F1 score has not improved for more than 3 epochs. Binary cross entropy is the loss function for the MuLMS-AZ output layer, whereas cross entropy is the loss function used for optimizing the multi-task output heads corresponding to the other AZ datasets. Table 7 lists the various learning rates found during grid search. We tested different learning rates between 1e-4 and 1e-7. A refinement of the grid was done after an initial search, which almost always leads to a second search area within the range of 1e-6 to 9e-6. When using ML-ROS, we oversample by 20%. Training was performed on a single Nvidia A100 GPU or alternatively V100 GPU.

Method	LM	Learning Rate
No Oversampling	BERT	3e-6
	MatSciBERT	8e-6
	SciBERT	3e-6
ML-ROS	SciBERT	2e-6
+ MT (+PM)	SciBERT	7e-6
+ MT (+ART)	SciBERT	2e-6
+ MT (+AZ-CL)	SciBERT	2e-6
+ MT (+DRI)	SciBERT	1e-6
+ MT (+ART+AZ+DRI)	SciBERT	8e-6
Data Augm. (+PM)	SciBERT	8e-6

Table 7: Learning rates of the different model reported in Table 12

B Multi-Label Random Oversampling (ML-ROS) Algorithm

Figure 1 details our adaption of the multi-label random oversampling (ML-ROS) algorithm originally proposed by Charte et al. (2015). In the initialization (lines 3-7), for each label, all the instances that carry a particular label are collected in what Charte et al. call a *bag*. The main part of the algorithm (lines 10-24) does the following: For each label y, the *Imbalance Ratio per label* (IRLbl), which is the ratio between the count of the most frequent label and the count of y, is calculated:

$$\mathrm{IRLbl}(y) = \frac{\max_{y' \in L} \sum_{i=1}^{|D|} h(y', Y_i)}{\sum_{i=1}^{|D|} h(y, Y_i)}$$

D is the dataset, L is the label set, Y_i is the set of labels assigned to the i-th sample and h is an indicator function evaluating if $y \in Y_i$. Hence, the larger the value, the less frequently y occurs compared to the most frequent label.

The per-label values are then used to determine the *mean imbalance ratio* (MeanIR):

$$\mathrm{MeanIR} = \frac{1}{|L|} \sum_{y' \in L} \mathrm{IRLbl}(y')$$

For each of the labels with an imbalance ratio exceeding the current MeanIR, a random instance of this label is duplicated.

The main part is repeated until the pre-specified size of the oversampled dataset is reached. Our implementation differs from Charte et al. in that we update meanIR in each iteration step and also oversample labels originally not being a minority label when their IRLbl exceeds MeanIR at the beginning of an iteration step.

C List of Materials Science Journals

We used the list of materials-science related journals collected on Wikipedia to filter for abstracts in the PubMed Medline corpus published in journals.[5]

D Further Corpus Statistics for MuLMS-AZ

Table 8 gives the counts of sentences carrying a particular AZ label. Distributions are similar across data splits. Table 8 also lists counts for ABSTRACT, which we decide to exclude from our modeling experiments because including it resulted in performance decreases due to confusion with other labels. Locating the abstract in a document can usually be solved in rule-based ways as abstracts of publications are commonly available in a machine-readable format.

During annotation, we introduced two subtypes of EXPLANATION, HYPOTHESIS and ASSUMPTION, distinguishing between scientific hypotheses and assumptions made by the author in cases where often choices are possible. As the overall counts

[5] https://en.wikipedia.org/w/index.php?title=List_of_materials_science_journals&oldid=1078212543

```
Inputs: <Dataset> D, <Percentage> P
Outputs: Oversampled dataset
samplesToDuplicate <-- |D|/100 * P # P % size increment
L <-- labelsInDataset(D) # Obtain the full set of labels
for each label in L do # Bags of samples for each label
    Bag_label <-- getSamplesPerLabel(label)
end for

while samplesToDuplicate > 0 do # Loop duplicating instances
    MeanIR <-- calculateMeanIR(D, L)
    # Gather minority bags (bag: all instances of a given label)
    minBags = []
    for each label in L do
        IRLbl_label <-- calculateIRperLabel(D, label)
        if IRLbl_label > MeanIR then
            minBags += Bag_label
        end if
    end for
    # Duplicate a random sample from each minority bag
    for each minBag_i in minBags do
        x <-- random(1, |minBag_i|)
        duplicateSample(minBag_i, x)
        -- samplesToDuplicate
    end for
end while
```

