TDAC, the First Time-Domain Astrophysics Corpus: Analysis and First Experiments on Named Entity Recognition

Atilla Kaan Alkan, Cyril Grouin, Fabian Schüssler, Pierre Zweigenbaum

TDAC, the First Time-Domain Astrophysics Corpus: Analysis and First Experiments on Named Entity Recognition

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

The increased interest in time-domain astronomy over the last decades has resulted in a substantial increase in observation report publication leading to a saturation of how astrophysicists read, analyze and classify information. Due to the short life span of the detected astronomical events, information related to the characterization of new phenomena has to be communicated and analyzed very rapidly to allow other observatories to react and conduct their follow-up observations. This paper introduces TDAC: a Time-Domain Astrophysics Corpus. TDAC is the first corpus based on astrophysical observation reports. We also present the NLP experiments we made for named entity recognition based on annotations we made and annotations from the WIESP DEAL shared task.

1 Introduction

Time-domain astrophysics consists in observing and studying transient cosmic phenomena, *i.e.* unpredictable, short-lived, and the most violent phenomena occurring in the Universe, such as supernovae explosions or gamma-ray bursts (GRBs), which are highly energetic explosions lasting from milliseconds to a few hours or days only (Neronov, 2019). The short life span of these events requires a rapid sharing, analysis and synthesis of the information disseminated in observation reports. However, the increased interest in time-domain astronomy has led to a significant increase in observation reports, leading to a saturation of how astrophysicists analyze and classify information in observation reports. As the current manual reading and analyzing of available information is approaching saturation, new ways of handling information are necessary.

One of the most promising approaches is to build Natural Language Processing (NLP) methods that tackle the challenges of extracting and summarizing information on observation reports by detecting, for example, named entities. Named Entity Recognition (NER) can identify and extract information about an astrophysical object, such as the date of detection, its coordinates in the Universe, and numerous information, such as intensity and magnitude, to let astrophysicists know if they can trigger a follow-up observation. To train and evaluate an NER system, a corpus must first be created and annotated.

This paper presents TDAC: a Time-Domain Astrophysics Corpus for NLP, based on observation reports. To our knowledge, no existing resources and studies so far are based on time-domain astrophysics observation reports, and therefore there are no studies characterising the discourse used in astrophysics. Our objective is twofold: The first objective of our study, with the creation of this corpus, is to highlight differences between astrophysics corpora. What are their properties, and are they all the same? We provide some elements characterizing and revealing the specificity of the formulations used in astrophysics by conducting a corpus analysis (Section 4). Secondly, we started building an NER system for the domain. Section 5 presents our annotations and the first NER experiments we conducted on a sub-corpus of TDAC (75 documents). The annotated section of TDAC is the first annotated and publicly available[1] corpus based on observation reports for named entity recognition in time-domain astrophysics.

2 Research and Language Resources in Astrophysics

The vast majority of the limited research performed so far in NLP for astrophysics studies papers from the Astrophysics Data System (ADS[2]). The ADS is a database for researchers in astronomy with more than 15 million records covering publications in astronomy, astrophysics, and general physics.

[1] https://github.com/AtillaKaanAlkan/TDAC

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

Proceedings of the 1st Workshop on Information Extraction from Scientific Publications, pages 131–139

Abstracts and full text of astronomy paper publications are indexed and searchable through ADS, making it a rich exploitable platform for creating NLP resources.

2.1 The Astronomy Bootstrapping Corpus

The Astronomy Bootstrapping Corpus (ABC) (Becker et al., 2005; Hachey et al., 2005) is one of the unique existing annotated corpora for astrophysical Named Entity Recognition (NER). ABC consists of 209 abstracts of astronomical papers extracted from the ADS. The built corpus aimed to explore an active learning approach to reduce annotation costs for a NER task by defining four astrophysical named entities: `instrument_name`, `source_name`, `source_type` and `spectral_feature`, with respectively 136, 111, 499 and 321 instances. To our knowledge, the corpus is not available.

