Overview of the First Shared Task on *Detecting Entities in the Astrophysics Literature (DEAL)*

Felix Grezes[1], **Thomas Allen**[1], **Tirthankar Ghosal**[2], **Sergi Blanco-Cuaresma**[1]
[1]Center for Astrophysics, Harvard & Smithsonian, USA
[2]Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Czech Republic
[1](felix.grezes,thomas.allen,sblancocuaresma)@cfa.harvard.edu
[2]ghosal@ufal.mff.cuni.cz

Abstract

In this article, we describe the overview of our shared task: Detecting Entities in the Astrophysics Literature (DEAL). The DEAL shared task was part of the Workshop on Information Extraction from Scientific Publications (WIESP) in AACL-IJCNLP 2022[1]. Information extraction from scientific publications is critical in several downstream tasks such as identification of critical entities, article summarization, citation classification, etc. The motivation of this shared task was to develop a community-wide effort for entity extraction from astrophysics literature. Automated entity extraction would help to build knowledge bases, high-quality meta-data for indexing and search, and several other use-cases of interests. Thirty-three teams registered for DEAL, twelve of them participated in the system runs, and finally four teams submitted their system descriptions. We analyze their system and performance and finally discuss the findings of DEAL.

1 Introduction

A good amount of astrophysics research makes use of data coming from missions and facilities such as ground observatories in remote locations or space telescopes, as well as digital archives that hold large amounts of observed and simulated data. These missions and facilities are frequently named after historical figures or use some ingenious acronym which, unfortunately, can be easily confused when searching for them in the literature via simple string matching. For instance, "Planck" can refer to the person, the mission, the constant, or several institutions. Automatically recognizing entities such as missions or facilities would help tackle this word sense disambiguation problem. In our DEAL shared task, we instigate a community initiative to extract "entities of interest" from astrophysics publications.

[1]https://ui.adsabs.harvard.edu/WIESP/

2 Task

2.1 Definition

The shared task *Detecting Entities in the Astrophysics Literature (DEAL)* (Grezes et al., 2022) consists of Named Entity recognition (NER) on samples of text extracted from astrophysics publications indexed by NASA ADS (Kurtz et al., 2000). The labels were created by domain experts and designed to identify entities of interest to the astrophysics community. They range from simple to detect (ex: URLs) to highly unstructured (ex: Formula), and from useful to researchers (ex: Telescope) to more useful to archivists and administrators (ex: Grant).

2.2 Evaluation

Submissions were scored using both the CoNLL-2000 shared task seqeval F1-Score at the entity level and scikit-learn's Matthews correlation coefficient method at the token level. We also encouraged authors to propose their own evaluation metrics. The task baseline was computed using the astroBERT model (Grezes et al., 2021).

3 Dataset Description

3.1 Data Collection and Creation

The dataset [2] consists of text fragments obtained from the astrophysical literature. The journals that the text fragments were obtained from are the Astrophysical Journal, Astronomy & Astrophysics, and the Monthly Notices of the Royal Astronomical Society. All text fragments are from recent publications, between the years of 2015 and 2021. Each text fragment originates from one of two parts of an article. The first are fragments from the full-text, consisting of all sections of the body of the article, excluding the abstract and acknowledgment

[2]The data is openly available under the CC-BY-4.0 licence huggingface.co/datasets/adsabs/WIESP2022-NER

Proceedings of the 1st Workshop on Information Extraction from Scientific Publications, pages 1–7
Nov 20, 2022.

sections. The second are fragments from the acknowledgment section of the article.

Thirty-three different entities, comprised of general and astrophysical entities, were manually labeled in each text fragment by a domain expert. The entities that were labeled cover a number of broad categories. One category contains common NER entities, such as Person, Organization, and Location. A second category contains entities related to astrophysical facilities, such as Observatory and Telescope. A third category contains entities related to research funding and proposals, such as Grant or Proposal. A fourth category contains entities relating to astronomical objects and regions. Finally there is a category that contains various entities that are found in the literature, such as URL's and citations.

3.2 Data Segmentation for Shared Task

The overall dataset was separated into four components: the development dataset, the training dataset, the testing dataset, and the validation dataset. The development dataset is a small dataset of only twenty text fragments used to aid in the development of modeling systems. The training dataset consists of 1741 text fragments, 887 of which are from the full-text and 854 of which are from the acknowledgments. Table 3 shows the the number of labeled entities and origin of the text fragment for these entities. The testing dataset consists of 2495 text fragments, 1201 of which are from the full-text and 1294 of which are from the acknowledgments. Table 3 shows the the number of labeled entities and origin of the text fragment for these entities. Finally, the validation dataset consists of 2505 text fragments for the purpose of scoring the submitted models.