Figure 1: Pseudocode for adapted (dynamic) ML-ROS algorithm.

and agreement were low, we decided to only use the supertype EXPLANATION in all experiments.

Figure 2a shows the label coincidence matrix between the two annotators in the inter-annotator agreement study, i.e., how often each pair of labels co-occurred on an instance. For all labels except MOTIVATION, the majority of coincidences occur on the diagonal. RESULTS is the label most mixed up with others, possibly because these sentences often are long and also contain interpretative information of the other rhetorical types.

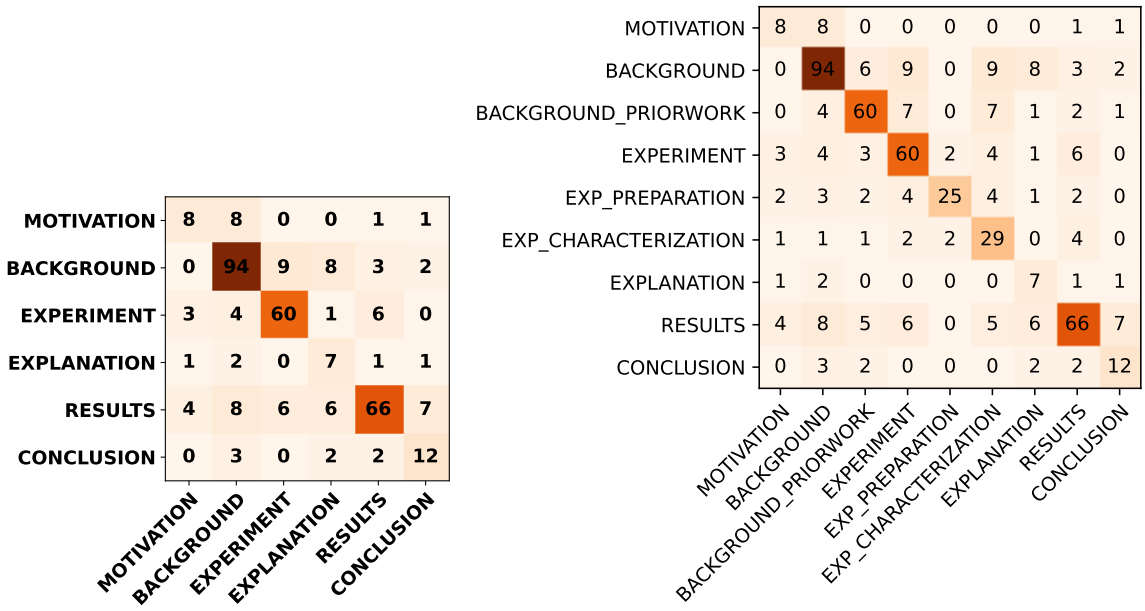
Figure 2a breaks this information down the level including subtypes. CHARACTERIZATION and PREPARATION are rarely confused by the domain experts. Similarly, BACKGROUND and PRIORWORK are reliably distinguished.

Agreement on sub-labels. Our agreement study contained only 12 CAPTION instances. Data inspection showed that the additional (not the main) annotator neglected to use this tag where appropriate, using only content-related tags on these instances. There were also not enough instances of the subtypes PREPARATION and EXPERIMENT_CHARACTERIZATION to measure agreement. On identifying the subtype BACKGROUND_PRIORWORK, annotators achieve a κ of 0.8, with (minor) disagreements mainly with regard to using BACKGROUND or its subtype.

