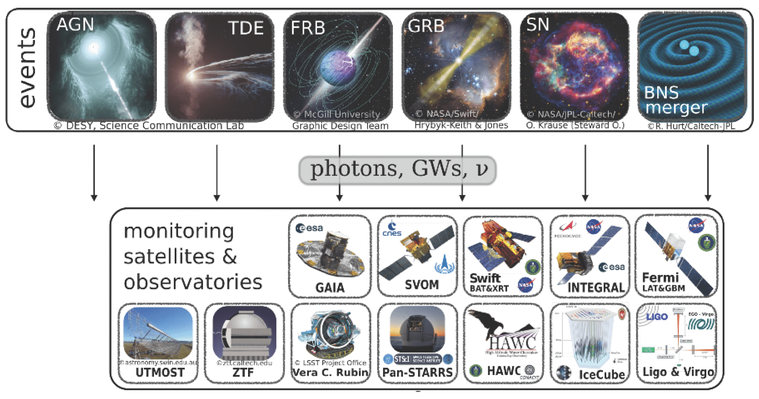
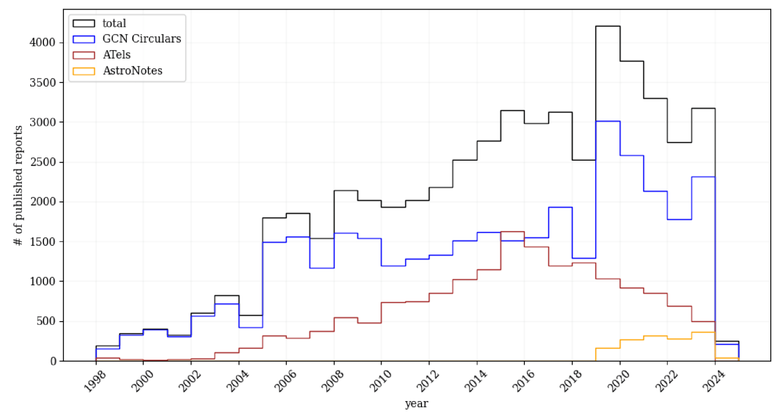
Enriching a Time-Domain Astrophysics Corpus with Named Entity, Coreference, and Astrophysical Relationship Annotations

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

Astronomical Observations are Increasing





astroECR Corpus Creation

- Source Corpus: TDAC (75 documents);
- Creation of astroECR (300 documents);
- Annotations: named entities, coreferences, and semantic relations.

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

An Annotated Astronomical Observation Report

The TDAC Corpus for Named Entity Recognition:

We report the spectroscopic confirmation of four supernovae with observations obtained with the 1.82-m Copernico Telescope (Telescope) in Asiago (obs) (+ AFOSC (instr); range 340-820 nm, resolution 1.3 nm). A spectrogram of PSN J23012936+0653381 (Celestial Object) obtained on Dec. 17.72 UT (Date) suggests that this is a type-Ia at redshift z~0.039 (Redshift). The best match was found with several type-Ia supernovae about 1 week after maximum light. The expansion velocity as deduced from the position of the Si-II 635.5nm absorption is about 10800 km/s. A spectrogram of PSN J06434189+5212337 (Celestial Object) obtained on Dec. 19.01 UT (Date) shows that this is a type-Ia at redshift z~0.043 (Redshift). The best match was found with type-Ia supernova 1994D (Patat et al. 1996, MNRAS 278, 111) (Citation) about two weeks after maximum light. The expansion velocity as deduced from the position of the Si-II 635.5nm absorption is about 8600 km/s. A spectrogram of PSN J09413802+4840255 (Celestial Object) (= SNhunt162 (obj)) obtained on Dec. 19.12 UT (Date) shows that this is a type II supernova. Together with Balmer lines showing P-Cygni profiles, Fe II lines and the Na I doublet feature are detected. Adopting for the host galaxy MCG+8-18-23 (CelestialObjectRegion) the redshift z = 0.038054 (Redshift) (SDSS 2004 Data Release 3, via NED (base)), the best fit is obtained with type IIP SN 1999gi (obj) (Leonard et al. 2002, AJ 124, 2490 (Citation); Smartt et al. 2001, ApJ 556L, 29 (Citation)) at maximum light. A spectrogram of PSN J10354824+3900279 (Celestial Object) obtained on Dec. 19.33 UT (Date) suggests that this is a type-Ia at redshift z~0.044 (Redshift). The best match was found with several type-Ia supernovae around maximum light. The expansion velocity as deduced from the position of the Si-II 635.5nm absorption is about 11200 km/s. The Asiago (obs) classification spectra are posted at this website: URL http://graspa.oapd.inaf.it (URL) ; classification was made via GELATO (sofw) (Harutyunyan et al. 2008, A.Ap. 488, 383 (Citation)) and SNID (sofw) (Blondin and Tonry 2007, Ap.J. 666, 1024 (Citation)).

