Enriching a Time-Domain Astrophysics Corpus with Named Entity, Coreference, and Astrophysical Relationship Annotations

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

Enriching a Time-Domain Astrophysics Corpus with Named Entity, Coreference, and Astrophysical Relationship Annotations

Atilla Kaan Alkan[*,†], **Felix Grezes**[‡], **Cyril Grouin**[*], **Fabian Schüssler**[†], **Pierre Zweigenbaum**[*]

[*] Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, 91405, Orsay, France.
{atilla.alkan, cyril.grouin, pz}@lisn.upsaclay.fr

[†] IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France.
fabian.schussler@cea.fr

[‡] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
felix.grezes@cfa.harvard.edu

Abstract

Interest in Astrophysical Natural Language Processing (NLP) has increased recently, fueled by the development of specialized language models for information extraction. However, the scarcity of annotated resources for this domain is still a significant challenge. Most existing corpora are limited to Named Entity Recognition (NER) tasks, leaving a gap in resource diversity. To address this gap and facilitate a broader spectrum of NLP research in astrophysics, we introduce astroECR, an extension of our previously built Time-Domain Astrophysics Corpus (TDAC). Our contributions involve expanding it to cover named entities, coreferences, annotations related to astrophysical relationships, and normalizing celestial object names. We showcase practical utility through baseline models for four NLP tasks and provide the research community access to our corpus, code, and models.

Keywords: Corpus Annotation, Information Extraction, Astrophysics

1. Introduction

In recent years, the field of Natural Language Processing (NLP) has witnessed a significant surge of interest in astrophysics. The development of pretrained language models and generative language models such as astroBERT (Grezes et al., 2021) and astroLLaMa (Nguyen et al., 2023) has primarily fueled this increase. These models are used for detecting entities in the astrophysics literature (Grezes et al., 2022) and for information extraction purposes (Sotnikov and Chaikova, 2023). Despite these developments, a notable challenge persists in the availability of resources for astrophysical NLP research. Most existing corpora (Becker et al., 2005; Hachey et al., 2005; Murphy et al., 2006) within the domain are not accessible and primarily serve Named Entity Recognition (NER) tasks, leaving a significant gap in resource diversity. Among accessible corpora, we find the DEAL shared task corpus (Grezes et al., 2022), and TDAC (Alkan et al., 2022), a corpus that we previously built for detecting entities in time-domain astrophysics[1]. Both corpora share common classes and consist of named entity annotations only. However, corpora often need more comprehensive annotations for complete information extraction. To facilitate a broader spectrum of research in astrophysical NLP, our work aims to create a more complete annotated corpus encompassing named entities, coreferences, astrophysical relationships annotation, and celestial object names normalization. We aim to provide the astrophysics and NLP communities with a valuable resource that enables the development of advanced information extraction models. To do so, we extended our existing TDAC corpus to build a new one, astroECR, richer in annotations. The main contributions of this work are:

- We enriched the TDAC corpus size, expanding from 75 to 300 documents. This augmentation contains a comprehensive set of annotations covering astrophysical named entities, celestial object name normalization (entity linking), coreference annotations, and astrophysical relationships;
- We used the same named entity categories used on the TDAC corpus, and we defined five additional named entity categories;
- We demonstrate the practical utility of this corpus by conducting experiments on four distinct NLP tasks, for which we have developed baseline models. These models facilitate the automated annotation of documents, showcasing the potential of our enriched corpus;
- We make our corpus, annotation guidelines, associated code, and models accessible on GitHub[2].

[1] Branch of astrophysics that studies supernova explosions and gamma-ray bursts.

[2] https://github.com/AtillaKaanAlkan/

LREC-COLING 2024, pages 6177–6188
20-25 May, 2024.

2. Related Work

This section highlights works dedicated to creating annotated resources in astrophysics. We have categorized these existing resources into two distinct groups: first, corpora annotated for named entity detection, and second, those annotated for coreference resolution.

2.1. NER-oriented Astrophysics Corpora

Becker et al. (2005) and Hachey et al. (2005) built the Astronomy Bootstrapping Corpus (ABC) to explore named entity recognition. The corpus is composed of 209 abstracts of published radio astronomical papers. The annotation guide comprises four named entity categories, covering astronomical instrument names (111 instances), celestial object names (136), types (499), and their spectral features (321). In their joint study, authors explored an active learning approach for leveraging the annotation cost. They demonstrated that using committee-based metrics to quantify disagreement between classifiers can optimize the selection of informative data points, resulting in substantial cost savings compared to random sampling (Hachey et al., 2005). To our knowledge, the corpus is not accessible.

Similarly, Murphy et al. (2006) focuses on NER by building a corpus of 7840 sentences from astronomical papers. Compared to the ABC, the concepts covered in this corpus are broader, with 43 defined named entities, expanding to categories characterizing celestial objects, such as their coordinates and physical properties (frequency, luminosity). The authors proposed a system achieving an F1-score of 0.878. To our knowledge, this corpus is not accessible either.

