Astro-mT5: Entity Extraction from Astrophysics Literature using mT5 Language Model

Madhusudan Ghosh [*], **Payel Santra** [*], **Sk Asif Iqbal, Partha Basuchowdhuri**
Indian Association for the Cultivation of Science, Kolkata
{madhusuda.iacs, payel.iacs, skasifiqbal31}@gmail.com,
partha.basuchowdhuri@iacs.res.in

Abstract

Scientific research requires reading and extracting relevant information from existing scientific literature in an effective way. To gain insights over a collection of such scientific documents, extraction of entities and recognizing their types is considered to be one of the important tasks. Numerous studies have been conducted in this area of research. In our study, we introduce a framework for entity recognition and identification of NASA astrophysics dataset, which was published as a part of the DEAL SharedTask. We use a pre-trained multilingual model, based on a natural language processing framework for the given sequence labeling tasks. Experiments show that our model, **Astro-mT5** [1], outperforms the existing baseline in astrophysics related information extraction.

1 Introduction

Extracting information about entities and their relationships from unstructured text is an important area in natural language processing (NLP). Fast evolution of many scientific disciplines has led to a continuous influx of a large number of research papers into the publication repositories (e.g., Arxiv [2], Anthology [3] and Biorxiv [4]). Literature study is an important step in any scientific study. Till now, this has been limited to human effort i.e., the amount of previous literature that an individual is exposed to is limited to human capabilities. This may lead to a few fundamental problems, such as, the researcher's inability to find out relevant previous works and identify suitable baselines for performance comparison. It poses a significant problem due to the limitation of human abilities, unless a framework could be designed to obtain the literature corpus by using machine learning methods. To overcome this problem, Luan et al. (2018); Jain et al. (2020); Hou et al. (2021); Mondal et al. (2021) proposed an end-to-end information extraction (IE) system from AI-based scientific documents for preparing suitable knowledge graph (KG). Recently, fine-tuning of pre-trained language models (PLMs) have shown remarkable performance on IE task such as named entity recognition (NER), and relation extraction (RE) from unstructured text in NLP (Baldini Soares et al., 2019). Self-supervised pre-training allows these PLMs to learn highly accurate linguistic, semantic, and factual information from a significant quantity of unlabeled data (Wang et al., 2022).

While tremendous progress has been made in the field of AI, the area of astrophysics has been rarely explored as an area of application by AI researchers. The current search engine of NASA Astrophysics Data System (ADS) shows poor performance on information retrieval (IR) tasks due to the absence of suitable KG (Grezes et al., 2021). Grezes et al. (2021) have recently developed astroBERT, a language model pre-trained on astrophysics literature, in order to apply it to downstream tasks of NLP in the astrophysics domain.

The first edition of the shared task, named Detecting Entities from Astrophysics Literature (DEAL) took place in 2022, for building a system that is capable of extracting fine-grained entities of different *categories* such as CelestialObjectRegion, CelestialRegion, Instrument (Grezes et al., 2022).

We participated in DEAL SharedTask 2022 and proposed a neural architecture based model to identify the required entities from a collection of astrophysics articles. Our proposed model namely, Entity Extraction from Astrophysics Literature using mT5 Language Model (**Astro-mT5**), seeks to devise a transfer learning strategy by fine-tuning the mT5 (Xue et al., 2020) model. Furthermore, we apply conditional random field (CRF) decoder to

[*] Equal Contribution to this work
[1] Our source code is available at https://github.com/MLlab4CS/Astro-mT5.git
[2] http://arxiv.org
[3] http://anthology.org
[4] http://www.biorxiv.org

Proceedings of the 1st Workshop on Information Extraction from Scientific Publications, pages 100–104
Nov 20, 2022. ©2022 Association for Computational Linguistics

implement the entity extraction task. Experimental results show that the proposed framework achieves state-of-the-art (SOTA) performance on this task.

