A Majority Voting Strategy of a SciBERT-based Ensemble Models for Detecting Entities in the Astrophysics Literature (Shared Task)

Atilla Kaan Alkan[*,†], **Cyril Grouin**[*], **Fabian Schüssler**[†], **Pierre Zweigenbaum**[*]
[*]Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, 91405, Orsay, France
[†]IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France
{atilla.alkan, cyril.grouin, pierre.zweigenbaum}@lisn.upsaclay.fr
fabian.schussler@cea.fr

Abstract

Detecting Entities in the Astrophysics Literature (DEAL) is a proposed shared task in the scope of the first Workshop on Information Extraction from Scientific Publications (WIESP) at AACL-IJCNLP 2022. It aims to propose systems identifying astrophysical named entities. This article presents our system based on a majority voting strategy of an ensemble composed of 32 SciBERT models. The system we propose is ranked second and outperforms the baseline provided by the organisers by achieving an F1 score of 0.7993 and a Matthews Correlation Coefficient (MCC) score of 0.8978 in the testing phase.

1 Introduction

Astronomy and astrophysics consist of observing and studying various cosmic phenomena such as tidal disruption events, gamma-ray bursts, and many other messengers such as neutrinos and gravitational waves (Neronov, 2019; Abbott et al., 2016). Missions and observations performed by astronomical facilities worldwide significantly increase the number of astrophysics papers. Most published papers are freely available and accessible through the Astrophysics Data System (ADS[1]), where researchers can search and access more than 15 million records covering astronomy, astrophysics, and general physics publications. However, some domain keywords can be easily confused when searching for articles in the literature. For instance, "*Planck*" can refer to the person, the mission, the constant, or several institutions. One approach for this word sense disambiguation problem would be automatically recognised entities. Named Entity Recognition (NER) consists of recognising mentions of entities from text belonging to predefined semantic types: person, location or organisation (Yadav and Bethard, 2018). It is, therefore, an essential technique to extract relevant information from unstructured human-written data.

Detecting Entities in the Astrophysics Literature (DEAL) is a shared task that tackles the challenge of developing a system that identifies named entities in the astrophysics literature (Grezes et al., 2022). The shared task was organised in two stages: validation and test. Evaluation metrics used were both the CoNLL-2000 shared task seqeval[2] F1-Score at the entity level and scikit-learn's Matthews correlation coefficient (MCC[3]) method at the token level. Organisers provided the NER system's baseline (see Table 3 in Appendix) using astroBERT (Grèzes et al., 2021), a deep contextual language model pre-trained on 395 499 publications (3 819 322 591 tokens, 16GB on disk) from the ADS database. The model astroBERT is not available yet, but preliminary results are exposed in the companion paper.

As part of this shared task, we used and explored an ensemble of contextual Pre-Trained Language Models (PLTMs) for NER purposes.

The paper is organised as follows: Section 2 briefly presents existing methods and approaches for named entity recognition in astrophysics and other scientific domain. Section 3 provides information about the corpus. Section 4 describes our system as well as the experimental setup. Section 5 presents our results.

2 Strategies for Entities Detection

2.1 State-of-the-Art Methods

The use of neural networks constitutes the current state-of-the-art in many tasks of NLP, including NER. Indeed, for a few years, word embeddings and the combination of two algorithms: bi-

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

[2] https://github.com/chakki-works/seqeval

[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html

Proceedings of the 1st Workshop on Information Extraction from Scientific Publications, pages 145–150
Nov 20, 2022.

directional LSTM and Conditional Random Fields (CRF), have been widely used for sequence tagging (Huang et al., 2015). The use of PLTMs (Devlin et al., 2019), and their domain-adapted version such as SciBERT for scientific literature (Beltagy et al., 2019), or BioBERT for the biomedical field (Lee et al., 2019) give state-of-the-art results on NER tasks. Some studies in the biomedical domain have shown that combining multiple PLTMs instead of a single prediction system help to increase performances on NER (Schneider et al., 2022; Dang et al., 2020).

2.2 What About Astrophysics?

Becker et al. (2005); Hachey et al. (2005) built the Astronomy Bootstrapping Corpus (ABC) composed of 209 abstracts of astronomical papers extracted from the ADS. This study explored an active learning approach to detect relevant features and reduce annotation costs for NER using a conditional Markov model tagger (Finkel et al., 2004).