More recently, the DEAL shared task corpus (Grezes et al., 2022) has been released, establishing one of the first accessible[3] corpus in astrophysics. It comprises full-text fragments and acknowledgment sections extracted from astrophysics papers and specifically annotated for the shared task. The corpus consists of three subsets: a training set with 1753 documents, a development set with 1366, and a test set with 2505. The authors defined 31 named entity categories. Among the shared task participants, Ghosh et al. (2022) proposed a NER system ranked first, achieving an F1-score of 0.8364 on the test set of this corpus.

In a previous paper, we introduced TDAC Alkan et al. (2022), the only NER-oriented corpus based on different astronomical documents. It comprises 75 observation reports (short textual messages), constituting one of the main ways of information sharing and communication for time-domain astronomers. The training and test sets comprise 59 and 16 documents. These reports consist of 25 *circulars* from the GCN Network (Barthelmy et al., 1995), 25 *telegrams* from the ATel system (Rutledge, 1998) and 25 *astronotes* from the Transient Name Server (Gal-Yam, 2021). Unlike DEAL and previously cited corpora, this corpus focuses on a specific branch of astronomy (time-domain astrophysics), possessing thus a particular vocabulary and discourse not necessarily found in general astrophysics papers. The TDAC corpus is accessible[4].

2.2. Coreference-oriented Corpora

Compared to NER, the task of coreference resolution in astrophysics documents has received less attention. Kim and Webber (2006), dealt with anaphora resolution in astrophysics literature. More precisely, their study was restricted to the automatic linking of pronouns to their corresponding citations.

Brack et al. (2021) built a coreference resolution corpus comprising ten different scientific disciplines (including eleven annotated abstracts in astrophysics). The system proposed by the authors achieved a CoNLL F1-score of 0.611 in astrophysics papers.

2.3. Summary

Annotated corpora in astrophysics are limited. These resources primarily feature NER-oriented corpora, which limits their utility for broader NLP tasks within astrophysics. Moreover, the nature of these resources predominantly comprises scholarly-type papers, restricting the variety of data sources that researchers can leverage. To fill this gap, we based our work on the existing TDAC corpus to build a richer corpus and extend it to cover non-treated NLP tasks such as coreference resolution, astrophysical relation detection, and celestial object names normalization (entity linking).

3. Corpus Annotation

In this section, we present text processing we applied to the TDAC corpus (Alkan et al., 2022), which we used as a starting point to build our new corpus (3.1). We then present the annotation process into named entities (3.2), celestial object names normalization (3.3), coreferences (3.4) and astrophysical relations annotations (3.5). We finish this section by presenting the difficulties encountered

astroECR

[3] https://huggingface.co/datasets/adsabs/WIESP2022-NER/

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

when annotating our corpus (3.6). We used BRAT (Stenetorp et al., 2012) as the annotation tool.

3.1. Corpus Pre-Processing

We noticed some persistent noises in the TDAC corpus coming from the content of HTML tags that were not entirely removed (e.g., `href="https://url.url"` or `data-show-count='false'`). Thus, we first conducted an HTML code cleanup to rectify this and improve URL text processing. For instance, we cleaned the following text:

- *It is almost identical to* ***href="https://wis-tns.weizmann.ac.il/object/2010gx" target="_blank">SN2010gx*** *at +4D after peak (Pastorello et al. 2010, ApJ, 724, L16).*

into,

- *It is almost identical to* ***SN2010gx*** *at +4D after peak (Pastorello et al. 2010, ApJ, 724, L16).*

3.2. Named Entity Annotation

3.2.1. Extending Named Entity Categories

We expanded the annotation guideline of TDAC with five new categories considered essential by astronomers.

- Non-numerical-type categories:
  - **`Date`**: This category includes dates and temporal expressions referring to a detection date or the duration of an observation. Example: *We report the discovery of a probable nova in M31 on a co-added 990-s R-band CCD frame taken under poor conditions on* ***2019 Mar. 12.791 UT*** *[Date] with the 0.65-m telescope at Ondrejov.*
  - **`Reference`**: This covers references in the text to other observation reports. This will be helpful to gather all reports dealing with the same celestial object. Example: *In comparison to the optical region (ref: the SALT spectrum in* ***ATel #3289*** *[Reference]), few strong NI lines are expected in the JHK bands*.
- Equation and numerical-type categories:
  - **`Magnitude`**: This category includes equations and numerical values that characterize the brightness of celestial bodies, a useful property for astronomers to determine the visibility of celestial objects. Example: *As reported to CBAT, this nearby-M31 object was discovered by Koichi Itagaki at* ***16.5 mag.*** *[Magnitude]*
  - **`Flux`**: A numerical value that characterizes the energy passing through a unit area per unit time. Example: *The flux values ranged from* ***1.01 +/-0.06 E+11 cgs*** *[Flux] to* ***1.71 +/- 0.04 E+11 cgs.*** *[Flux]*
  - **`Redshift`**: This category includes equations and numerical values characterizing the distance of a celestial body relative to an observer. Example: *The host KUG 0180+227 is an E+A galaxy at* ***z=0.022.*** *[Redshift]*

For better organization and comprehension, we proposed a taxonomy (see Table 11 in the Appendix) that categorizes these named entities into generic or astrophysics-related classes and subclasses.