4 Participant Systems

Ghosh et al. (2022) proposed an Astro-mT5 model for entity recognition from Astrophysics publications. Primarily, they fine-tune a multilingual Text-To-Text Transfer Transformer (T5) model on the downstream task followed by sequence-labelling using Conditional Random Field (CRF) to get the probability sequence over the possible sequence labels.

Huang (2022) propose a system that uses data augmentation as a low-cost method of teacher-student training to transfer domain-specific knowledge to a larger adapter-based model. The author introduce a framework that uses data augmentation from domain-specific pre-trained models to transfer domain-specific knowledge to larger general pre-trained models for the underlying DEAL task. Specifically, they use the adapter architecture of the DeBERTaV3-large model as the backbone model, and CosmicRoBERTa (a further pretrained version of SpaceRoBERTa, a domain-specific model), as the augmentation teacher model.

Dai and Karimi (2022) investigate two different NER methods, word-based tagging and span-based classification for the DEAL task. They show that their span-based method using RoBERTa-large pre-trained models outperform the widely used word-based sequence tagging method (which uses BIO annotation schema).

Kaan Alkan et al. (2022) proposed a majority voting strategy of a SciBERT-based ensemble models for the DEAL task. Specifically, they used outputs from 32 different SciBERT-based classifiers for the majority voting strategy.

5 astroBERT Baseline

The shared task submissions were evaluated using F-1 score and the Matthews correlation coefficient (MCC) metrics. The F-1 score is a standard measure of model quality and was computed using seqeval (Nakayama, 2018), which uses micro-averaging and ignores the 'O' label. The MCC takes into account every value in the confusion matrix and is generally regarded as a balanced measure; it was computed using scikit-learn (Pedregosa et al., 2011). The F-1 score was computed at the entity level and the MCC score was computed at the token level. Using two metrics help prevent with a submission overfitting by optimizing for a single score.

As a baseline, we finetuned three BERT variants on the shared task. The original BERT from Google (Devlin et al., 2018), SciBERT from AllenAI (Beltagy et al., 2019), and astroBERT from NASA/ADS (Grezes et al., 2021). Each variant was finetuned on the training dataset for 1000 epochs (~5 hours each on dual V100 NVIDIA GPUs).

Table 2 provides the scores of the baselines on the WIESP datasets. Additionally, a model making random predictions based on label frequency was

model	astroBERT		Augmentation (Huang, 2022)		Word vs Span (Dai and Karimi, 2022)		Ensemble (Kaan Alkan et al., 2022)		**Astro-mT5** (Ghosh et al., 2022)	
Metric \ Split	val	test	val	test	val	test	val	test	val	test
MCC	0.8104	0.7939	0.9063	0.8928	0.9138	0.8946	0.9139	**0.8978**	0.9129	0.8954
F-1	0.5779	0.5561	0.7988	0.7799	0.8307	0.7990	0.8262	0.7993	0.8364	**0.8057**
Precision	0.5508	0.5387	0.7854	0.7854	0.8249	0.8076	0.8145	0.8013	0.8296	**0.8137**
Recall	0.6077	0.5746	0.8126	0.7744	0.8366	0.7906	0.8382	0.7972	0.8434	**0.7979**
Accuracy	0.9389	0.9308	0.9692	0.9633	0.9718	0.9640	0.9718	**0.9651**	0.9714	0.9642

Table 1: Main DEAL@WIESP 2022 Shared Task Results. F-1, Precision and Recall are computed using micro-averaging.

included for comparison. Additional standard metric scores (overall precision, recall and accuracy) are included as well. These additional metrics were also computed for the shared task submissions and provided to the participants but were not used to rank them. For each metric, astroBERT outscored BERT and SciBERT. A finer comparison between astroBERT and SciBERT is provided in the appendix figure 1, as well as the confusion matrix between labels for astroBERT on the testing dataset in appendix figure 2.

6 Results and Analysis

We report the results of the four teams that submitted their system papers in table 1 which were also the best performers of the twelve shared task participants on both F-1 score and MCC metrics. All three systems significantly outperform the astroBERT baseline, and are built on top of pre-existing publicly available language models.

7 Findings of DEAL

Each participants system significantly outperformed the baseline using different techniques. Below are the findings each system that we believe to be of importance to the community. From the top participant system astro-mT5 by Ghosh et al. (2022), we highlight the use of Conditional Random Fields (CRF), which validate other studies showing that CRFs help on NER tasks. Kaan Alkan et al. (2022) found that using ensemble methods to combine multiple models made for more robust predictions. Dai and Karimi (2022) concluded that span-based methods outperform word-based. They also showed that non-astrophysics tokenizer may suffer from over-segmentation when applied to astronomy papers. Finally, Huang (2022) highlighted the usefulness of data augmentation when applied to a dataset the size of WIESP.