Murphy et al. (2006) built a larger corpus than the ABC for named entities. The annotated corpus consists of 7840 sentences. Similarly, the study investigates the features improving the performances of a NER system based on an adaptation of a Maximum Entropy tagger (Curran and Clark, 2003).

NER studies are limited in astrophysics, and the explored approaches are feature-based only. Since methods presented in the previous section (2.1) have been successfully applied to other specific domains, such as the biomedical one, we were confident that their application to the astrophysics domain would be successful. That is why we explored a method based on an ensemble of PLTMs for NER purposes as part of this shared task.

3 The Corpus

The shared task corpus comprises full-text fragments and acknowledgements sections extracted from ADS papers. Three sets of corpus were accessible for participants[4]: training, development and testing sets. Some statistics of the corpora are provided in Table 1.

The annotation guide comprises 31 named entities and covers the entities of interest, such as astronomical facilities, celestial objects, coordinates, formulae or observational techniques. Detailed tags list is presented in Table 5 (Appendix).

[4]Data are accessible for participants only. We do not know how organisers will make the collection publicly available.

Corpus	Docs	Tokens
Train	1753	573 132
Validation	1366	447 366
Test	2505	794 739

Table 1: Corpus statistics.

For the shared task, only labels of the training corpus were provided. Figure 1 shows entities' distribution in the training corpus. The training corpus comprises full-text fragments (blue) and acknowledgements sections (orange) of approximately equal size. Most frequent categories are `Citation`, `Organization`, `Grant` or `Person`, but classes' distribution within the type of document (acknowledgments vs. full-text fragments) is not similar.

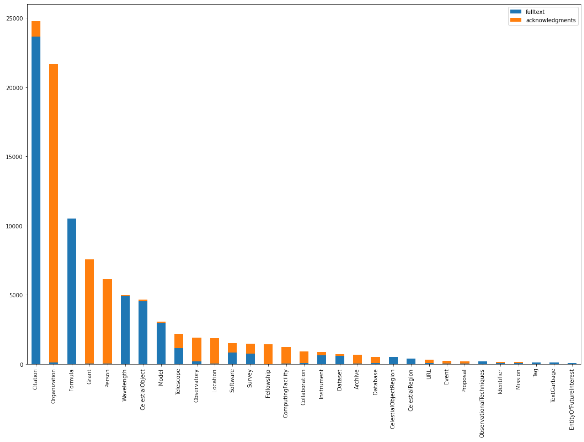

Figure 1: Entities' distribution in the training corpus. In blue are full-text fragments, and in orange are acknowledgements sections.

4 System Description

4.1 The SciBERT-cased Model

We did not apply text preprocessing to the original tokens provided by the organisers. Since some entities, such as astronomical facilities, organisations, and people's names, are proper names and therefore written in the upper-case letter, we decided to opt for the PyTorch HuggingFace's `scibert_scivocab_cased` version of SciBERT model (Beltagy et al., 2019). We assumed that preserving the type case would help the system distinguish these specific entities from standard terms. A first experiment demonstrated our assumption : the SciBERT's cased version performed better than the uncased by increasing the F1-score from 0.797 to 0.801 on the official validation set.

4.2 Setup

Internal Training and Validation Data Since we were limited to 15 daily submissions (and 100 in total) for the validation phase, we decided to create our internal validation set by splitting the original training set and conducting several experiments. Thus, our internal training set consists of 1653 annotated documents (542 550 tokens), and the internal development set comprises 100 documents (30 582 tokens).

Entities Filtering Among the defined categories, two were difficult to interpret (`TextGarbage` and `EntityOfFutureInterest`). Moreover, their low distribution in the training corpus did not make the system efficient in predicting these classes. These two reasons led us to remove them from the fine-tuning phase. Deleting these classes did not impact the overall performance since the evaluation metric was based on the micro F1-score.