3.3. Celestial Object Name Normalization

We focus on linking celestial object names like stars, planets, and galaxies to their specific entries in astronomical catalogs. Due to diverse naming conventions and cataloging systems in astronomy, this linking is crucial for preventing confusion and misinterpretation when integrating data from different papers and observation reports. For instance, the Andromeda Galaxy has at least 39 designations[5] that all need correct association. This process would thus avoid misunderstandings and determine the number of distinct objects mentioned in the text. For this purpose, we used SIMBAD (Wenger et al., 2000), NED (Mazzarella et al., 2001), and TNS (Gal-Yam, 2021) astronomical catalogs to link/normalize celestial object names in our corpus.

3.4. Coreferences Annotation

The details of our annotation guideline are accessible in our GitHub repository. Here, we give a broad overview of the type of coreferences annotated in our corpus.

3.4.1. Scope of Coreference Annotation

We created a class `CorefExp` gathering anaphoric and coreferential relations without distinction for the annotation process. We annotated mentions exclusively as coreferential relations when linked to a celestial object (the named entity of type `CelestialObject`). We also decided to annotate the cases where a celestial object is designated with another of its names within the text as a coreferential relation. However, we excluded mathematical

[5] http://simbad.cds.unistra.fr/simbad/sim-id?Ident=Andromeda+Galaxy&NbIdent=1&Radius=2&Radius.unit=arcmin&submit=submit+id

expressions, numerical quantities, and other coreferential relationships not associated with a celestial object from the annotation process. To illustrate this distinction, consider the following examples:

- In-scope coreferences:
  - *We discovered **PS19did**[1] on MJD 58666.31 = 2019-07-02.31, at w=19.9 +/- 0.1 [...] **The new transient source**[1] is in the galaxy **UGC 11003**[2] [...] Adopting **the host galaxy**[2] redshift z=0.03566 (NED) yields an expansion velocity [...] Followup observations of **this intrinsically faint transient**[1] are encouraged.*
  - *We report on the discovery and follow-up of a very bright and highly magnified microlensing event **Gaia19bld**[1]. [...] **It**[1] has been detected and announced by the Gaia Science Alerts program.*
  - *We report on the NIR brightening of the intermediate redshift quasar **PKS0735+17**[1] (z=0.424), also known as **CGRaBSJ04738+1742**[1].*
- Out-of-scope coreferences:
  - *Analysis of **the data**[1] is ongoing. We remind the community that all **Swift data**[1] are public, and encourage **their**[1] use.*
  - ***The observations**[1] continued until 2019-04-26 20:15 UT, when **they**[1] were aborted to begin followup of.*
  - *The estimated AB magnitude is **17.6**[1]. **This magnitude**[1] is not corrected for the host galaxy contribution.*

3.5. Astrophysical Relationship Annotation

3.5.1. Relations Between Celestial Bodies and Physical Properties

In our annotation scheme, we are not connecting all named entities. Only `CelestialObject`-type mentions can be connected to entity mentions describing the physical attributes of celestial objects such as `CelestialObjectRegion` (coordinates in the sky), `Wavelength`, `Magnitude`, etc. We defined the `related_to` relationship to link entity pairs.

3.5.2. Within and Cross-Sentences Relationship

The information is spread throughout the observation report, and the described physical properties of celestial objects (coordinates, wavelength, magnitude, etc.) are not necessarily within the same sentence. This, in turn, increases the distance between entity mentions in the relationship, making the relation detection difficult if the sequence is too long. To mainly address a within-sentence relationship detection task, we decided to link the properties described in the text to the closest mention of the celestial object's coreference in the text. Therefore, the coreference mentions inheriting the properties of celestial objects. For instance, as illustrated in Figure 1, named entities of type `CelestialRegion` are linked to the coreference mention "*the object*" as it is the closest, dealing thus with a within-sentence relation detection.

3.6. Annotation Difficulties

Whether for named entities annotation, coreferences, or semantic relationships annotation, the nature of astrophysics texts makes the annotator task challenging.

Areas of Ambiguity surrounding specific terms complicates the task of annotators. For instance, "*Gaia*" may refer to the `Mission` or the `Telescope`, as well as "*Fermi*" may refer to the `Telescope` or a group of scientists. Additionally, astrophysical NER involves dealing with intricately linked concepts, and this proximity can pose difficulties when deciding how to assign labels to these entities.

Surface Forms Certain concepts described in texts can be misleading. For example, the coordinates of celestial objects are often described in multiple forms, one of which, commonly encountered, closely looks to a time and date format, posing a challenge for non-expert annotators. Example: *We report the discovery of a nova (RA = **00h45m02.36s**, Dec (2000)= **+41d14'39.8"***.