2.2 The Astro Corpus

Murphy et al. (2006) built a larger corpus than the ABC for named entities detection by downloading all the astronomical journal articles and conference papers (52 658 documents) from the astrophysics section (astro-ph) of arXiv. The annotated corpus consists of 7840 sentences (approximately 200 000 words) with an average of 26.1 tokens per sentence. There are 43 astrophysical named entities, including celestial objects, telescope names and categories related to astrophysical sources' properties. To our knowledge, this corpus is not available either.

2.3 The DEAL Shared Task Corpus

The Detecting Entities in the Astrophysics Literature (DEAL) shared task[3] consists of developping a system that identifies named entities in the astrophysics literature (Grezes et al., 2022). The organisers provided a baseline NER system 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 astroBERT model is not available yet, but preliminary results (F1-score of 0.902 on an NER task) are exposed in the above-cited paper. The DEAL corpus comprises full-text fragments and acknowledgements sections extracted from ADS papers for the shared task. The corpus was split into train, development and test sets, with 1753, 1366 and 2505 documents, respectively. During the shared task, only the labels for the training set were provided. We participated in the shared task and had access to the entire annotated collection[4] (train+development+test) at the end of the shared task. It is, therefore, the only annotated corpus we have for comparison with our TDAC corpus. We provide more detailed statistics on the DEAL corpus in the rest of the paper.

2.4 Other Studies

Information Retrieval and Recommendation System Kerzendorf (2019) downloaded astrophysics papers from the arXiv Bulk Data Access to build a corpus (201.997 articles). Their study aims to develop a robust text-based similarity tool to recommend articles given a reference input paper. Mukund et al. (2018) built and deployed another information retrieval and recommendation system, "Hey LIGO", an open access NLP-based web application for LIGO and VIRGO observatories (both aiming to detect gravitational waves). Documents used are extracted from the open source logbook data from both observatories. Therefore, to our knowledge, this is the only study not based on astrophysics papers. Data have been recorded since 2010, and the logbook consists of 83.911 entries, and an automatic check for new data entries is periodically done to update the models regularly.

Anaphora Resolution Kim and Webber (2006) used astrophysics articles from the Monthly Notices of the Royal Astronomical Society (MNRAS) to constitute a small corpus (it consists of more than a hundred articles) for anaphora resolution. To conclude this literature review, most NLP resources for astrophysics are mainly created and exploited using scientific papers. This paper presents TDAC, the first annotated corpus based on observation reports for named entity recognition in time-domain astrophysics.

3 Material for the TDAC Corpus

3.1 The resource platforms used

Reports are written and published on mainly three platforms by an extensive network of professional observers worldwide (astronomical observatories and satellites) and are accessible in open source to the entire research community. In this study, we

[3] https://ui.adsabs.harvard.edu/WIESP/2022/SharedTasks

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

use these platforms to have a good coverage of the domain.

The Gamma-Ray Burst Coordinates Network The GCN[5] platform is dedicated mainly to the gamma-ray bursts astrophysicists community, where observers report their observations and analysis of GRBs in the form of "GCN Circulars" (Barthelmy et al., 1995).

The Astronomer's Telegram This system is a communication channel[6] that allows instantaneously sharing and reporting information to the astrophysicists' community in the form of astronomer's telegrams or "ATel" (Rutledge, 1998). Observers report discoveries regarding a large variety of astronomical sources with no restrictions on the type of discoveries (black holes, blazars, neutron stars etc.).

The Transient Name Server The TNS[7] is mainly a dedicated platform for the astronomers' community interested in confirmed supernovae candidates. Astrophysicists report their observations in the form of "AstroNotes" (Gal-Yam, 2021).

3.2 Collecting the raw corpus

An archive with the complete set of published GCN circulars in text files is available on the GCN website. Thus, to collect raw data and build up our corpus, we downloaded it. However, unlike GCN circulars, there is no direct way to bulk download all past ATel and AstroNotes. Therefore, we set up a Python script using the BeautifulSoup package to perform an automated extraction of the HTML code of all reports published from 1997 to 2021 and parsed the content into a text file. Figure 1 shows the evolution of reports published annually.