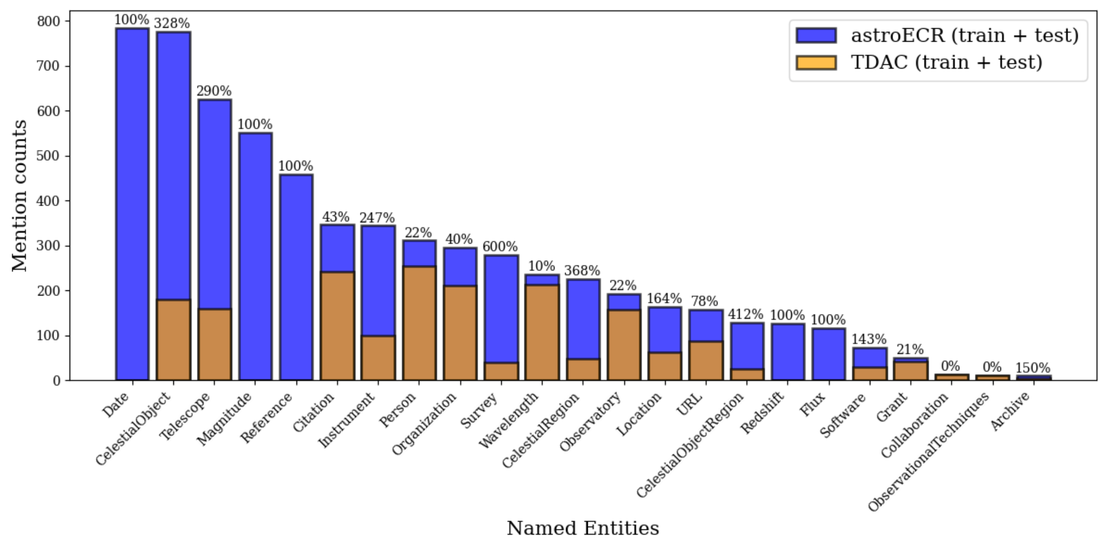
astroECR: an Astrophysics Corpus Annotated with Entities, Coreference, and Semantic Relations:

We report the spectroscopic confirmation of four supernovae with observations obtained with the 1.82-m Copernico Telescope (Telescope) in Asiago (loc) (+ AFOSC (instr); range 340-820 nm, resolution 1.3 nm). A spectrogram of PSN J23012936+0653381 (Celestial Object) obtained on Dec. 17.72 UT (Date) suggests that this (CorefExp) is a type-Ia at redshift z~0.039 (Redshift). The best match was found with several type-Ia supernovae about 1 week after maximum light. The expansion velocity as deduced from the position of the Si-II 635.5nm absorption is about 10800 km/s. A spectrogram of PSN J06434189+5212337 (Celestial Object) obtained on Dec. 19.01 UT (Date) shows that this (CorefExp) is a type-Ia at redshift z~0.043 (Redshift). The best match was found with type-Ia supernova 1994D (Patat et al. 1996, MNRAS 278, 111) (Citation) about two weeks after maximum light. The expansion velocity as deduced from the position of the Si-II 635.5nm absorption is about 8600 km/s. A spectrogram of PSN J09413802+4840255 (Celestial Object) (= SNhunt162 (obj)) obtained on Dec. 19.12 UT (Date) shows that this (CorefExp) is a type II supernova. Together with Balmer lines showing P-Cygni profiles, Fe II lines and the Na I doublet feature are detected. Adopting for the host galaxy MCG+8-18-23 (CelestialObjectRegion) the redshift z = 0.038054 (Redshift) (SDSS 2004 Data Release 3, via NED (base)), the best fit is obtained with type IIP SN 1999gi (obj) (Leonard et al. 2002, AJ 124, 2490 (Citation); Smartt et al. 2001, ApJ 556L, 29 (Citation)) at maximum light. A spectrogram of PSN J10354824+3900279 (Celestial Object) obtained on Dec. 19.33 UT (Date) suggests that this (CorefExp) is a type-Ia at redshift z~0.044 (Redshift). The best match was found with several type-Ia supernovae around maximum light. The expansion velocity as deduced from the position of the Si-II 635.5nm absorption is about 11200 km/s. The Asiago (loc) classification spectra are posted at this website: URL http://graspa.oapd.inaf.it (URL) ; classification was made via GELATO (sofw) (Harutyunyan et al. 2008, A.Ap. 488, 383 (Citation)) and SNID (sofw) (Blondin and Tonry 2007, Ap.J. 666, 1024 (Citation)).

TDAC vs astroECR: Global Annotation Statistics

Objects	TDAC		astroECR	
	Train	Test	Train	Test
# documents	59	16	210	90
# tokens	15374	3638	43481	10578
# annotated tokens	4338	1014	17392	3173
# coreferent mentions	-	-	412	101
# coreferences chains	-	-	257	65
# Within sentence relations	-	-	490	143
# Inter sentential relations	-	-	154	26
# Total relations	-	-	644	169

TDAC vs astroECR: Named Entities Distribution



Experimental Setup

- **Named Entity Recognition:** Evaluation of an astroBERT-based model on $\text{TDAC}_{\text{test}}$ using different training sets ($\text{TDAC}_{\text{train}}$ and $\text{astroECR}_{\text{train}}$);
- **Coreference Resolution:** Evaluation on $\text{astroECR}_{\text{test}}$ of FastCoref (baseline model) with astroFastCoref (trained on 50 epochs using $\text{astroECR}_{\text{train}}$);
- **Relation Detection:** Evaluation on $\text{astroECR}_{\text{test}}$ of a biLSTM model trained on 20 epochs using $\text{astroECR}_{\text{train}}$.

Experiments & Results

- **Improvement of named entity recognition on $\text{TDAC}_{\text{test}}$:**

Category	$\text{TDAC}_{\text{train}}$				$\text{astroECR}_{\text{train}}$				$\Delta F1$ (%)
	N	P	R	F1	N	P	R	F1	
CelestialObject	130	0,88	0,94	0,90	519	0,94	1,0	0,97	+ 7,7
CelestialRegion	20	0,31	0,23	0,26	149	0,64	1,0	0,78	+ 200
Observatory	60	0,54	0,58	0,64	101	0,80	0,67	0,72	+ 12,49

- **Performances of Coreference Resolution Systems on $\text{astroECR}_{\text{test}}$:**

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

- **Evaluation of a BiLSTM System for Relation Detection on $\text{astroECR}_{\text{test}}$:**

Precision	Recall	F1
0.77	0.80	0.79

Conclusion

- Enriching TDAC improved domain entities detection;
- Future directions for coreference resolution and relation extraction: explore BERT-based models;
- Deployement in Astro-COLIBRI for real-time analysis of observation reports.

References

- TDAC, the First Time-Domain Astrophysics Corpus: Analysis and First Experiments on Named Entity Recognition. *Proceedings of the first Workshop on Information Extraction from Scientific Publications (AACL-IJCNLP 2022)*.
- Astro-COLIBRI Platform: `https://astro-colibri.science`

Corpus + Code