Sliding Window for Long Sequences We used `BertTokenizerFast`, one of BERT's tokenizers. During the fine-tuning stage, Transformer-based models segment original tokens into subwords (or word pieces), extending thus an original sequence of N tokens into a sequence of length $N + n_{subwords}$, where $n_{subwords}$ is the number of sub-words generated by the tokenizer. This extension can exceed the size of 512, the limit sequence length that a Transformer-based model can handle. The standard way to deal with this is to apply a sliding window across the input sequence, where each window contains a passage of tokens that fit in the model's context.

4.3 Hyper-Parameters Tuning

When we started our experiments, we wanted to know the optimal combination of hyper-parameters. To do so, we proceeded to a grid search by varying two hyper-parameters: the learning rate α ($[1.10^{-5}, 2.10^{-5}, 5.10^{-5}]$) and the training batch size ($[4, 8, 16]$), representing a total of nine combinations. In order to ensure reliable results regarding the impact of hyper-parameters, each combination of hyper-parameters was used five times with five different seeds randomly chosen ($[0, 123, 762, 5000, 6822]$). We fine-tuned all models on 15 epochs using our internal training corpus and evaluated them on the internal validation set at each epoch. On average, one epoch lasts approximately 170 seconds. The ranking of the nine combinations is in Table 4 (appendix).

4.4 Ensemble Strategy

In our study, we wanted to test the influence of an ensemble approach composed of several NER classifiers. Therefore, we conducted experiments comparing the performance of a single system to an ensemble of multiple systems. We used the different models fine-tuned during the grid search to design our ensemble. We wonder two main questions:

- Which different models should we use, and how many models should be included in the system?
- What method should we use to combine the predictions of the different models in our ensemble?

Regarding the first question, we first rank the combinations of the models by performances according to their hyper-parameters during the grid search stage (Table 4, appendix). Then, we proceeded by adding models progressively to the ensemble.

Regarding the second question, related studies showed that there are mainly two approaches: the first consists of a soft strategy, where each model returns its predicted probabilities, and the class label is obtained by applying the argmax function to the sum of all probabilities (Schneider et al., 2022). The second is a majority voting strategy where the system selects the majority class of the class labels predicted by each classifier (Dang et al., 2020). We opted for the majority voting strategy.

5 Results on Official Sets

The official validation and test corpora results (Table 2) show that an ensemble composed of classifiers leads to a higher F1 score.

To determine the number of models to include in our ensemble, we progressively formed an ensemble consisting of the five models of the first performant combination (C2), then added the five models of the second performant combination (C5) and so on. We notice that the performance decreases beyond a certain number of models. Our ensemble comprises the first six combinations that gave the best results during the grid search. This represents 30 models (6 combinations $*$ 5 models / combination). A last submission in the validation phase