The increase in published reports is due to the number of observations monitored by various observers, particularly with the launch of the Swift telescope in 2004, leading to a significant increase in GCN circulars regarding GRB detection. However, we note a slight decrease in the number of ATel telegrams since 2015. A migration of publications to the TNS platform could be the reason for the decrease in the number of ATel published per year. Another explanation for this decrease could be that the types of objects processed in the telegrams have been less observed in recent years.

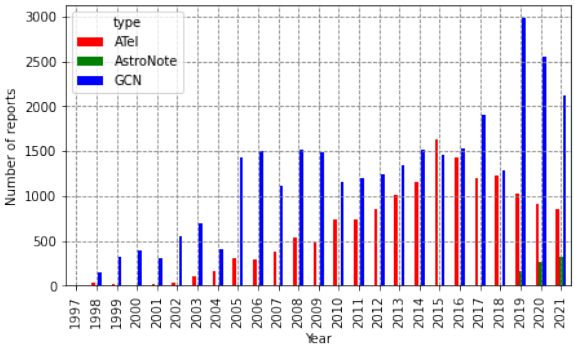

Figure 1: Number of published reports from 1997 to 2021 (GCN circulars in blue, ATel in red, AstroNotes in green)

4 Corpus Analysis

4.1 Statistics

As descibed in Table 1, within the TDAC corpus, AstroNotes are the least numerous, as the platform is more recent than GCN and ATel platforms. It explains the significant difference in the total number of tokens for each type of document. However, although AstroNotes are less numerous, they have the highest lexical diversity. GCN circulars and ATels seem to be quite similar in terms of vocabulary richness. We notice that the DEAL corpus has the least lexical diversity. Perhaps the documents in this corpus are all from the same theme, or all deal with the same types of astrophysical phenomena. Among observation reports, GCN circulars are the longest.

Corpus	# Doc	# Tokens	Lex.	Len.
ATel	15 108	3 250 292	0.068	260
GCN	31 964	7 283 252	0.065	319
AstroNotes	741	165 303	0.076	277
DEAL	5624	1815237	0.057	322

Table 1: Astrophysics corpus statistics comparison (number of focuments, number of tokens, lexical diversity, and average length)

4.2 Most Frequent Word N-grams

Before counting the most frequent unigrams and bigrams characterising the observation reports, we proceeded to some text preprocessing[8]. Results in

[5] https://gcn.gsfc.nasa.gov/gcn3_archive.html

[6] https://astronomerstelegram.org/

[7] https://www.wis-tns.org/astronotes/

[8] We removed stopwords, normalised all digits/numerical values, and lemmatised each word.

TDAC	Unigrams	Bigrams
ATel	num_val, *source*, *observation*, *atel*, *spectrum*, *flux*, *telescope*, *image*, *x-ray*, *transient*	*atel* link, *dec* num_val, *apj* num_val, *reference image*, num_val *mcrab*, *unfiltered magnitude*, *host galaxy*, *autodetection system*, *redshift* num_val
GCN	num_val, *grb*, *gcn*, *observation*, *report*, *burst*, *team*, *kev*, *swift*	*grb* num_val, *gcn* num_val, num_val gmt, num_val kev, *light curve*, *upper limit*, *fermi gbm*, *swift-xrt team*, *grb observation*, *photon index*
AstroNotes	*transient*, *atlas*, *survey*, *object*, *related*, *report*, *telescope*, *observation*, *classification*, *search*, *supernova*, *system*, *galaxy*	*related files*, num_val *mpc*, *grant* num_val, *near earth*, *transient name*, *iau transient*, num_val *arcsec*, *queens university*, *zwicky transient*, *follow-up observation*