Label	total	train	dev	test
MOTIVATION	363	273	44	46
BACKGROUND	3155	2423	440	292
-PRIORWORK	1824	1387	265	172
EXPERIMENT	2579	1896	394	289
-CHARACTERIZATION	1347	982	200	165
-PREPARATION	962	705	146	111
EXPLANATION	603	430	91	82
RESULTS	2953	2146	440	367
CONCLUSION	680	507	106	67
ABSTRACT	269	190	28	51
CAPTION	485	309	91	85
HEADING	702	536	96	70
METADATA	210	142	40	28

Table 8: **Label counts** on the complete dataset and on data split subsets. **Multi-label counts:** Number of sentences in which the label is present. Due to multi-labeling, the sum of these columns exceeds the total amount of sentences. For hierarchical labels, the super-label count includes all sub-label counts.

Agreement on HEADING. As it should be straightforward to identify headings, we looked at the 6 cases that one annotator labeled as HEADING but not the other. We found 4 cases to result from broken formatting. One METADATA sentence was wrongly labeled HEADING, and the remaining HEADING sentence was missed by the other annotator.



(a) Coincidence matrix for coarse AZ labels.

(b) Coincidence matrix for AZ labels with subtypes.

Figure 2: Coincidence matrices of inter-annotator agreement study for AZ labels on 357 sentences.

	Precision	Recall	F1	support
micro avg.	77.3	80.0	78.7	
macro avg.	75.0	76.1	74.9	
HEADING	100.0	81.2	89.7	32
METADATA	75.0	81.8	78.3	11
MOTIVATION	50.0	44.4	47.1	18
BACKGROUND	77.0	89.5	82.8	105
-PRIORWORK	77.9	90.9	83.9	66
EXPERIMENT	80.0	84.5	82.2	71
-PREPARATION	92.6	73.5	82.0	34
CHARACTERIZATION	61.7	78.4	69.0	37
RESULTS	85.7	70.2	77.2	94
CONCLUSION	50.0	66.7	57.1	18

Table 9: Human agreement computed in terms of hierarchical precision, recall, and F1.

Human "upper bound". In order to provide a *rough* estimate of how humans would perform on the classification task, we use the data from the agreement study to compute hierarchical precision, recall, and F1 scores. Due to insufficient data for the remaining labels, we only compute the scores over the following labels: HEADING, METADATA, MOTIVATION, BACKGROUND, PRIORWORK, EXPERIMENT, PREPARATION, CHARACTERIZATION, RESULTS, and CONCLUSION. Table 9 reports detailed scores per label. Scores have been computed using scikit-learn[6].

[6] https://scikit-learn.org/stable

E Description and Comparison of AZ Datasets.

In this section, we provide a detailed description and comparison of existing AZ datasets. The various corpora try to capture very similar information. However, each corpus defines its set of labels in a slightly different way. Table 10 lists the various labels used in the datasets and groups labels used for the same or very similar purpose. Table 11 shows the label distributions of the corpora.

AZ-CL corpus. The Argumentative Zoning (AZ, Teufel et al., 1999; Teufel and Moens, 1999) corpus[7] consists of 80 manually annotated open-access **computational linguistics** research articles. Sentences are marked according to their argumentative zone or rhetorical function as one of the following classes: AIM, BACKGROUND, BASIS, CONTRAST, OTHER, OWN or TEXT. Inter-annotator agreement is reported as substantial ($\kappa = 0.71$). The distribution of classes is quite skewed towards OTHER and OWN.

ART corpus. The ART corpus[8] (Soldatova and Liakata, 2007) covers topics in **physical chemistry** and **biochemistry**. Articles are annotated according to the CISP/CoreSC annotation scheme (Liakata and Soldatova, 2008). Sentences are