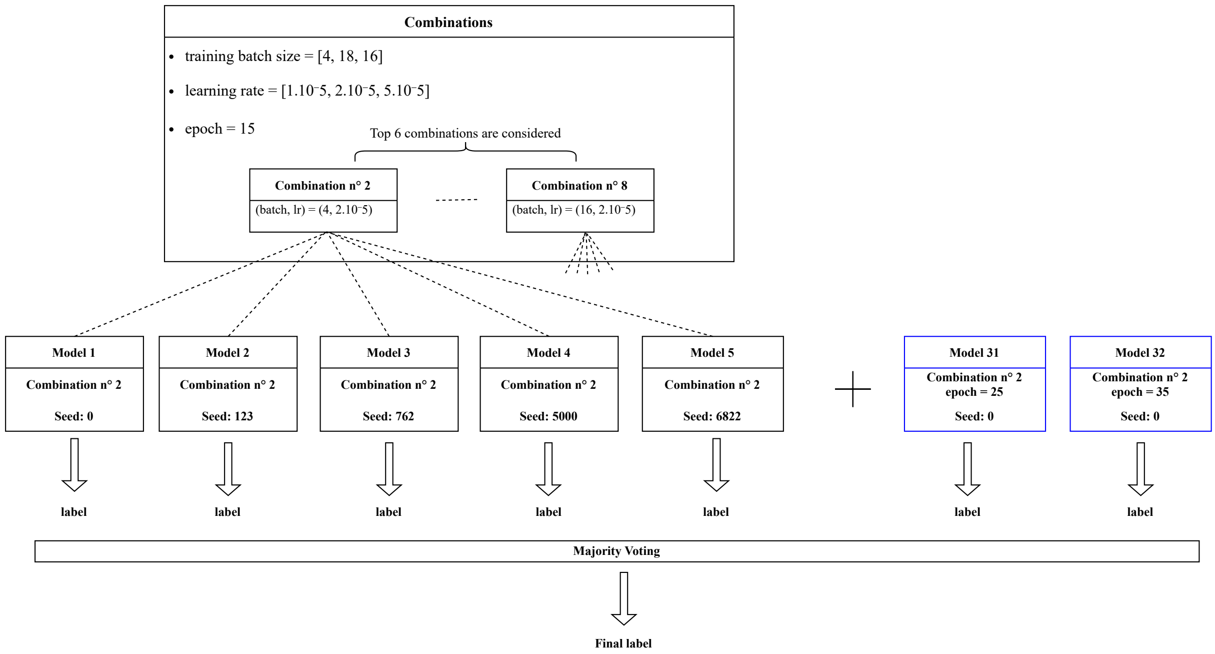

Figure 2: Final architecture of our NER ensemble based on a majority voting strategy.

Ensemble	**Validation**					**Test**				
	P	**R**	**F1**	**MCC**	**s**	**P**	**R**	**F1**	**MCC**	**s**
Single system	0.7751	0.8284	0.8009	0.9025	4	0.7990	0.7957	0.7973	0.8968	1
$\sum_{i=1}^{6} S_i$	0.8140	0.8366	0.8251	0.9132	17	0.8008	0.7966	0.7988	0.8974	2
$\sum_{i=1}^{6} S_i$ + 2 models	0.8145	0.8383	**0.8262**	0.9140	24	0.8013	0.7972	**0.7993**	0.8978	4

Table 2: Results on official validation and test sets with the corresponding submission number (s) on the Codalab platform. Metrics used are Precision (P), Recall (R), F1-score and MCC.

(s=24) showed us that adding two additional models from combination n°2 (fine-tuned on a few additional epochs) increases the F1 score. Ultimately, our ensemble consists of 32 models. Figure 2 illustrates our architecture.

6 Conclusion

This shared task aimed to tackle the challenge of detecting entities in the astrophysics literature by proposing a NER system. We exposed in this paper our approach, which first consists of identifying the different hyper-parameters combination giving the highest F1-score. To do so, we proceeded to do a grid search on our internal training and validation sets. In the second stage, we built an ensemble of classifiers based on the top 6 combinations identified during the grid search. Our submissions on the official validation and test sets show that adopting a majority voting strategy of an ensemble of SciBERT-based classifiers gives better results than a single model approach. Finally, we ranked second, achieving an F1 score of 0.7993 and an MCC coefficient of 0.8978 using an ensemble of 32 SciBERT models.

References

B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. 2016. Observation of gravitational waves from a binary black hole merger. *Physical Review Letters*, 116(6).

Markus Becker, Ben Hachey, Beatrice Alex, and Claire Grover. 2005. Optimising selective sampling for bootstrapping named entity recognition. In *In Proceedings of the ICML Workshop on Learning with Multiple Views*, pages 5–11.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*. Association for Computational Linguistics.

James Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the Seventh Conference on Natu-*

ral Language Learning at HLT-NAACL 2003, pages 164–167.

Huong Dang, Kahyun Lee, Sam Henry, and Özlem Uzuner. 2020. Ensemble BERT for classifying medication-mentioning tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 37–41, Barcelona, Spain (Online). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jenny Finkel, Shipra Dingare, Huy Nguyen, Malvina Nissim, Christopher Manning, and Gail Sinclair. 2004. Exploiting context for biomedical entity recognition: From syntax to the web. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 91–94, Geneva, Switzerland. COLING.

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.

Félix Grèzes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Golnaz Shapurian, Edwin A. 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. *CoRR*, abs/2112.00590.

Ben Hachey, Beatrice Alex, and Markus Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 144–151, Ann Arbor, Michigan. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *ArXiv*, abs/1508.01991.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Tara Murphy, Tara McIntosh, and James R. Curran. 2006. Named entity recognition for astronomy literature. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 59–66, Sydney, Australia.