Table 2: Most frequent unigrams and bigrams in the TDAC corpus

Table 2 show that more digits and numerical values (num_val token) exist in the GCNs and ATels compared to the AstroNotes. We note and identify different astronomical facilities and objects according to the report's type, such as *swift* and *fermi* telescopes in the GCNs, or even *atlas* and *zwicky transient* facility in the AstroNotes. We note different energy ranges and measurement units (*kev*, *mcrab*, *arcsec*), or different wavelengths (*x-ray*) depending on the type of report. Astrophysicists we are collaborating with confirmed our conclusion: in astrophysics, each community uses dedicated platforms according to the discoveries that interest them. Finally, the main thing we notice when analyzing the bigrams is the strong interconnection inside ATel and GCN circulars. Indeed, there are many explicit references between the observation reports (*gcn* num_val and *atel* link) regarding detected events. Since the information concerning an astrophysical event is disseminated across several linked documents, it is essential to gather all the documents and aggregate them by the event.

4.3 Syntactic Analysis

Campbell and Johnson (2001) showed the usefulness of the Pointwise Mutual Information (PMI) and the chi-square χ^2 distance to compare syntactic complexity between corpora. Thus, we decided to compute these two metrics to characterise the discourse used in astrophysics. We computed the positive PMI (see equation 1) on parts-of-speech (POS) bigrams between two corpora: our TDAC corpus composed of observation reports and the DEAL challenge corpus.

$$PMI(x, y) = \log_2\left(\frac{P(xy)}{P(x) * P(y)}\right) \quad (1)$$

The mutual information allows highlighting the proximity between two corpora. We also compared the frequency of occurrence of single POS and POS bigrams between corpora using the χ^2 metric (see equation 2).

$$\chi^2 = \sum\left(\frac{Observed - Expected}{Expected}\right)^2 \quad (2)$$

We used SciSpacy (Neumann et al., 2019) for POS tagging after conducting performance tests[9] of POS labelling, and obtaining better performance than NLTK, TreeTagger, Spacy and Genia tools.

4.3.1 Pointwise Mutual Information of POS

We divided each corpus into ten sections of the same size in order to ensure stability of results. We only considered the positive mutual information and then set the negative values to zero. Table 3 reports the average postitive PMI for POS bigrams.

These results seem to point to a less complicated syntactic structure in the DEAL corpus compared to the TDAC one. Indeed, the average PMI value of the DEAL corpus is slightly higher than the average PMI score of the TDAC corpus. When looking inside the TDAC corpus, we notice that compared to ATels et GCNs, the occurrence of POS bigrams

[9]To compare tagging performances, we manually annotated 20 documents from the TDAC corpus and compared performances on POS tagging of 5 different tools to determine the appropriate one for astrophysics texts.

Corpus	# token/section	Avg PMI
TDAC	1 500 000	0.469 (0.028)
– ATel	450 000	0.554 (0.050)
– GCN	960 000	0.524 (0.009)
– AstroNotes	21 000	0.961 (0.044)
DEAL	210 000	0.622 (0.026)

Table 3: Average Positive PMI for POS bigrams (standard deviation of mean in parentheses)

in AstroNotes seems more dependent than those in ATels and GCNs, as seen by the higher score in the positive PMI. The syntactic structure seems to be less complicated in AstroNotes.

4.3.2 Frequency Distributions of POS

We computed the chi-square metric to calculate the distances between each corpora. The chi-square distances for single and POS bigrams comparisons are reported in Table 4. POS and POS bigrams

Corpus	χ^2 POS	χ^2 POS bigram
ATel-GCN	1 075 610.83	1 234 413.99
ATel-AstroNotes	1 594 932.63	1 597 152.56
GCN-AstroNotes	4 017 655.62	4 012 353.21
TDAC-DEAL	3 986 047.13	4 053 795.68

Table 4: χ^2 distance comparison for single POS and POS bigram frequencies.

distributions are relatively different between the TDAC and DEAL corpus, which explains these large χ^2 values between the two corpora. Within the TDAC corpus, we can see a high distance between GCN circulars and the AstroNotes, whereas it is less marked between the ATel and AstroNotes. These first results regarding syntactic analysis show a diversity between the corpora used, but further analysis is needed to qualify these differences.