[7] https://github.com/WING-NUS/RAZ

[8] https://www.aber.ac.uk/en/cs/research/cb/projects/art/art-corpus/

PubMed	AZ-CL	ART	DRI	MuLMS-AZ	Description
OBJECTIVE	AIM	HYPOTHESIS MOTIVATION GOAL	CHALLENGE	MOTIVATION	A sentence describing the research target, goal, aim or the motivation for the research.
BACKGROUND	BACKGROUND CONTRAST BASIS	BACKGROUND	BACKGROUND	BACKGROUND PRIORWORK	A statement concerning the knowledge domain or previous related work.
METHOD	OWN	OBJECT, METHOD MODEL EXPERIMENT OBSERVATION	APPROACH	EXPERIMENT PREPARATION CHARACTERIZ. EXPLANATION	A sentence describing the research procedure, models used, or observations made during the research.
RESULT	OWN	RESULT	OUTCOME	RESULTS EXPLANATION	A sentence describing the study findings, effects, consequences, and/or analysis of the results.
CONCLUSION	OWN	CONCLUSION	OUTCOME FUTUREWORK	CONCLUSION	A statement concerning the support or rejection of the hypothesis or suggestions of future research.
–	TEXT OTHER	–	SENTENCE UNSPECIFIED	–	Example sentences, broken sentences, etc.

Table 10: AZ Corpus Zones Mapping and Descriptions. Compare to Table 3.

labeled with one of the categories HYPOTHESIS, MOTIVATION, GOAL OF INVESTIGATION, BACKGROUND, OBJECT OF INVESTIGATION, RESEARCH METHOD, MODEL, EXPERIMENT, OBSERVATION, RESULT or CONCLUSION. The annotation scheme also defines subcategories for some of these. The corpus has been annotated by domain experts. In a preliminary study, κ was measured as 0.55, however, for the final corpus, only the annotators that had the highest average agreement were selected. Hence, the agreement in the final corpus is expected to be higher.

DRI corpus. The Dr. Inventor Multi-Layer Scientific Corpus[9] (DRI, Fisas et al., 2016, 2015), contains 40 scientific articles taken from the domain of **computer graphics**. Each of the 10,784 sentences was annotated with one of the rhetorical categories: CHALLENGE, BACKGROUND, APPROACH, OUTCOME or FUTUREWORK. They have also included two other categories SENTENCE for sentences that are characterized by segmentation or character encoding errors and UNSPECIFIED for sentences where identification is not possible. Also to note was the possibility to annotate a combination of two different categories as seen in the example of: OUTCOME_CONTRIBUTION, CHALLENGE_GOAL and CHALLENGE_HYPOTHESIS. Manual annotation reaches a κ value of 0.66.

PubMed corpus. The PubMed corpus[10] (de Moura and Feltrim, 2018) contains abstracts of papers in the **biomedical** domain extracted from PUBMED/MEDLINE. The collected abstracts were written in English and annotated with predefined section names by their authors; based on the mapping provided by the U.S. National Library of Medicine (NLM), the section names were collapsed into five rhetorical roles: BACKGROUND, OBJECTIVE, METHODS, RESULTS, and CONCLUSIONS. The abstracts that did not contain the five mentioned rhetorical roles were removed from the dataset with the resulting corpus containing close to 5 million sentences. The dataset is not particularly challenging: a simple CRF model achieves an F-score of 93.75, an LSTM-based model achieves 94.77 according to de Moura and Feltrim (2018).

F Examples

In this section, we present and discuss several examples from our dataset.

F.1 Example Sentences

- MOTIVATION: *Therefore, it is highly desirable to develop an innovative technology to raise the mass activity of Ir-based OER catalysts to the targeted level.*

[9] http://sempub.taln.upf.edu/dricorpus

[10] https://github.com/dead/rhetorical-structure-pubmed-abstracts

Dataset	Label	Count
AZ-CL	OWN	8624
	OTHER	2019
	BACKGROUND	789
	CONTRAST	600
	AIM	313
	BASIS	246
	TEXT	227
ART	RESULT	7373
	BACKGROUND	6657
	OBSERVATION	4659
	METHOD	3751
	MODEL	3456
	CONCLUSION	3083
	EXPERIMENT	2841
	OBJECT	1190
	HYPOTHESIS	656
	GOAL	548
	MOTIVATION	466
DRI	APPROACH	5038
	BACKGROUND	1760
	SENTENCE	1247
	OUTCOME	1175
	UNSPECIFIED	759
	CHALLENGE	351
	OUTCOME_CONTRIBUTION	219
	FUTUREWORK	136
	CHALLENGE_HYPOTHESIS	7
MatSci PubMed	RESULTS	1282
	OBJECTIVE	1264
	METHODS	1198
	CONCLUSION	380
	BACKGROUND	60

Table 11: Label counts for the different AZ corpora.