Acronyms The domain extensively employs acronyms to designate celestial bodies, adding complexity to the annotation process. Deciphering what these acronyms refer to requires domain expertise. Annotating coreferential relations is more complex with these acronyms. For instance, in the example below, recognizing that "*PSN*" stands for "*Probable SuperNova*" makes the annotator task challenging to link it with the celestial object name in the text). The task is even more complicated when multiple sources are mentioned in the text, and for which we use distinct acronyms and expressions (see example below).

- *MASTER-SAAO auto-detection system discovered **MASTER OT J105440.86-391319.0**[1]. [...]. **This PSN**[1] is in 2.9"E,7"N from the center of **PGC600519**[2]. [...] **MASTER OT***

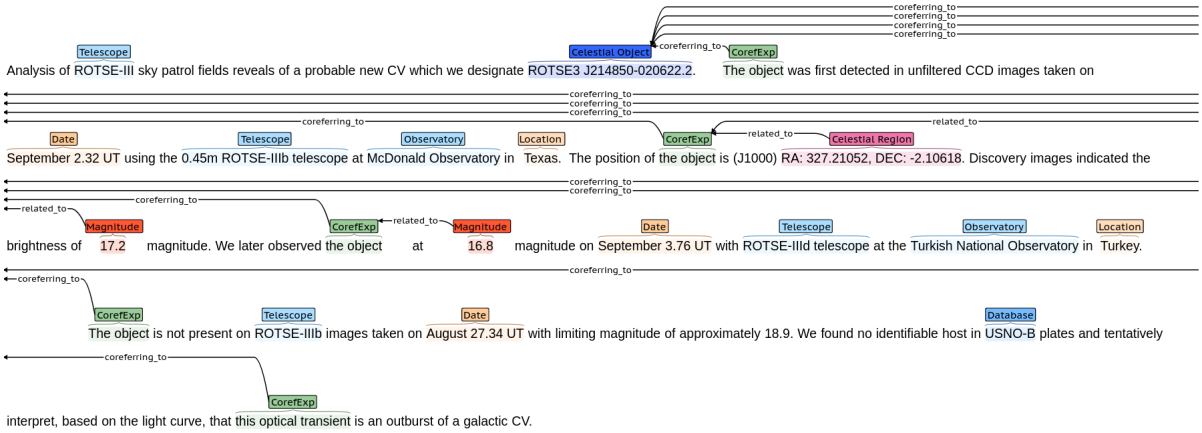

Figure 1: Example of an annotated observation report with BRAT. In light blue, the `Telescope` and `Observatory`-type mentions. In dark blue, the mention "*ROTSE3 J214850-020622.2*" of type `CelestialObject` indicates the main subject of this observation report. The solid black lines represent coreference between a referring mention such as "*The object*" (`CorefExp`, illustrated in green) to its antecedent (the celestial object "*ROTSE3 J214850-020622.2*"). In pink, the coordinates of the celestial object. In dark orange, the measured magnitudes (`Magnitude`) of the celestial source. We created a semantic relation ("`related_to`") that links these physical properties to mentions of the corresponding celestial object. The `Date`-type mentions are in orange.

> ***J155546.00-734455.8***$_{[3]}$*. [...].* ***The OT***$_{[3]}$ *is seen in 4 images.*

In this example, identifying which celestial object the mentions "*PSN*" and "*The OT*" are referring can be a non-trivial task for a non-expert due to the potential for multiple interpretations and associations: "*The OT*" is it referring to "*MASTER OT J105440.86-391319.0*", "*MASTER OT J155546.00-734455.8*" or to the galaxy "*PGC600519*" ?

4. Statistics of the Resulting Annotated Corpus

In this section, we describe the main characteristics of the resulting annotated corpus (inter-annotator agreement and statistics), and we provide comparative tables between TDAC and astroECR corpora.

4.1. Inter-Annotator Agreement and Adjudication

To establish an Inter-Annotator Agreement (IAA), we selected a subset of 30 documents (6499 tokens). These documents were provided to two annotators (one astrophysicist and one NLP expert) for their independent annotations. Annotation errors have been tracked and solved by a consensus/adjudication phase, allowing errors to be handled and thus producing the reference/gold annotation. In addition, we compared both annotators' annotations with the produced reference set using precision, recall, and F1 scores as previously done by Galibert et al. (2012). Results in Table 1 show that the astrophysics expert achieved a higher F1 score (0.94) than the NLP (0.91) compared to the consensus, that is why the astrophysics expert conducted the remaining annotations (270 documents) to ensure the annotated concepts' correctness.

4.2. Statistics

Source Data Statistics TDAC comprises three types of observation reports: *telegrams* (ATels), *circulars* (GCN) and *AstroNotes*, equally represented in the corpus. However, when building astroECR, we included more ATels because we realized that the latter contained a greater diversity of astrophysical concepts during annotation. Table 2 shows the compositions of the two corpora.