5 Named Entity Recognition

5.1 Astrophysical Named Entities

We used the same categories defined in the DEAL shared task. This annotation guide comprises 31 named entities and covers the entities of interest, such as astronomical facilities, celestial objects, coordinates, formulae or observational techniques contained in observation reports. Detailed tags list is presented in Table 8 in Appendix. Figure 2 shows the normalised distribution of annotated named entities on the TDAC and DEAL corpora for comparison purposes.

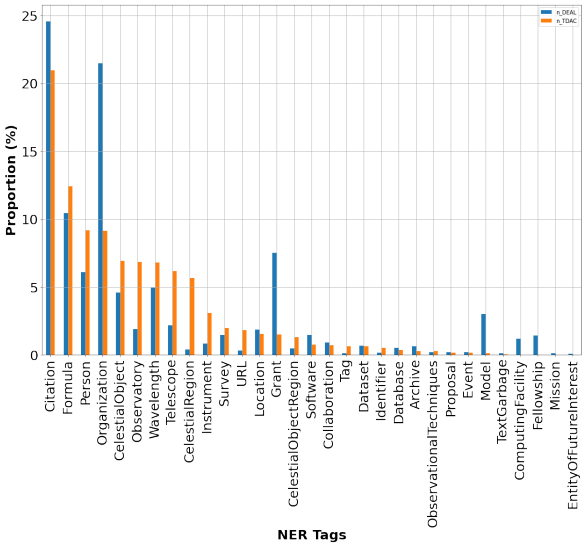

Figure 2: Normalised distribution of named entities in the TDAC (orange) and DEAL (blue) corpus.

Classes' distribution within the two corpora is not similar. Indeed, in the TDAC corpus, the most frequent categories *e.g.* `Formula`, `CelestialObject`, `Observatory`, or `CelestialRegion`. These are particular categories in the astrophysics domain. Most of these specific classes are less present in the DEAL corpus, in which we find mainly the classes of types: `Citation`, `Organization`, `Grant` or `Person`, which seems to be more generic named entity categories.

5.2 Annotation Procedure

The reports used to build the TDAC corpus for NER were randomly selected from the extracted observation reports and annotated in two stages. First, we used one of the models fine-tuned for the DEAL shared task to perform an automatic pre-annotation of 75 observation reports, followed by a manual correction stage by a PhD student with a background in astrophysics. The evaluation of the quality of the pre-annotation using the fine-tuned model corresponds to experiment 1 presented in the rest of this article (see Table 5). During the manual correction phase of the 75 documents, a double annotation was carried out on 30 documents (*i.e.* 7584 double annotated tokens) between the PhD student and a senior in NLP. The average time spent per document is about 4.5 minutes for the PhD student and 5.7 min for the NLP expert. This double annotation allowed us to calculate an inter-annotator agreement (IAA) using recall, precision,

and F1 score; metrics considered adapted for computing IAA in several studies (Grouin et al., 2011). After a first double annotation of the 30 documents between the two annotators, we obtained an overall F1 score of 0.7839. After a second pass, we reached an F1 score of 0.8490, high enough for the PhD student to continue annotating the remaining documents alone.

5.3 Experiments

The Baseline Model We used one of the models fine-tuned as part of the DEAL shared task to perform an automatic pre-annotation of the TDAC corpus. It corresponds to the PyTorch HuggingFace's `scibert_scivocab_cased` version of SciBERT model (Beltagy et al., 2019). It has been fine-tuned on the DEAL corpus that we split into train and development sets. The training set consists of 1653 annotated documents (542 550 tokens), and the development set comprises 100 documents (30 582 tokens). For the shared task, the model has been tested on 1366 documents (447 366 tokens). Fine-tuning was performed on 11 epochs, with a learning rate $\alpha = 2.10^{-5}$ and a training batch size of 4. One epoch took approximately 170 seconds. More information is provided in the corresponding system description paper (Alkan et al., 2022).