- BACKGROUND: *For photocatalytic water splitting using photoelectrochemical cells (PECs), the charge carriers are created from the photovoltaic effect close to the catalytic site.*
- PRIORWORK: *Proton exchange membrane (PEM) electrolysis, which occurs in acidic electrolytes (pH 0–7), has better efficiency and enhanced ramping capability over other types of electrolysis [7].*
- EXPERIMENT: *In order to find an optimum efficiency of the PV–electrolysis, different combinations of the electrolyzer with A-CIGS-based thin film solar cell modules with different band gaps of the cell were examined.*
- PREPARATION: *Pre-sputtering was performed for 5 min in argon plasma in order to remove surface impurities.*
- CHARACTERIZATION: *The current density-potential (j–V) characteristics of the A-CIGS cells were recorded under simulated AM 1.5G sunlight in a set-up with a halogen lamp (ELH).*
- EXPLANATION: *A possible explanation for the superior ECSA-specific activity in the 3D WP-structured catalysts is efficient removal of oxygen bubbles from the catalyst layer.*
- RESULTS: *The load curves were similar for the electrolyzers with different WO3 thin films and the lowest potential needed for 10 mA cm-2 in the overall reaction was 1.77 V.*
- CONCLUSION: *The Cu-N- rGO demonstrated superior catalytic activity to the counterpart N-rGO, and enhanced durability compared to commercial Pt/C.*

Structural tags are used, for example, in the following cases.

- HEADING: *4. Discussion and concluding remarks*
- METADATA: *This research was funded by Hubei Superior and Distinctive Discipline Group of "Mechatronics and Automobiles" (No.XKQ2019009).*
- CAPTION: *Figure 8. Enlarged view of the shaded portion of Figure 7.*

F.2 Multi-Label Examples

In contrast to earlier works on AZ, our approach to labeling AZ in materials science publications uses a multi-label approach. In this section, we discuss some multi-label examples.

- BACKGROUND, PRIORWORK, RESULTS: *This indicates that the HER follows a rate-determining Volmer or Heyrovsky step for different sputtering conditions without any order [40,41].* In this example, a result obtained in the current paper confirms a result known from prior work.
- EXPERIMENT, CHARACTERIZATION, RESULTS, EXPLANATION: *Attributing this enthalpy release exclusively to the removal of grain boundaries in stage B, a specific grain boundary energy(2)γ=Hρ3dini-1-dfin-1=0.85$\pm$0.08Jm-2is estimated using the initial and final crystallite diameters of stage B, as given above (dini=222nm, dfin=764nm),*

as well as the Cu bulk value of 8.92gcm-3 for the mass density ρ. The first subordinate clause of this sentence (*Attributing ... stage B*) is an EXPLANATION. The remainder of the sentence states a CHARACTERIZATION.

- BACKGROUND, PRIORWORK, RESULTS, CONCLUSION: *Furthermore, the fatigue life decreased approximately by more than 12% when the pre-corroded time was doubled, and the fatigue life decreased approximately by more than 11% when the applied stress level was doubled, indicating that both pre-corroded time and applied stress level can significantly affect the fatigue life of specimens, which shows a good agreement with the previous works [37,38].* This example illustrates a case where our simplification of labeling entire sentences comes to its limits: The first part of the sentence (*Furthermore ... was doubled*) reports RESULTS while the second part draws a CONCLUSION drawing connections to specific PRIORWORK.

G Detailed Results

In this section, we provide detailed results for the experiments presented in the main part of the paper.