Training and Test Sets Annotation Statistics For our corpus's train/test split, we based it on the distribution of coreferential and astrophysical relations (to have enough relations in each set). We ensured we had around 80% of relations in the training set and 20% in the test set. Table 3 shows the number of annotated documents and the number of annotated tokens (in named entities) in each set of both corpora.

Since coreferences and semantic relations are not annotated in the TDAC corpus, we only report in Table 4 and Table 5 statistics on our corpus.

Table 5 shows the overall number of annotated semantic relations in each set of our corpus. Our choice of annotating astrophysical relations between celestial objects' coreferential mentions and

Annotation Task	Annotation Comparison	Exact-Match			Inexact-Match		
		P	R	F1	P	R	F1
Named Entities	Astro vs. NLP	0,65	0,59	0,62	0,84	0,92	0.88
	Astro vs. consensus	0,83	0,86	0,84	0,93	0,96	0,94
	NLP vs. consensus	0,73	0,69	0,71	0,94	0,89	0,91
Coreferences	Astro vs. TAL	0,77	0,88	0,82	0,78	0,89	0,83
	Astro vs. consensus	0,97	1,00	0,98	0,97	1,00	0,98
	TAL vs. consensus	0,74	0,89	0,81	0,75	0,90	0,82

Table 1: Inter-annotator agreement for the annotation of named entities, mentions of coreferences between the two annotators, and comparison with the consensus. The astrophysicist annotator is called "*Astro*", and the NLP expert is called "*NLP*". The metrics used are Precision (P), Recall (R), and F-measure (F1). We applied two evaluation modes : exact match and inexact match. In the exact-match, the extracted entity is considered as a true positive if both entity type and boundaries are correctly extracted, as a false positive if it was wrongly labeled, and as a false negative if it was not annotated. The inexact-match evaluation setting, allows entities to match if their boundaries overlap: an extracted entity is counted as a true positive if it shares half of the tokens with the gold entity.

Source Data	TDAC	astroECR
ATels	25	175
GCN	25	100
AstroNotes	25	25

Table 2: Composition of TDAC and astroECR corpora, with details of the data sources used.

Parameters	TDAC		astroECR	
	Train	Test	Train	Test
# documents	59	16	210	90
# tokens	15374	3638	43481	10578
# ann. tokens	4338	1014	17392	3173

Table 3: Number of annotated documents and number of annotated tokens in each set of both corpora.

Parameters	astroECR	
	Train	Test
# coref. ment.	412	101
# coref. clust.	257	65
avg. clust. len.	3.5 (+/- 2.26)	3.4 (+/- 1.61)

Table 4: Number of coreferent mentions, number of clusters including singleton clusters, and the average cluster length (with standard deviation).

their physical properties shows that most relations are within sentences.

Parameters	astroECR	
	Train	Test
# within-sent rel.	490	143
# cross-sent rel.	154	26
overall ann. sem. rel.	644	169

Table 5: Number of annotated semantic relations in the train and test sets of our astroECR corpus with details by type of relations (within and cross-sentences).

5. First Experiments on the Resulting Annotated Corpus

In this section, we present the use cases of our corpus following the annotations we made. We provide initial methods to demonstrate how to use this corpus and its annotations for four different NLP tasks: named entity recognition (5.1), celestial object-named entity linking (5.2), coreference resolution (5.3), and relation detection (5.4).

5.1. Named Entity Recognition

Named Entity Recognition (NER) identifies mentions of entities from text belonging to predefined semantic types: person, location, or organization (Yadav and Bethard, 2018). This task has proven to be useful for information retrieval (Banerjee et al., 2019) and also for building question-answering systems (Mollá Aliod et al., 2006).

Experimental Setup We fine-tuned two transformer models on the training set of our corpus, astroBERT (Grezes et al., 2021), and SciBERT (Beltagy et al., 2019), and evaluated them on the test set. We fine-tuned them using the same hyperparameters used by Alkan et al. (2022), i.e., on 20 epochs, with a learning rate $\alpha = 2.10^{-5}$, and a training batch size of 4.

5.2. Celestial Object Name Linking

Named Entity Linking (NEL) is a task that involves linking named entities mentioned in the text to specific entries or entities in a knowledge base or reference dataset (Sevgili et al., 2020). The primary objective of NEL is to disambiguate named entities and connect them to their corresponding real-world entities or concepts. In our case, celestial objects

have multiple designations. These different designations are often used in the same documents or others, making entity linking a challenging task.