Experiment 1: Testing directly on TDAC This first experiment evaluates the baseline model fine-tuned on the DEAL corpus directly to the TDAC corpus and analyzes whether performances stay maintained when applying to another type of corpus of the same specialised domain. Thus, we evaluate the model on the 75 annotated documents.

Experiment 2: Continue Model's Fine-Tuning using TDAC We will continue the model's fine-tuning on 9 additional epochs in this second experiment using the TDAC corpus. We split the TDAC corpus into training and test sets (approximately 80%-20%), *i.e.* 59 documents for training (18 ATels, 21 GCNs and 20 AstroNotes) which represents a total of 15 374 tokens and 16 documents for evaluation (7 ATels, 4 GCNs, and 4 AstroNotes) which represents a total of 3638 tokens. Since the corpus size is still small, one epoch lasts about 6 seconds when fine-tuning on TDAC.

Experiment 3: Fine-Tuning a New Model From Scratch on TDAC For this third experiment, we fine-tuned from scratch on TDAC the `scibert_scivocab_cased` with same hyper-parameters configuration than the baseline model, *i.e.* ($epoch = 20$, $\alpha = 2.10^{-5}$, $batch = 4$). We used the same training and test sets as experiment 2.

5.4 Results

For evaluation we used both the CoNLL-2000 shared task seqeval[10] F1-Score at the entity level and scikit-learn's Matthews correlation coefficient (MCC[11]) method at the token level.

Experiment 1 For comparison purposes, we also reminded the performances of the system trained and tested on the DEAL corpus as part of the shared task. The performances of the NER system on the TDAC corpus (75 documents) are given in Table 5.

Corpus	P	R	F1	MCC
DEAL	0.7752	0.8284	0.8009	0.9025
TDAC	0.4993	0.7043	0.5843	0.7760
– ATel	0.5809	0.7325	0.6480	0.8213
– GCN	0.5236	0.7230	0.6074	0.7653
– AstroNotes	0.3952	0.6421	0.4893	0.7474

Table 5: Performance of the baseline NER system fine-tuned on DEAL (as part of the shared task) and tested on our TDAC corpus (with details by type of document). Metrics used are Precision (P), Recall (R), F1-score and MCC.

Experiment 2 Table 6 shows the performance of the baseline NER system we fine-tuned on 9 additional epochs.

Corpus	P	R	F1	MCC
TDAC	0.720	0.796	0.756	0.855
– ATel	0.667	0.703	0.684	0.854
– GCN	0.745	0.822	0.781	0.842
– AstroNotes	0.874	0.891	0.882	0.943

Table 6: Performance of the baseline NER system after fine-tuning on 9 additional epochs using our TDAC corpus (with details by type of document). Metrics used are Precision (P), Recall (R), F1-score and MCC.

Experiment 3 Table 7 shows the performance of the NER system we built and fine-tuned from scratch on the TDAC corpus.

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

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

Corpus	P	R	F1	MCC
TDAC	0.693	0.777	0.733	0.814
– ATel	0.672	0.733	0.701	0.846
– GCN	0.728	0.793	0.759	0.796
– AstroNotes	0.684	0.792	0.734	0.877

Table 7: Performance of a NER system after fine-tuning from scratch on 20 epochs using our TDAC corpus (with details by type of document). Metrics used are Precision (P), Recall (R), F1-score and MCC.

6 Discussion and Outlook

Experiment 1 is not comparable to experiments 2 and 3 because the test sample size is not the same. However, it allows us to first appreciate the baseline model's robustness by testing it on our TDAC corpus. When tested on the TDAC corpus, we noticed a considerable drop in performance (a loss of 0.2166 on the F1 score globally). The results may appear low or moderate. This could be explained by a strict evaluation (identical label and border). With experiments 2 and 3, we notice relatively similar results. Overall, the model fine-tuned from scratch performs slightly worse than the baseline model for which we continued the fine-tuning over nine additional epochs. Experiment 3 shows that the system performs better on the ATels when fine-tuning from scratch. These preliminary results on this first small annotated corpus nevertheless show that the DEAL corpus is a good starting point for building an entity detection system and can be adapted to other types of documents in the astrophysical domain. However, it is necessary to analyze whether this behaviour is confirmed on a larger scale.