2 Related Works

Recently, researchers have explored many directions for applying information extraction from the unstructured text of scientific articles written in english. Adding token-level classifiers or CRFs above the sentence encoders is a popular strategy for NER task (Chiu and Nichols, 2015; Strubell et al., 2017; Ma and Hovy, 2016). Pennington et al. (2014) empirically showed that using pre-trained word embeddings such as GloVe along with CNN-based (Ma and Hovy, 2016) and LSTM-based models (Lample et al., 2016) produced better results on the NER task. Recent releases of transformer based PLMs such as BERT (Devlin et al., 2019), SciBERT (Beltagy et al., 2019), T5 (Raffel et al., 2019), RoBERTa (Liu et al., 2019), BigBIRD (Zaheer et al., 2020), ALBERT (Lan et al., 2019), RemBERT (Chung et al., 2020), and Longformer (Beltagy et al., 2020) showed a significant performance improvement in many downstream tasks in NLP such as NER and RE. Additionally, Akbik et al. (2018) developed an easy-to-use interface namely FLAIR, that allows users to fine-tune any word embedding and any PLMs to produce improved results on the NER task. There has been a recent surge in proposing multilingual pre-trained language models such as mBERT (Devlin et al., 2018), mBART (Liu et al., 2020), XLM-R (Conneau et al., 2019), mT5 for achieving SOTA results in many downstream IE tasks in NLP.

3 System Description

Given a sentence from the astrophysics documents, we follow a sequence labeling approach for the fine-grained entity extraction task. Formally, given a sentence (word sequence) $s = (\mathrm{w}_1, \mathrm{w}_2 ... \mathrm{w}_\mathrm{n})$, the objective is to learn a function f_θ (parameterized by θ) that maps an observed sequence of embedded vectors to a sequence of labels $f_\theta : (\mathbf{w_1}, \ldots, \mathbf{w_n}) \rightarrow (y_1, \ldots, y_n)$, where each $\mathbf{w_i} \in \mathbb{R}^d$ is an embedded vector of the token w_i, and each $y_i \in \{\mathrm{B}, \mathrm{I}, \mathrm{O}\}$ stands for a label, which indicates if it is the beginning, continuation or the end of a text span (in the context of our work - predicted entity category). Given a set of examples of such $\mathcal{D} = \{(s, y)\}$ sequence pairs, the parameters θ of a sequence classification models are learned by optimizing the cross-entropy loss. In our work we use FLAIR [5], a neural framework, proposed by Akbik et al. (2018). In this framework, we employ **mT5** as a base pre-trained language model, which gets fine-tuned on our downstream entity extraction sequence labeling task. Say, the internal output representation produced at the fine-tuning stage is $\mathbf{x_i} \in \mathbb{R}^{d_1}$. Then, we pass it to the CRF decoder layer and get the probability sequence over the possible sequence labels $\mathbf{y}$ by using the Eqns. 1 and 2.

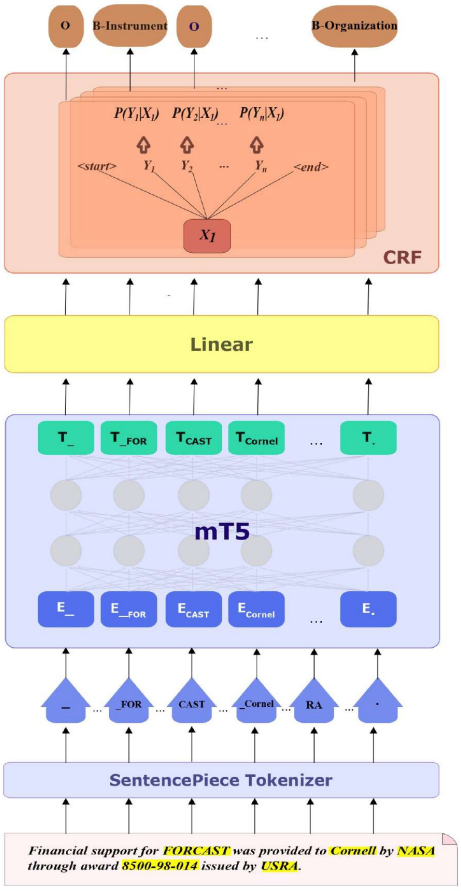

Figure 1: The overall architecture of our proposed model **Astro-mT5**

$$P(\mathbf{y_{0:n}}|\mathbf{x_{0:n}}) \propto \prod_{i=1}^{n} \phi_i(\mathbf{y_{i-1}}, \mathbf{y_i}, \mathbf{x_i}) \quad (1)$$