Querying Knowledge Bases The baseline system we propose solely relies on external knowledge querying. We implemented a system that first queries the SIMBAD (Wenger et al., 2000) database, utilizing ADQL queries (similar to SQL) to identify celestial sources by their names. For each source identified within the SIMBAD catalog, we extract its unique identifier and compile an exhaustive list of all its designations and canonical names. In cases where the source remains elusive within SIMBAD, our system further extends its search to the NED database (Mazzarella et al., 2001) and, if necessary, the TNS (Gal-Yam, 2021). These three catalogs collectively encompass many events observed in time-domain astrophysics, substantially augmenting our probability of identifying the source. Notably, NED specializes in extragalactic sources, while TNS is mainly used for identifying supernovae, making each database a valuable complement to the others in our celestial object identification endeavor.

5.3. Coreference Resolution

Coreference resolution is an information extraction task aiming to identify all the mentions in a text that refer to the same entity (Zheng et al., 2011).

Performance of Existing System We have tested an existing coreference system, F-coref (Otmazgin et al., 2022), on the test set of our corpus. F-coref is based on the LingMess architecture (Otmazgin et al., 2023). We opted for F-coref because the model is easily callable through its Python package *fastcoref*[6]. We first evaluated the model without fine-tuning by comparing its predictions with our annotations. Then, we fine-tuned the model on 50 epochs using the train set of our corpus and evaluated it on the test set. Each experiment has been run five times with different random seeds.

5.4. Relation Detection

Relation Detection (or extraction) is a task that aims to identify and classify the relationships between pairs of entities (Bassignana and Plank, 2022). In the case of binary relation detection, the objective is to determine whether a relationship exists between two entities in a given sequence. Researchers commonly model the task as a classification problem. In our study, we aim to detect whether a relation exists between celestial objects and physical properties and to accurately link those physical properties to their respective celestial objects, even when a text describes multiple objects.

[6] https://pypi.org/project/fastcoref/

Sentences	Label
The median magnitude of ***@@FO Aqr$_{[CelestialObject]}$$*** *in 498 ASAS-SN observations from 2012-2015 was* ***@@V=13.54$_{[Magnitude]}$$***.	1
A spectrum was obtained using the SPRAT spectrograph on the ***@@Liverpool Telescope$_{[Telescope]}$$***. *Classification indicates it is a type Ia supernova with estimated redshift* ***@@z=0.078$_{[Redshift]}$$***	0

Table 6: Examples of positive (denoted by the label 1) and negative (label 0) relations. For the training process, entities in sentences are marked with the symbols @@ and $$.

Generating Negative Examples Table 5 shows that our training and test sets comprise 644 and 169 annotated relations, respectively. To build a binary relation detection system to categorize whether a relation exists between two entities, we must include negative sequences in both sets (i.e., examples where no relations appear in the sequence). To ensure a balanced class distribution during training and evaluation, we maintained an approximately equal number of negative examples to positive ones. We leveraged the annotated named entities to create these negative example sequences for both sets. Specifically, we randomly selected sequences marked by two named entities lacking semantic relationships (e.g., we paired a mention of type `Redshift`, characterizing the distance of an object, with a `Telescope`). In some instances, we selected documents involving multiple celestial objects to establish connections between an object and the physical properties of another object described in the text. The training set comprises 644 positive samples and 712 negative examples for the relationship detection task, while the test set includes 169 positive and 180 negative examples. Table 6 illustrates examples from the training set.

Model Setup As the first experiments on astrophysical relation detection between celestial objects and physical properties, we fine-tuned a bidirectional Long Short-Term Memory (biLSTM) neural network. We fine-tuned our system on the train set of our corpus during 20 epochs with a learning rate $\alpha = 10^{-3}$ and a training batch size set to 128. We evaluated the system performance on the test set.

6. Results and Error Analysis

This section presents and analyzes our results on the four NLP tasks.

6.1. Named Entity Recognition

Results are presented in Table 7 below.

Model	Precision	Recall	F1
SciBERT	0.85	0.75	0.79
astroBERT	0.83	0.81	0.82

Table 7: Performance of a SciBERT-based and an astroBERT-based NER system fine-tuned and evaluated on our corpus. Metrics used are Precision, Recall, and F1-score.

The results of our fine-tuned NER systems, SciBERT and astroBERT, for astrophysical named entity recognition show satisfying performance in identifying key domain-specific concepts. Notably, astroBERT exhibits high F1 scores for recognizing the names of celestial objects (F1 score of 0.84) and their associated physical properties (e.g., an F1 score of 0.74 for wavelength) as well as for detecting astronomical facilities (F1 score of 0.79 for observatories). However, there is still room for improvement in detecting references to other observation reports, a class specifically created for our task, as the F1 score in this category stands at 0.64. This suggests that further refinements are needed to enhance the recognition of such references within the astrophysical context.

6.2. Linking Celestial Object Names

Table 8 shows that catalogs significantly enhance the accuracy of linking celestial object names.

Catalogue	Accuracy
SIMBAD	60.39
SIMBAD + NED	71.28
SIMBAD + NED + TNS	80.19

Table 8: Celestial object linking accuracy.