Table 13 (no oversampling and ML-ROS) and Table 14 (multi-task AZ-CL) show the results in terms of precision, recall and (hierarchical) F1 for each label individually. We report the results on both dev and test of the specific model that performed best on dev compared to all other models.

First, we compare the difference between no oversampling at all and when using ML-ROS. As shown in Table 1, MOTIVATION, METADATA, and CAPTION are the least frequent labels in our dataset. Except for METADATA on the test set, there is always an increase in terms of F1-score when applying ML-ROS on minority labels during training. The biggest increase of 5.8 happened for MOTIVATION on the test set. Furthermore, there is also an improvement of 1.2 points on dev and 2.5 points on test in terms of F1-score for EXPLANATION, which is fourth in the list of rarest AZ labels.

During our experimentation, we observed that ML-ROS tends to be especially helpful for models that show strong performance on majority labels, but not on minority labels. Other models with different hyperparameters achieve even better scores on minority labels without oversampling; however, they tend to have worse overall performance.

Method	LM	mic.-F1	mac.-F1
No Oversampling	BERT	$72.6_{\pm 1.0}$	$65.5_{\pm 0.7}$
	MatSciBERT	$76.3_{\pm 0.7}$	$70.1_{\pm 0.7}$
	SciBERT	$76.2_{\pm 0.9}$	$70.2_{\pm 0.6}$
ML-ROS	SciBERT	$76.7_{\pm 0.7}$	$70.6_{\pm 0.9}$
+ MT (+PM)	SciBERT	$76.5_{\pm 0.4}$	$69.5_{\pm 0.5}$
+ MT (+ART)	SciBERT	$75.0_{\pm 0.9}$	$68.9_{\pm 1.1}$
+ MT (+AZ-CL)	SciBERT	$\mathbf{77.2}_{\pm 0.3}$	$\mathbf{71.1}_{\pm 0.5}$
+ MT (+DRI)	SciBERT	$76.6_{\pm 0.3}$	$70.5_{\pm 0.4}$
+ MT (+ART+AZ+DRI)	SciBERT	$76.4_{\pm 0.6}$	$70.2_{\pm 0.5}$
Data Augm. (+PM)	SciBERT	$77.1_{\pm 0.8}$	$70.8_{\pm 1.3}$
*human agreement**		*78.7*	*74.9*

Table 12: Results on MuLMS-AZ test set, hierarchical micro/macro F1: MT=Multi-Task models, *not directly comparable.

Next, we describe the effects of **multi-task training** with the AZ-CL dataset. We also apply ML-ROS to MuLMS-AZ in our multi-task experiments. Both micro-F1 and macro-F1 increase by 0.5 points in terms of micro- and macro-F1 when using multi-tasking instead of ML-ROS only. Most of the per-label F1-scores increased when using multi-tasking, interestingly with notable differences for CHARACTERIZATION (4.8) and METADATA (5.6). We conclude that multi-tasking with AZ-CL helps supporting common majority labels even though the domain of this dataset is clearly different from ours.

In contrast, multi-task learning with the other datasets consistently resulted in *decreases* of performance. The chemistry domain is intuitively closest to that of materials science, hence, we would have expected ART to be a good additional dataset in multi-task learning. Brack et al. (2022) provide some insights into cross-domain learning of AZ categories using datasets from biomedicine, chemistry, and computer graphics. Our MuLMS-AZ, alongside AZ-CL, opens up new research opportunities.

In addition, we perform a **data augmentation** experiment using AZ data from scientific abstracts of the PubMed Medline corpus[11], filtering for abstracts that were published in journals related to the materials science domain (see Appendix C). We map the four PubMed AZ labels BACKGROUND, OBJECTIVE, RESULTS, and CONCLUSIONS to our four AZ labels BACKGROUND, MOTIVATION, RESULTS and CONCLUSION. Augmenting with data from the PubMed Medline dataset also helps to

[11] https://www.nlm.nih.gov/databases/download/pubmed_medline.html

achieve better performance. However, the micro-F1 score is 0.1 lower and the macro-F1 score is 0.3 lower compared to the MT (+AZ-CL) model. On the other hand, training is much more time-efficient since a low augmentation percentage of 10% is sufficient to get good results.