where,

$$\phi_i(\mathbf{y_{i-1}}, \mathbf{y_i}, \mathbf{x_i}) = exp(\mathbf{W}_{\mathbf{y_{i-1}},\mathbf{y_i}}\mathbf{x_i} + \mathbf{b}_{\mathbf{y_{i-1}},\mathbf{y_i}}) \quad (2)$$

Here, W, $\mathbf{b} \in \mathbb{R}^{d_2}$ are the required parameters, which are trained during end-to-end training of our model.

[5] https://github.com/flairNLP/flair

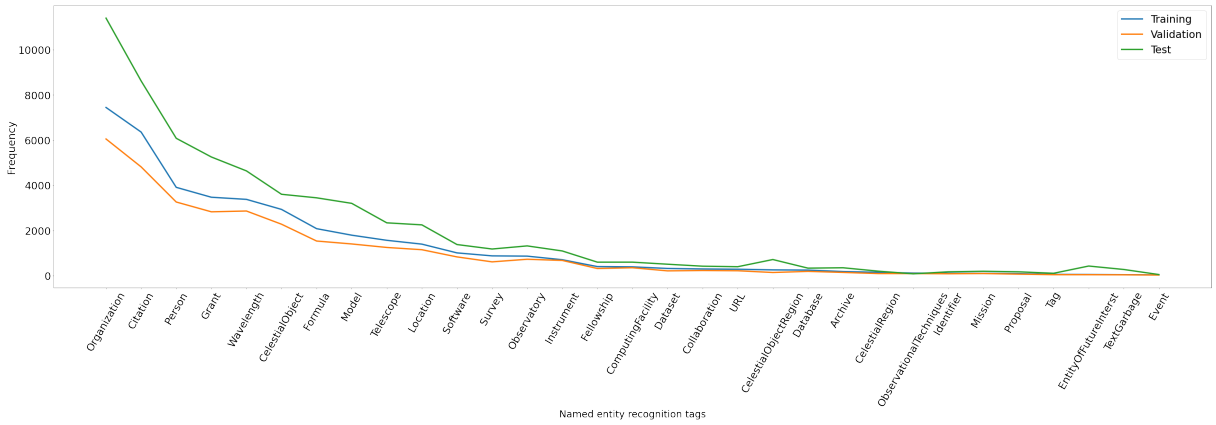

Figure 2: Frequency of each NER tag in the training, testing, and validation dataset.

The overall architecture diagram of our proposed model, **Astro-mT5**, has been shown in Fig. 1.

4 Experimental Setup

4.1 Data Description

Throughout our experiments, we have used the astrophysics dataset released for the DEAL Shared-Task. The dataset mainly consists of the full texts and the acknowledgement sections of a collection of astrophysics articles. The total number of entity categories used for the sequence classification task is 32 and the frequency of the categories (NER tags) has been depicted in Fig. 2. The dataset statistics has been shown in Table 1. A sample snippet of the dataset has been presented in Table 2.

Dataset	# of Samples
Training Data	1753
Validation Data	1366
Test Data	2505

Table 1: DEAL Dataset statistics

DEAL Dataset
The question of whether the Sun$^{\text{CelestialObject}}$ acts (magnetically) as other SLS$^{\text{EntityofFutureInterest}}$ is difficult to answer. If all such stars are indeed magnetically similar, it implies that stars have a consistent magnetic variability over stretches only 0.01 million years into the past (Wu et al. 2018$^{\text{Citation}}$).