8 Conclusion and Future Directions

All the participant systems were built on top of existing language models (i.e. general English, not tailored to a domain), and significantly beat the baseline scores. This begs the question: how would these systems performs when built on top of a language model and tokenizer tailored to astronomy? Based on the competition results, the use of CRFs seems especially promising. Furthermore, the wide variety in the methods used by the successful participant systems indicate that the task is far from solved, and that many improvements can be made to the astroBERT baseline.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv e-prints*, page arXiv:1903.10676.

Xiang Dai and Sarvnaz Karimi. 2022. Detecting entities in the astrophysics literature: A comparison of word-based and span-based entity recognition methods. In *Proceedings of the 1st Workshop on Information Extraction from Scientific Publications*, Taipei, Taiwan. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, page arXiv:1810.04805.

Madhusudan Ghosh, Payel Santra, Sk Asif Iqbal, and Partha Basuchowdhuri. 2022. Astro-mt5: Entity extraction from astrophysics literature using mt5 language model. In *Proceedings of the 1st Workshop on Information Extraction from Scientific Publications*, Taipei, Taiwan. Association for Computational Linguistics.

Felix Grezes, Thomas Allen, Tirthankar Ghosal, and Sergi Blanco-Cuaresma. 2022. Overview of the first shared task on detecting entities in the astrophysics literature (deal). In *Proceedings of the 1st Workshop on Information Extraction from Scientific Publications*, Taipei, Taiwan. Association for Computational Linguistics.

model	Random			BERT			SciBERT			**astroBERT**		
Metric \ Split	train	val	test	train	val	test	train	val	test	train	val	test
MCC	0.1037	0.1083	0.1057	0.7542	0.7405	0.7229	0.8159	0.8019	0.7844	0.8296	0.8104	**0.7939**
F-1	0.0170	0.0166	0.0162	0.4920	0.4739	0.4513	0.5867	0.5601	0.5355	0.6138	0.5779	**0.5561**
Precision	0.0122	0.0119	0.0116	0.4995	0.4780	0.4622	0.5753	0.5463	0.5313	0.5889	0.5508	**0.5387**
Recall	0.0278	0.0273	0.0269	0.4848	0.4698	0.4409	0.5986	0.5745	0.5398	0.6409	0.6077	**0.5746**
Accuracy	0.7146	0.7059	0.6876	0.9256	0.9188	0.9094	0.9430	0.9366	0.9280	0.9468	0.9389	**0.9308**

Table 2: Evaluation of the three BERT baselines. F-1, Precision and Recall are computed using micro-averaging.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S. Grant, Donna M. Thompson, Roman Chyla, Stephen McDonald, Timothy W. Hostetler, Matthew R. Templeton, Kelly E. Lockhart, Nemanja Martinovic, Shinyi Chen, Chris Tanner, and Pavlos Protopapas. 2021. Building astroBERT, a language model for Astronomy & Astrophysics. *arXiv e-prints*, page arXiv:2112.00590.

Po-Wei Huang. 2022. Domain specific augmentations as low cost teachers for large students. In *Proceedings of the 1st Workshop on Information Extraction from Scientific Publications*, Taipei, Taiwan. Association for Computational Linguistics.

Atilla Kaan Alkan, Cyril Grouin, Fabian Schussler, and Pierre Zweigenbaum. 2022. A majority voting strategy of a scibert-based ensemble models for detecting entities in the astrophysics literature (shared task). In *Proceedings of the 1st Workshop on Information Extraction from Scientific Publications*, Taipei, Taiwan. Association for Computational Linguistics.

Michael J. Kurtz, Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Stephen S. Murray, and Joyce M. Watson. 2000. The NASA Astrophysics Data System: Overview. , 143:41–59.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

