System Development: Data Enrichment
Named Entities & Language Model

Felix Grezes, Thomas Allen and the ADS Team

ADSUG Meeting, November 2022
Why Use Machine Learning/AI at ADS

- What we are accomplishing
  - Text enrichment:
    - Identify Entities in the Literature (e.g., observatories, instruments)
    - Identify Planetary Feature Names
  - Build an Astronomy specific language model to automate the enrichment
  - Provide models and datasets for other researchers in the field

(2021 - 2026)
# Entities of Interest to the Astronomical Community

## Traditional NER
- Person
- Organization
- Location
- EntityOfFutureInterest

## Facilities
- Observatory
- Telescope
- Instrument
- Wavelength/Filter
- Archive
- Mission
- Collaboration
- Survey
- Database
- ComputingFacility

## Astronomical Objects
- CelestialObject
- CelestialRegion
- CelestialObjectRegion

## Other
- Citation
- Event
- Formula
- URL
- Identifier (e.g., DOIs or other document identifiers)
- Tag (e.g., `<whatever>`some valuable text`</whatever>`)
Labeled Entities Astronomy Dataset

- **Astrophysical Literature**
  - Full-text - 3009 snippets
  - 71631 labelled entities
- **Acknowledgements**
  - 3004 snippets
  - 76230 labelled entities
- **Years between 2015 and 2020**
- **Working to automate labeling process**

2018ApJ...854.1645S. 1645 the protosolar disk. In addition, D/H ratios can readily be modified by non-nebular processes. For example, the D/H ratio of Rumuruti type and ordinary chondrites is higher than that in comets, which Alexander et al. (2012)\(^\text{citation}\) attribute to iron oxidation in their parent bodies. 4.5. Water in Chondrites If carbonaceous and non-carbonaceous chondrites formed on opposite sides of the snowline, where Jupiter\(^{\text{crumhops}}\) formed, why do some carbonaceous chondrites have relatively low water and carbon contents like those in some non-carbonaceous chondrites? CM, CI, and CR chondrites have the highest contents of water and carbon among chondrites, but CO and CV have lower amounts that are comparable to those in some of the least metamorphosed ordinary chondrites (Krot et al. 2014\(^\text{citation}\)). Krot et al. (2015)\(^\text{citation}\) infer that CL, CM, and CR chondrites were altered under water/rock mass ratios of up to 0.6, whereas CV and CO chondrites were altered at lower water/rock ratios of 0.1–0.2. We do not know the answer to this issue but note that water contents of carbonaceous chondrites are related to matrix contents and that the coarse-grained matrix material likely contained more water than the fine-grained matrix rims. Thus, the low water content of CO and CV chondrites may reflect their low content of coarse-grained matrix material relative to CL, CM, and CR chondrites (Scott Krot 2014\(^\text{citation}\)). We also note that comets can contain more rock than ice (Brownlee 2014\(^\text{citation}\); Davidsson et al. 2016\(^\text{citation}\)), much less ice than the predicted theoretical water-rock ratio beyond the snowline of 1.2 (Krot et al. 2014\(^\text{citation}\); Palme et al. 2014\(^\text{citation}\)).

5. How and When Were the Two Isotopic Populations Mixed Together in the Asteroid Belt?\(^\text{citation}\) The isotopic dichotomy in the solar system lasted until at least 3–4 Myr after CAI formation, the likely accretion time of CR chondrites, which are the least metamorphosed and altered chondrites and have the highest concentration of presolar grains (Zhao et al. 2013\(^\text{citation}\); Buddle et al. 2017\(^\text{citation}\); Krot Nagashima 2017\(^\text{citation}\)). The two isotopic populations were then intermixed, most likely by the migration of the giant planets before the gas had dissipated in the disk. In the Grand Tack model\(^\text{citation}\), once Jupiter\(^\text{crumhops}\) and Saturn\(^\text{crumhops}\) reached masses of ~50 M\(_{\oplus}\), they migrated inward across the asteroid belt so that it was first emptied and then repopulated with S-type asteroids, which formed in the belt, and C-type, which formed outside Jupiter\(^{\text{crumhops}}\). Alternatively, planetesimals from the giant planet region may have been scattered into the asteroid belt as a natural side-effect of the growth of giant planets (Kretke et al. 2017\(^\text{citation}\); Raymond Izidoro 2017\(^\text{citation}\)). Raymond Izidoro (2017)\(^\text{citation}\) show that scattering during the rapid growth of the gaseous envelopes of giant planets would have scattered nearby planetesimals in all directions and may have delivered water to the growing Earth\(^{\text{crumhops}}\). However, the Grand Tack model\(^\text{citation}\) has the advantage that it creates a strongly mass-depleted belt with roughly equal masses of S-
Labeled Entities Astronomy Dataset

- Astrophysical Literature
  - Full-text - 3009 snippets
    71631 labelled entities
  - Acknowledgements - 3004 snippets
    76230 labelled entities
  - Years between 2015 and 2020
- Working to automate labeling process
automating the enrichment: astroBERT

astroBERT aims to provide automated enrichment services, similar to SIMBAD and NED but with broader scope.

Under the hood, astroBERT is language model tailored to astrophysics.