Andrii Neronov. 2019. Introduction to multi-messenger astronomy. *Journal of Physics: Conference Series*, 1263(1):012001.

Elisa Schneider, Renzo M. Rivera-Zavala, Paloma Martinez, Claudia Moro, and Emerson Paraiso. 2022. UC3M-PUCPR at SemEval-2022 task 11: An ensemble method of transformer-based models for complex named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1448–1456, Seattle, United States. Association for Computational Linguistics.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A Appendix

Model	P	R	F1	MCC
random	0.119	0.0274	0.0166	0.1089
BERT	0.4779	0.4697	0.4738	0.7405
SciBERT	0.5457	0.5741	0.5595	0.8016
astroBERT	**0.5511**	**0.6080**	**0.5781**	**0.8104**

Table 3: Baseline scores for the DEAL shared task. Metrics used are Precision (P), Recall (R), F1-score and MCC.

Rank	Comb.	Designation	Hyp.-params.
1	C2	S1	$(4, 2.10^5)$
2	C5	S2	$(8, 2.10^5)$
3	C9	S3	$(16, 5.10^5)$
4	C6	S4	$(5, 5.10^5)$
5	C1	S5	$(4, 1.10^5)$
6	C8	S6	$(16, 2.10^5)$
7	C3	S7	$(4, 5.10^5)$
8	C4	S8	$(8, 1.10^5)$
9	C7	S9	$(16, 1.10^5)$

Table 4: Grid search: ranking of the combination (Comb.) giving the best results. After having ranked the different combinations, we denote by S_i the set of five models (having the same hyper-parameters) ranked in position i

Category	Definition	Example
Person	A named person or their initials	Andrea M. Ghez, Ghez A.
Organization	A named organization that is not an observatory.	NASA, University of Toledo
Location	A named location on Earth.	Canada
Observatory	A, often similarly located, group of telescopes.	Keck Observatory, Fermi
Telescope	A "bucket" to catch light.	Hubble Space Telescope, Discovery Channel Telescope
Instrument	A device, often, but not always, placed on a telescope, to make a measurement.	Infrared Array Camera, NIRCam
Survey	An organized search of the sky often dedicated to large scale science projects.	2MASS, SDSS
Mission	A spacecraft that is not a telescope or observatory that carries multiple instruments	WIND
CelestialObject	A named object in the sky	ONC, Andromeda galaxy
CelestialRegion	A defined region projected onto the sky, or celestial coordinates.	GOODS field, l=2, b=15
CelestialObjectRegion	Named area on/in a celestial body.	Inner galaxy
Wavelength	Portion of the electromagnetic spectrum	656.46 nm, H-alpha
ObservationalTechniques	Methods/technqiues for observation	Spectroscopic, helioseismic
Model	Mathematical/Physical model	Gaussian, Keplerian
Software	Software, IT tool	NuSTAR, healpy, numpy
ComputingFacility	Server, cluster for computation	Supercomputer, GPU
Dataset	Astronomical catalogues	3FGL catalog
Database	A curated set of data	Simbad database
Archive	A curated collection of the literature or data.	NASA ADS, MAST
Identifier	A unique identifier for data, images, etc.	ALMA 123.12345
Citation	A reference to previous work in the literature.	Allen et al. 2012
Collaboration	Name of collaboration	Fermi LAT Collaboration
Event	A conference, workshop or other event that often brings scientests together.	Protostars and Planets VI
Grant	An allocation of money and/or time for a research project.	grant No. 12345, ADAP grant 12345
Fellowship	A grant focused towards students and/or early career researchers.	Hubble Fellowship
Formula	Mathematical formula or equations.	$F = Gm1m2/r^2, z = 2.3$
Tag	A HTML tag.	<bold>
TextGarbage	Incorrect text, often multiple punctuation marks with no inner text.	,,,
EntityOfFutureInterest	A general catch all for things that may be worth thinking about in the future.	Earth-like, Solar-like
URL	A link to a website.	$https://www.astropy.org/$

Table 5: Classification of the named entities in the annotation guideline. The HuggingFace repository containing the annotated data and the annotation guide is only accessible to participants of the shared task. Thus, we have reproduced the same list of named entities with their definition.