A Appendix

	Training		Testing		Total
Label \ Section	Ack	Full Text	Ack	Full Text	
Archive	628	30	1119	50	1827
CelestialObject	110	4521	113	5615	10359
CelestialObjectRegion	0	488	7	1344	1839
CelestialRegion	8	390	27	581	1006
Citation	1097	23665	1650	31923	58335
Collaboration	855	49	1214	45	2163
ComputingFacility	1188	20	1644	9	2861
Database	461	54	649	152	1316
Dataset	102	594	182	1005	1883
EntityOfFutureInterest	0	77	52	724	853
Event	213	8	340	7	568
Fellowship	1426	0	2096	0	3522
Formula	0	10521	4	17856	28381
Grant	7532	26	14610	24	22192
Identifier	68	75	156	145	444
Instrument	224	630	367	1064	2285
Location	1843	28	2932	55	4858
Mission	56	81	143	161	441
Model	64	2980	174	6110	9328
O	59549	412758	86353	553386	1112046
ObservationalTechniques	4	194	1	141	340
Observatory	1713	195	2469	378	4755
Organization	21562	97	31954	87	53700
Person	6081	41	9539	97	15758
Proposal	176	24	312	40	552
Software	679	810	1050	883	3422
Survey	707	751	969	1003	3430
Tag	0	120	0	148	268
Telescope	1044	1136	1699	1627	5506
TextGarbage	14	92	3	483	592
URL	262	44	342	110	758
Wavelength	61	4906	106	7210	12283
Total	107727	465405	162276	632463	1367871

Table 3: Counts of labels in training and test datasets according source origination. Note, that "O" refers to unlabeled words. (note: 'Ack' stands for Acknowledgment)

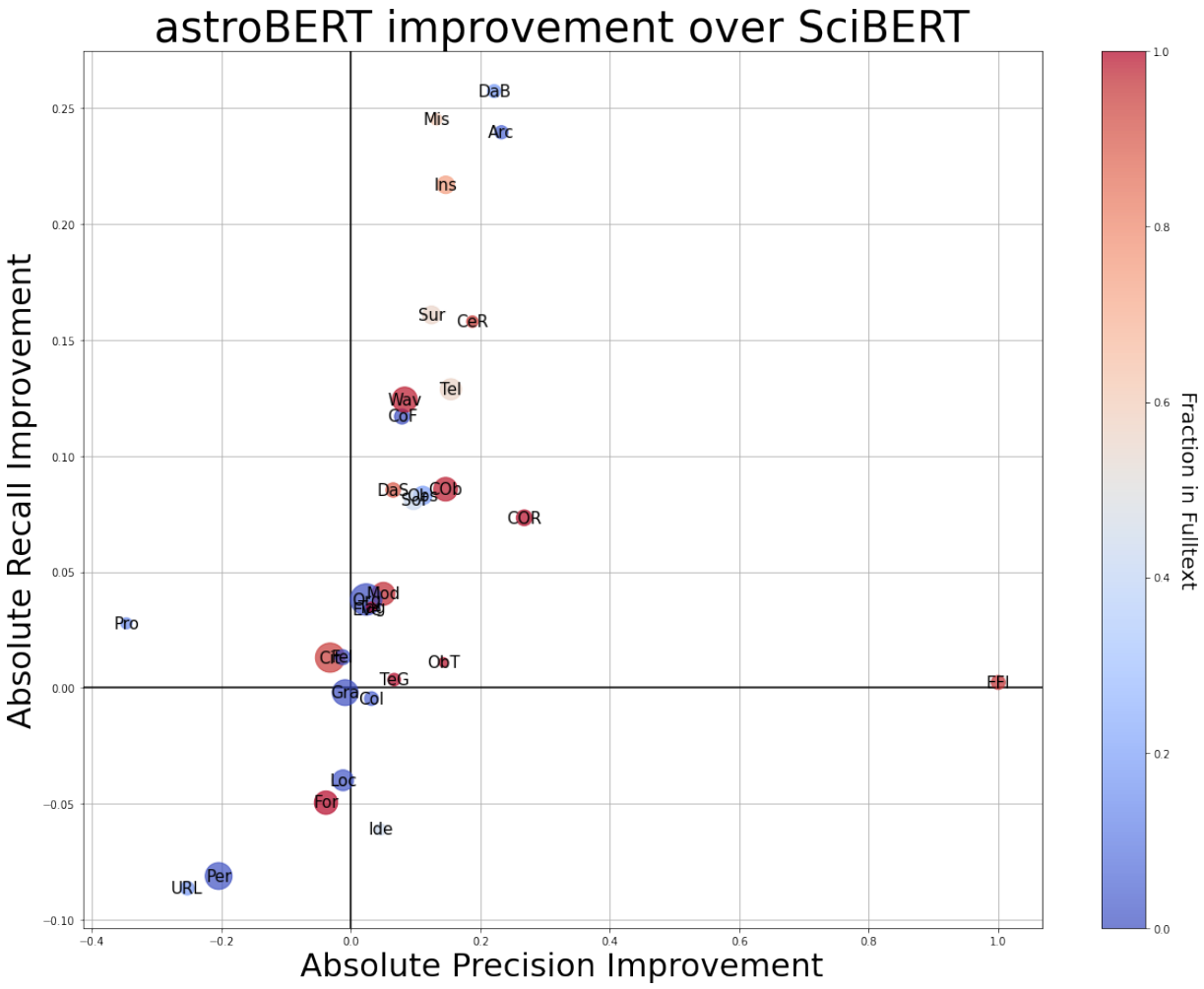

Figure 1: Absolute improvement from astroBERT over SciBERT in precision and recall for each class over the WIESP-TESTING data set, colored by predominance of that class body or acknowledgment sections.