The three catalogs, namely SIMBAD, NED, and TNS, show a cumulative effect in improving accuracy, indicating that they are complementary in their contributions. Our system achieves 80.19% accuracy in linking celestial object names in the corpus. Our system encounters challenges with certain types of objects, such as Gravitational Waves (e.g., "*S190426c*") and objects with non-standard naming schemes (e.g., with punctuation in the name "*GRB210303.42*").

6.3. Coreference Resolution

Table 9 reports the results we achieved when applying the F-coref (Otmazgin et al., 2022) coreference resolution system on our test set. We made two evaluations: first, using the existing tool, and second, fine-tuning the tool on our training set (astroFastCoref).

Model	CoNLL		
	Precision	Recall	F1
F-coref	0.09 (0)	0.26 (0)	0.13 (0)
astroFastCoref	0.67 (0.01)	0.44 (0.01)	0.53 (0.01)

Table 9: Mean precision, recall, and F1-score (with standard deviation) of the F-coref system evaluated on the test set of our corpus with and without fine-tuning. Each experiment has been run five times (on 50 epochs when fine-tuning) with different random seeds.

Our evaluation shows that F-coref has a very low precision (0.09) and a high recall (0.26), resulting in a low F1 score (0.13). The model lacks the domain-specific knowledge to resolve coreferences in this context accurately. However, fine-tuning the model (astroFastCoref) allowed us to learn astrophysics-specific patterns, making it more efficient in resolving coreferences related to celestial objects by reaching a CoNLL F1-score of 0.53.

6.4. Relation Detection

Table 10 presents the results we obtained by applying a biLSTM neural network for relation detection. The performance of our baseline system is satisfying. Our system can detect whether a relation exists or not in almost 80% of the cases.

Label	Precision	Recall	F1
0	0.82	0.79	0.81
1	0.77	0.80	0.79

Table 10: Performance of a biLSTM binary relation detection system on our corpus. We assigned Label 1 to sequences containing a relation between a celestial object and a physical property and Label 0 otherwise.

The results of our biLSTM binary relation detection system on astrophysical relation detection are encouraging, especially considering that this is a baseline model. The overall F1 score suggests a strong performance in identifying the presence and absence of relations between celestial objects and physical properties. Notably, our model achieves high precision and recall for both Label 0 and Label 1, indicating its proficiency in distinguishing

between sentences with and without such relations. However, the model faces challenges when dealing with longer sequences, particularly in the context of cross-sentence relation detection. Additionally, there is room for improvement in cases where multiple celestial objects are described in the text, as the system still needs to enhance its ability to correctly associate the relevant properties with the corresponding celestial objects. These observations provide valuable insights for further refining and optimizing our relation detection system in astrophysical contexts.

7. Conclusion and Outlook

The scarcity of annotated resources and the need for more diversity in existing corpora have been significant challenges in astrophysical NLP. Our primary objective was to bridge this data and diversity gap by expanding and enhancing the TDAC corpus (Alkan et al., 2022). Based on the 75 documents of TDAC, we introduced astroECR, a new corpus of 300 astrophysical observation reports. We annotated each text with named entities, coreferences, and astrophysical relations between celestial objects and their physical properties. Additionally, we normalized celestial object names of our corpus by linking them to astronomical databases. Through this endeavour, we conducted initial experiments across four NLP tasks to showcase the potential utility of our corpus. While these preliminary results show promising outcomes, areas still need refinement and improvement. With our corpus, we aim to provide the NLP community with a resource that can facilitate complementary studies on scientific coreference resolution or relation detection. Indeed, this corpus may represent an additional resource to complement and extend previous scientific corpora, such as Chaimongkol et al. (2014) and Brack et al. (2021). Our initial models could also be used to reduce the annotation cost for additional documents. From an astrophysical perspective, models trained on our corpus can be used for information extraction, as proposed by Sotnikov and Chaikova (2023). Specifically, we propose integrating these models into the Astro-COLIBRI platform, which processes real-time alerts from observers regarding transient sources (Reichherzer et al., 2021, 2023). By deploying our NLP models within Astro-COLIBRI, we enable astrophysicists to rapidly access pertinent information in observation reports. We make our corpus, annotation guidelines, associated code, and models accessible to both communities.

8. Bibliographical References

Partha Sarathy Banerjee, Baisakhi Chakraborty, Deepak Tripathi, Hardik Gupta, and Sourabh S. Kumar. 2019. A information retrieval based on question and answering and ner for unstructured information without using sql. *Wirel. Pers. Commun.*, 108(3):1909–1931.

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.

Elisa Bassignana and Barbara Plank. 2022. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 67–83, Dublin, Ireland. Association for Computational Linguistics.

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.

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

Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, and Ludovic Quintard. 2012. Extended named entities annotation on OCRed documents: from corpus constitution to evaluation campaign. In *International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

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 first Workshop on Information Extraction from Scientific Publications*, pages 100–104, Online. Association for Computational Linguistics.

Felix Grezes, 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.