- statistical model that captures the ambiguity of the text used by astrophysicists
- based on the proven technology of BERT and SciBERT
- core component that can be applied to many other tasks, not just labelling entities
astroBERT details

- astroBERT is now publicly available for all
  - 😊 [https://huggingface.co/adsabs/astroBERT](https://huggingface.co/adsabs/astroBERT)
  - Includes multiple versions, tutorials, and a demo

- creating astroBERT
  - ~400K astronomy documents
  - ~50 days of computation on dual Nvidia V100 GPUs
  - all open source technologies (Tensorflow, Numpy...)
Evaluating astroBERT

- To evaluate astroBERT, we created the DEAL challenge (Detecting Entities in the Astrophysics Literature)
  - part of a workshop\(^1\) at the AACL-IJCNLP\(^2\) 2022 conference
  - challenged participants to build labeling systems
  - using our dataset of labeled entities
  - measured against BERT, SciBERT, and astroBERT baselines

Each baseline model was finetuned for DEAL i.e. the core language models were adapted and retrained for ~12 hours each.

---

\(^1\) WIESP: The first Workshop on Information Extraction from Scientific Publications
\(^2\) The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing
**astroBERT on DEAL**

*astroBERT outperforms BERT and SciBERT on DEAL*

<table>
<thead>
<tr>
<th>Model</th>
<th>Metric</th>
<th>MCC</th>
<th>overall F1 score</th>
<th>overall precision</th>
<th>overall recall</th>
<th>overall accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Split</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>train</td>
<td>0.1037</td>
<td>0.0170</td>
<td>0.0122</td>
<td>0.0278</td>
<td>0.7146</td>
</tr>
<tr>
<td></td>
<td>val</td>
<td>0.1083</td>
<td>0.0166</td>
<td>0.0119</td>
<td>0.0273</td>
<td>0.7059</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>0.1057</td>
<td>0.0162</td>
<td>0.0116</td>
<td>0.0269</td>
<td>0.6876</td>
</tr>
<tr>
<td>BERT</td>
<td>train</td>
<td>0.7542</td>
<td>0.4920</td>
<td>0.4995</td>
<td>0.4848</td>
<td>0.9256</td>
</tr>
<tr>
<td></td>
<td>val</td>
<td>0.7405</td>
<td>0.4739</td>
<td>0.4780</td>
<td>0.4698</td>
<td>0.9188</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>0.7229</td>
<td>0.4513</td>
<td>0.4622</td>
<td>0.4409</td>
<td>0.9094</td>
</tr>
<tr>
<td>SciBERT</td>
<td>train</td>
<td>0.8159</td>
<td>0.5867</td>
<td>0.5753</td>
<td>0.5986</td>
<td>0.9430</td>
</tr>
<tr>
<td></td>
<td>val</td>
<td>0.8019</td>
<td>0.5601</td>
<td>0.5463</td>
<td>0.5745</td>
<td>0.9366</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>0.7844</td>
<td>0.5355</td>
<td>0.5313</td>
<td>0.5398</td>
<td>0.9280</td>
</tr>
<tr>
<td>astroBERT (WIESP)</td>
<td>train</td>
<td>0.8296</td>
<td>0.6138</td>
<td>0.5889</td>
<td>0.6409</td>
<td>0.9468</td>
</tr>
<tr>
<td></td>
<td>val</td>
<td>0.8104</td>
<td>0.5779</td>
<td>0.5508</td>
<td>0.6077</td>
<td>0.9389</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>0.7939</td>
<td>0.5561</td>
<td>0.5387</td>
<td>0.5746</td>
<td>0.9308</td>
</tr>
<tr>
<td>astroBERT (public release)</td>
<td>train</td>
<td>0.8250</td>
<td>0.5995</td>
<td>0.5701</td>
<td>0.6319</td>
<td>0.9442</td>
</tr>
<tr>
<td></td>
<td>val</td>
<td>0.8194</td>
<td>0.5907</td>
<td>0.5575</td>
<td>0.6282</td>
<td>0.9405</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td><strong>0.8302</strong></td>
<td><strong>0.6093</strong></td>
<td><strong>0.5846</strong></td>
<td><strong>0.6362</strong></td>
<td><strong>0.9418</strong></td>
</tr>
</tbody>
</table>

**astroBERT improvement over SciBERT**

![Graph showing improvement](image)

Absolute Recall Improvement

Absolute Precision Improvement

Fraction in fulltext
Labeling Results

thank the anonymous referee for detailed comments, which improved this paper significantly. The authors thank Damien Coffey for critical reading of the manuscript, especially for improving the language. Jürgen Schmitt provided us unpublished light curves from their TRXS source variability analysis. Joachim Paul performed detective work in our old archives for hints on attitude errors and has set up the 2RXS webpage and catalogue interface. We appreciate discussions with Damien Coffey and Mara Salvato on positional offsets and source identification procedures. This research has made extensive use of the SIMBAD database and of the VizieR catalogue access tool, both operated at CDS, Strasbourg, France (see descriptions in Wenger et al. 2000 and Ochsenbein et al. 2000). This work would have been impossible without the old ROSAT Telescope staff (H/W + S/W), who are too numerous to mention.
Plans to improve astroBERT

Current astroBERT is good but not production ready

- Improve DEAL performance
  - by using ideas from WIESP: Conditional Random Fields (CRF)
  - with iterative training and labeling
- Improve core language model
  - by training with a new semi-supervised task: Semantic Textual Similarity (STS)
- Evaluate astroBERT on new downstream tasks: UAT Concept Extraction
astroBERT Recap

- We are building astroBERT, a core language model for astrophysics.
- We plan on using it internally for text enrichment, a task in which it already outperforms other language models.
- We publicly released both astroBERT and a dataset of enriched text to the astrophysics research community.