Label	dev			test			
	P	R	H. F1	P	R	H. F1	Count
SciBERT, no oversampling							
MOTIVATION	65.5	46.8	54.4	68.5	36.5	47.6	363
BACKGROUND	89.2	80.0	**84.3**	85.0	76.6	80.6	3155
-PRIORWORK	97.0	84.5	**90.3**	92.9	67.9	78.4	1824
EXPERIMENT	82.1	85.8	83.9	80.6	82.6	**81.6**	2579
-CHARACTERIZATION	72.0	68.9	**70.3**	75.8	67.3	**71.1**	962
-PREPARATION	65.2	65.1	65.0	78.6	69.7	**73.7**	1347
EXPLANATION	46.3	33.0	38.4	55.0	35.9	43.4	603
RESULTS	75.0	84.6	79.5	79.9	85.9	82.8	2953
CONCLUSION	56.7	55.3	**56.0**	42.4	43.0	**42.6**	680
CAPTION	92.4	75.2	82.9	80.9	68.9	74.4	485
HEADING	84.8	97.9	90.9	87.4	96.6	91.7	702
METADATA	93.1	68.0	78.6	78.6	72.9	**75.2**	210
Average	**76.6**	70.4	72.9	**75.5**	67.0	70.2	
SciBERT, ML-ROS							
MOTIVATION	56.3	55.9	**55.9**	72.9	43.0	**53.4**	363
BACKGROUND	82.2	84.8	83.5	79.7	84.2	**81.9**	3155
-PRIORWORK	96.0	84.5	89.9	90.5	71.3	**79.7**	1824
EXPERIMENT	85.1	83.2	**84.1**	81.1	81.7	81.4	2579
-CHARACTERIZATION	73.3	67.3	70.1	73.2	67.5	70.2	962
-PREPARATION	69.4	63.4	**66.3**	73.8	69.5	71.5	1347
EXPLANATION	45.7	35.2	**39.6**	53.4	40.2	**45.9**	603
RESULTS	77.6	83.4	**80.4**	83.6	83.8	**83.7**	2953
CONCLUSION	60.6	44.5	51.3	46.8	35.2	40.1	680
CAPTION	91.7	79.6	**85.2**	77.9	73.6	**75.7**	485
HEADING	85.4	97.5	**91.1**	90.6	96.3	**93.4**	702
METADATA	89.3	70.5	**78.8**	61.9	80.0	69.8	210
Average	76.1	**70.8**	**73.0**	73.8	**68.9**	**70.6**	

Table 13: Per label scores on dev and test of MuLMS-AZ in terms of precision, recall, and hierarchical F1. **Bold**: best result for label. P, R, and F1 scores are averages over the P, R, F1 scores of 5 folds each.

Label	dev			test		
	P	R	H. F1	P	R	H. F1
MOTIVATION	62.7	54.1	58.0	71.2	43.9	54.3
BACKGROUND	85.6	82.1	83.8	80.9	81.6	81.2
-PRIORWORK	95.4	84.2	89.4	93.7	68.8	79.3
EXPERIMENT	83.6	82.8	83.2	83.1	83.0	83.0
-CHARACTERIZATION	73.7	65.9	69.3	77.4	73.0	75.0
-PREPARATION	69.4	55.6	61.7	79.4	67.2	72.8
EXPLANATION	42.6	35.8	38.8	51.2	35.9	41.7
RESULTS	76.6	84.4	80.3	81.5	85.1	83.2
CONCLUSION	61.8	49.6	55.0	41.0	32.8	36.4
CAPTION	90.5	77.6	83.5	79.2	76.2	77.7
HEADING	84.7	97.7	90.7	88.9	97.4	92.9
METADATA	84.3	72.0	77.6	70.6	81.4	75.4

Table 14: Per label scores on dev and test in terms of precision, recall, and hierarchical F1 using multi-task learning with the AZ-CL dataset, SciBERT, ML-ROS.