Felix Grezes, Sergi Blanco-Cuaresma, Thomas Allen, and Tirthankar Ghosal. 2022. Overview of the first shared task on detecting entities in the astrophysics literature (DEAL). In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 1–7, Online. 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.

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

Joseph M. Mazzarella, Barry F. Madore, and George Helou. 2001. Capabilities of the NASA/IPAC extragalactic database in the era of a global virtual observatory. In *SPIE Proceedings*. SPIE.

Diego Mollá Aliod, Menno van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Workshop, ALTA 2006, Sydney, Australia, November 30-December 1, 2006*, pages 51–58. Australasian Language Technology Association.

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.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, Josh Peek, Kartheik Iyer, Tomasz Różański, Pranav Khetarpal, Sharaf Zaman, David Brodrick, Sergio J. Rodríguez Méndez, Thang Bui, Alyssa Goodman, Alberto Accomazzi, Jill Naiman, Jesse Cranney, Kevin Schawinski, and UniverseTBD. 2023. Astrollama: Towards specialized foundation models in astronomy.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 48–56, Taipei, Taiwan. Association for Computational Linguistics.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. LingMess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2752–2760, Dubrovnik, Croatia. 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.

Patrick Reichherzer, Fabian Schüssler, Valentin Lefranc, Julia Becker Tjus, Jayson Mourier, and Atilla Kaan Alkan. 2023. Astro-colibri 2—an advanced platform for real-time multi-messenger discoveries. *Galaxies*, 11(1).

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.

Özge Sevgili, Artem Shelmanov, Mikhail Y. Arkhipov, Alexander Panchenko, and Chris Biemann. 2020. Neural entity linking: A survey of models based on deep learning. *CoRR*, abs/2006.00575.

Vladimir Sotnikov and Anastasiia Chaikova. 2023. Language models for multimessenger astronomy. *Galaxies*, 11(3).

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

M. Wenger, F. Ochsenbein, D. Egret, P. Dubois, F. Bonnarel, S. Borde, F. Genova, G. Jasniewicz, S. Laloë, S. Lesteven, and R. Monier. 2000. The SIMBAD astronomical database. *Astronomy and Astrophysics Supplement Series*, 143(1):9–22.

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.

Jiaping Zheng, Wendy W. Chapman, Rebecca S. Crowley, and Guergana K. Savova. 2011. Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, 44(6):1113–1122.

9. Language Resource References

Alkan, Atilla Kaan and Grouin, Cyril and Schussler, Fabian and Zweigenbaum, Pierre. 2022. *TDAC, The First Corpus in Time-Domain Astrophysics: Analysis and First Experiments on Named Entity Recognition*. Association for Computational Linguistics. [link].

Arthur Brack, Daniel Uwe Müller, Anett Hoppe, and Ralph Ewerth. 2021. Coreference resolution in research papers from multiple domains. *CoRR*, abs/2101.00884.

Chaimongkol, Panot and Aizawa, Akiko and Tateisi, Yuka. 2014. *Corpus for Coreference Resolution on Scientific Papers*. European Language Resources Association (ELRA). [link].

A. Appendix

Taxonomy of astrophysical named entities We propose a taxonomic representation of the named entities in Table 11.

Category	Definition
Generic Classes	
- Archive (Arc)	A curated collection of the literature or data.
- Organization (Org)	A named organization that is not an observatory.
- Person (Per)	A named person or their initials.
- Location (Loc)	A geographical location (city, country).
- Software (Sof)	Software, IT tool.
- URL (URL)	A link to a website.
- Date	Dates and temporal expressions referring to a detection date or the duration of an observation.
Astrophysics-Domain Specific Classes	
- **Celestial Bodies:**	
- CelestialObject (COb)	A named object in the sky.
- CelestialObjectRegion (COR)	Named area on/in a celestial body.
- **Physical Properties:**	
- CelestialRegion (CeR)	A defined region projected onto the sky, or celestial coordinates.
- Wavelength (Wav)	Portion of the electromagnetic spectrum.
- Flux (Wav)	A numerical value that characterizes the energy passing through a unit area per unit time.
- Magnitude (Wav)	Equations and numerical values that characterize the brightness of celestial bodies.
- Redshift (Wav)	Equations and numerical values characterizing the distance of a celestial body relative to an observer.
- **Observation Instruments:**	
- Observatory (Obs)	A, often similarly located, group of telescopes.
- Telescope (Tel)	A "bucket" to catch light.
- Instrument (Ins)	A device, often, but not always, placed on a telescope, to make a measurement.
- ObservationalTechniques (ObT)	A method used to observe celestial objects.
- **References and Collaborations:**	
- Citation (Cit)	A reference to previous work in the literature.
- Reference (Cit)	References in the text to other observation reports.
- Collaboration (Col)	Name of collaboration.
- Grant (Gra)	An allocation of money and/or time for a research project.
- Survey	An organized search of the sky often dedicated to large scale science projects.

Table 11: Proposed taxonomy of astrophysical named entities used in the astroECR corpus.