Table 2: A sample snippet of the tagged astrophysics data.

4.2 Implementation Details

We submitted three experimental results in different settings against the released test data. We apply the stratified train-test split[6] strategy with a splitting ratio of 80:10:10 on the released training dataset to train and tune our model accordingly. We fine-tune different transformer based language models and apply separate subtoken pooling strategy at the penultimate layer of the used language model. For our first submission, namely DEAL_1, we use `xlm-roberta-large` [7] language model by applying subtoken pooling operation namely 'first and last' on the internal transformer embeddings. For our second and third submissions, namely DEAL_2 and DEAL_3 (**Astro-mT5**), we fine-tune `mt5-large` [8] language model with different pooling strategies such as 'first' and also 'first and last' on the transformer embeddings. In all the experiments, we utilize the FLAIR framework and train all the neural models for 100 epochs with the batch size of 4 using AdamW (Loshchilov and Hutter, 2019) optimizer with a very small initial learning rate of $5e^{-5}$ and a stopping criterion as mentioned in Conneau et al. (2020). We use Google Colab PRO plus to carry out all the experiments.

Models	F1-Score	MCC
Random	0.0166	0.1089
BERT	0.4738	0.7405
SciBERT	0.5595	0.8016
astroBERT	0.5781	0.8104
DEAL_1	0.8168	0.9053
DEAL_2	0.8261	0.9085
Astro-mT5	**0.8364**	**0.9129**

Table 3: Validation results

[6] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[7] https://huggingface.co/xlm-roberta-large

[8] https://huggingface.co/google/mt5-large

Models	F1-Score	MCC
DEAL_1	0.7881	0.8874
DEAL_2	0.7977	0.8933
Astro-mT5	**0.8056**	**0.8954**

Table 4: Test results

4.3 Results

In our experiments, we adopt F1-Score and Matthews correlation coefficient (MCC score) as the required evaluation metrics for the given entity extraction task. We compare our results produced on the validation dataset with the previous baselines released by the DEAL SharedTask team. From Table 3, we can see that our model, **Astro-mT5**, outperforms all the baselines in terms of both F1-Score and MCC score on the validation dataset. Table 4[9] shows that our model also achieves SOTA results in terms of F1-Score against the test dataset released by the SharedTask team.

5 Conclusion

This study discusses a transformer-based deep neural architecture to identify named entities from an astrophysics literature dataset provided by the DEAL SharedTask team. Our model, **Astro-mT5**, has achieved F1-score of 80.58% and MCC of 89.54% on the test data, which remarkably outperforms previously reported models and all other competing models submitted in the DEAL SharedTask. Our future work will include more research on fine-grained NER and boundary detection in context of astrophysics to support a wide range of practical applications. We can plan to enhance our framework by introducing fine-grained NER in place of coarse-grained NER to handle a named entity with various types. We can also investigate data-driven factored modeling approaches to handle the class imbalancing problem.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

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.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jason P. C. Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *CoRR*, abs/1511.08308.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

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*

[9] It is notable that we cannot compare our results with astroBERT model on the test dataset due to unavailability of requried source code.

on Information Extraction from Scientific Publications, Taipei, Taiwan. Association for Computational Linguistics.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S Grant, Donna M Thompson, Roman Chyla, Stephen McDonald, et al. 2021. Building astrobert, a language model for astronomy & astrophysics. *arXiv preprint arXiv:2112.00590*.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 707–714, Online. Association for Computational Linguistics.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Ishani Mondal, Yufang Hou, and Charles Jochim. 2021. End-to-end construction of NLP knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1885–1895, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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. *CoRR*, abs/1910.10683.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2670–2680, Copenhagen, Denmark. Association for Computational Linguistics.

Liwen Wang, Rumei Li, Yang Yan, Yuanmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu. 2022. Instructionner: A multi-task instruction-based generative framework for few-shot ner. *arXiv preprint arXiv:2203.03903*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.