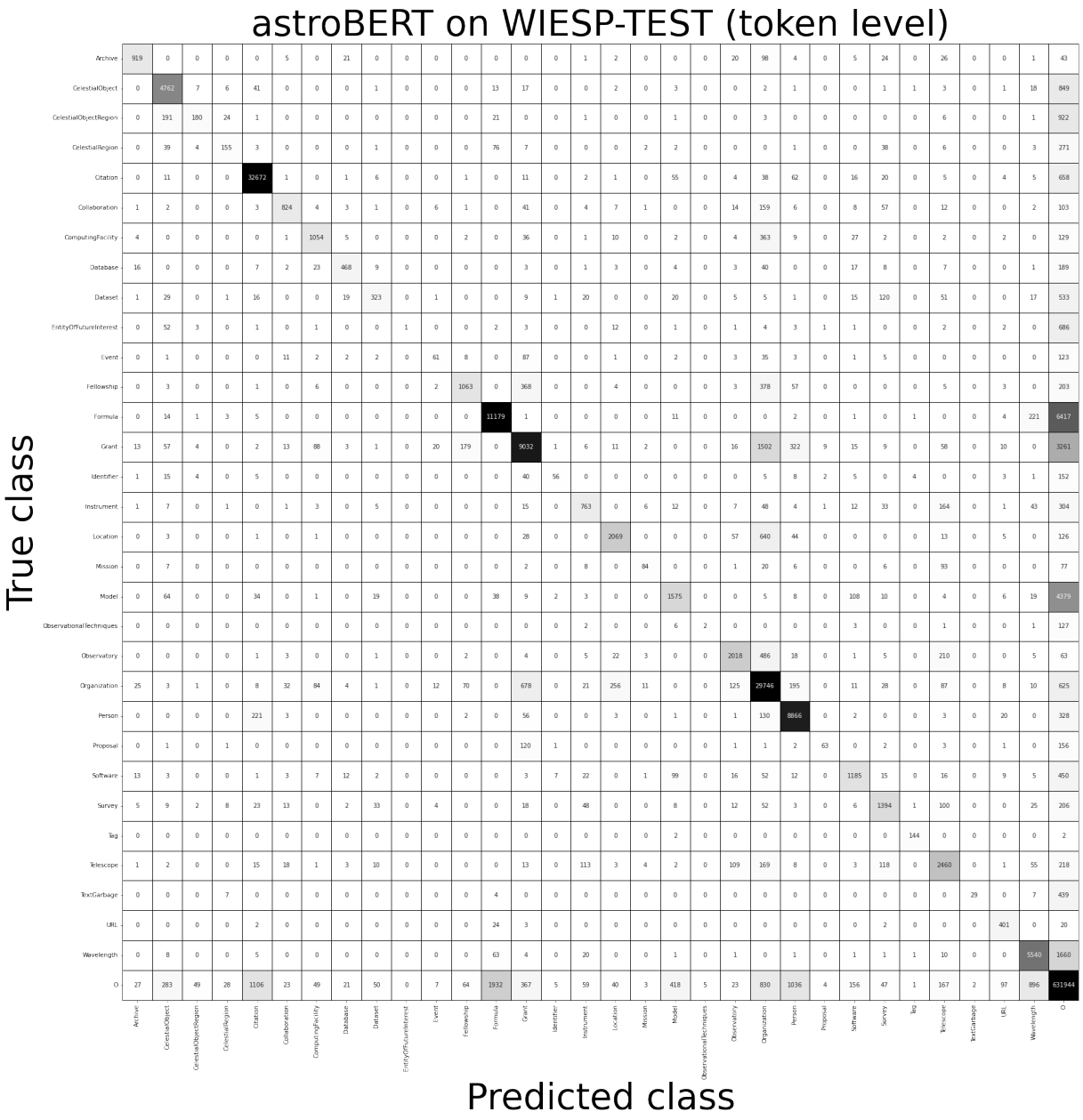

Figure 2: Confusion Matrix between actual labels and predicted labels from astroBERT on the WIESP-TEST tokens.