While the first annotations have been made by a PhD student with a background in astrophysics in order to make a proof-of-concept, we are now experiencing new annotations made by two senior experts, one in astrophysics, the other in NLP.

Joining the two corpora (DEAL+TDAC) would be complementary because of the distribution of classes in the two corpora (Figure 2). We observe that certain classes of entities are more present in the TDAC corpus than in DEAL (*e.g.* `Formula`, `CelestialObject`, `Observatory`, or `CelestialRegion`). The TDAC corpus thus makes it possible to fill the lack of specific classes and vice versa. Therefore, joining these two corpora would thus allow for building a more efficient system for a more significant number of classes.

7 Conclusion

In this paper, we presented the TDAC corpus, composed of astrophysics textual content from three sources (ATel, GCN circulars, and AstroNotes). Our corpus has been manually annoted in named entity, based on the annotation schema used in the DEAL corpus. We also presented the experiments we made in order to make it easier the manual annotation process, using a SciBERT-based model fine-tuned on the WIESP 2022 NLP Challenge. We observed that a model trained on the DEAL corpus is not sufficient since it obtained moderate results, while a quite light fine-tuning (9 additional epochs) on our TDAC corpus allows us to improve the performances of our NER system.

In the future, we plan to enrich the corpus with morpho-syntactic annotations and relations between named entities. We estimate this corpus would be a useful resource for NLP applications in astrophysics.

Once the information extraction system we are developing is considered reliable enough, we aim to deploy them in Astro-COLIBRI, a real-time platform that evaluates alerts sent by observers regarding transient sources (Reichherzer et al., 2021). The deployment of our NLP models in Astro-COLIBRI will allow both professional and amateur astronomers to access the most relevant information disseminated through GCN circulars, ATels and AstroNotes instantaneously.

References

Atilla Kaan Alkan, Cyril Grouin, Fabian Schüssler, 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.

Scott Douglas Barthelmy, Paul S. Butterworth, Thomas L. Cline, Neil Gehrels, Gerald J. Fishman, Chryssa Kouveliotou, and Charles A. Meegan. 1995. BACODINE, the real-time BATSE gamma-ray burst coordinates distribution network. *Astrophysics and Space Science*, 231:235–238.

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.

DA Campbell and S Johnson. 2001. Comparing syntactic complexity in medical and non-medical corpora. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 90–4.

A. Gal-Yam. 2021. The TNS alert system. *Bulletin of the AAS*, 53(1). Https://baas.aas.org/pub/2021n1i423p05.

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.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA. Association for Computational Linguistics.

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.

W. E. Kerzendorf. 2019. Knowledge discovery through text-based similarity searches for astronomy literature. *Journal of Astrophysics and Astronomy*, 40:1–7.

Yunhyong Kim and Bonnie Webber. 2006. Implicit reference to citations: a study of astronomy. *ERPANET*.

Nikhil Mukund, Saurabh Thakur, Sheelu Abraham, A. K. Aniyan, Sanjit Mitra, Ninan Sajeeth Philip, Kaustubh Vaghmare, and D. P. Acharjya. 2018. An Information Retrieval and Recommendation System for Astronomical Observatories. *Astrophysical Journal Supplement*, 235(1):22.

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.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

P. Reichherzer, F. Schüssler, V. Lefranc, A. Yusafzai, A. K. Alkan, H. Ashkar, and J. Becker Tjus. 2021. Astro-COLIBRI—the COincidence LIBrary for real-time inquiry for multimessenger astrophysics. *The Astrophysical Journal Supplement Series*, 256(1):5.

Robert E. Rutledge. 1998. The Astronomer's Telegram: A Web-based Short-Notice Publication System for the Professional Astronomical Community. *Publications of the Astronomical Society of the Pacific*, 110(748):754–756.

A Appendix

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 8: 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.