A Hybrid Approach to Labeling Datasets in Earth Science publications

Jacob Atkins
Irina Gerasimov
Mo Khayat
Code 610.2, NASA Goddard Space Flight Center, Greenbelt, MD, USA
with "Goddard Earth Sciences Data and Information Services Center (GES DISC)

Why this Project?

NASA GES DISC provides to the public over a thousand data collections. However, when scientists publish their work, these collections are not being cited with assigned Digital Object Identifiers (DOIs) preventing datasets from being found by an automated search. For example,

- Searching in Scopus, Google Scholar and DataCite with Microwave Limb Sounder (MLS/Aura) Carbon Monoxide dataset DOI returns only 4 publications
- We have already identified over 240 papers that referenced this dataset

Linking datasets with research publications based on these data collections enhances Findability, Accessibility, Interoperability, and Reusability (FAIR) principles in data management and stewardship

- Enables automated publication and dataset search
- Automated metric collection based on data use
- Allows users to easily find relevant datasets for their research
- Allows proper credit to be given to dataset creators

Example of Earth Science Dataset Referencing in Publication

Example: Arsenovic, P. (2019b, July 26). *Reactive nitrogen (NOy) and ozone responses to energetic electron precipitation during Southern Hemisphere winter.* https://doi.org/10.5194/acp-19-9485-2019

Mention: “We used two satellite datasets to evaluate our model results: MIPAS* for nitrogen species and the Microwave Limb Sounder for ozone”

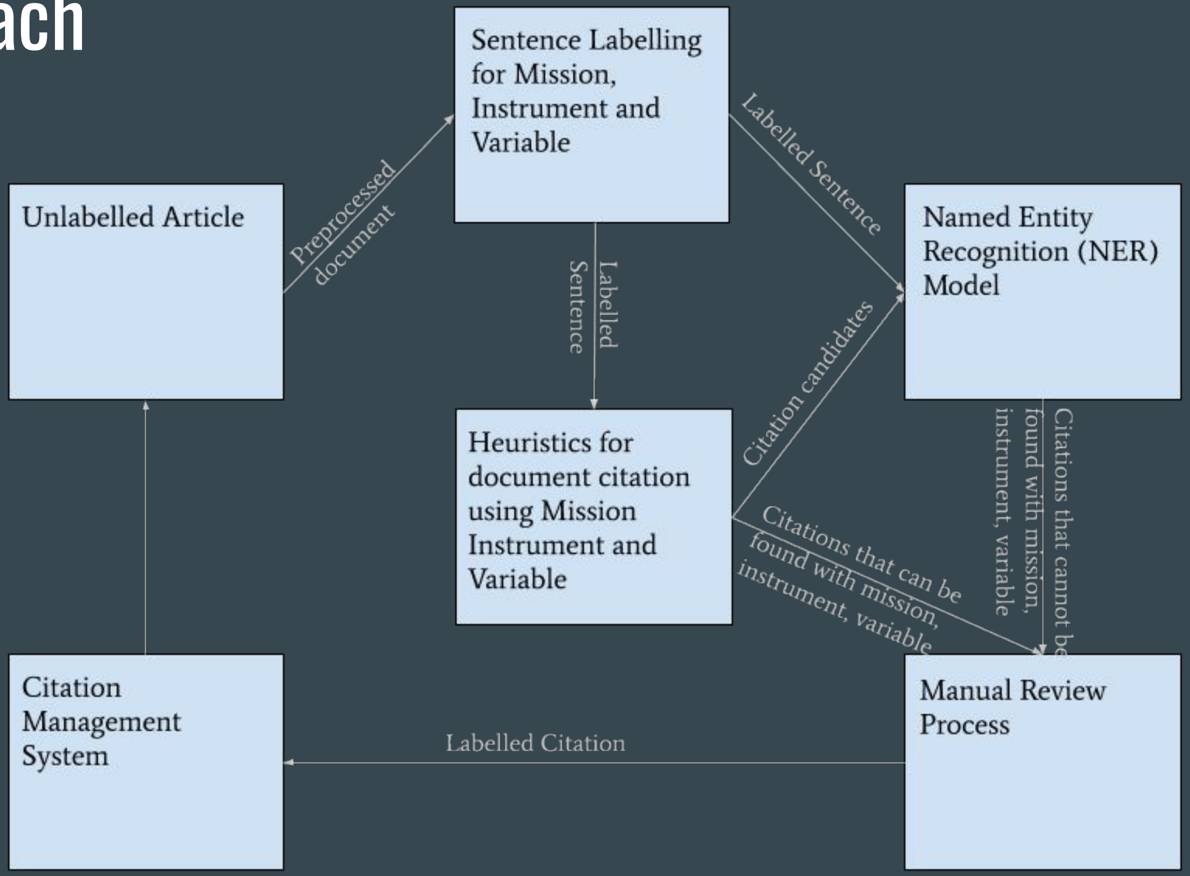
Proper DOI citation: Schwartz, M., Froidevaux, L., Livesey, N. and Read, W. (2015), MLS/Aura Level 2 Ozone (O3) Mixing Ratio V004, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: **[*Nov. 16 2020*]**, 10.5067/Aura/MLS/DATA2017

**Michelson Interferometer for Passive Atmospheric Sounding (MIPAS)*

Hybrid Approach Overview

- Label text sentences with the names of mission, instrument, re-analysis models, and science keywords taken from the Global Change Master Directory (GCMD) ontology.
- Based on the sentence labels, use heuristics to determine possible datasets or dataset groups:
  - Heuristics are successful at labeling datasets that do not require more information than mission, instrument, and variable names.
  - Heuristics are not sufficient to distinguish between datasets with different temporal or spatial resolution, processing level, and other product specifics.
- If needed, pass heuristics output to the Named Entity Recognition (NER) Natural Language Processing (NLP) model
- Use NER to generate possible dataset names and their probability.
- Manually review heuristics or NER outputs to correct labeling errors.
- Use newly-labeled citations as additional training data.

Hybrid Approach



Unlabelled Article

Preprocessed document

Sentence Labelling for Mission, Instrument and Variable

Labelled Sentence

Named Entity Recognition (NER) Model

Labelled Sentence

Heuristics for document citation using Mission Instrument and Variable

Citation candidates

Citations that can be found with mission, instrument, variable

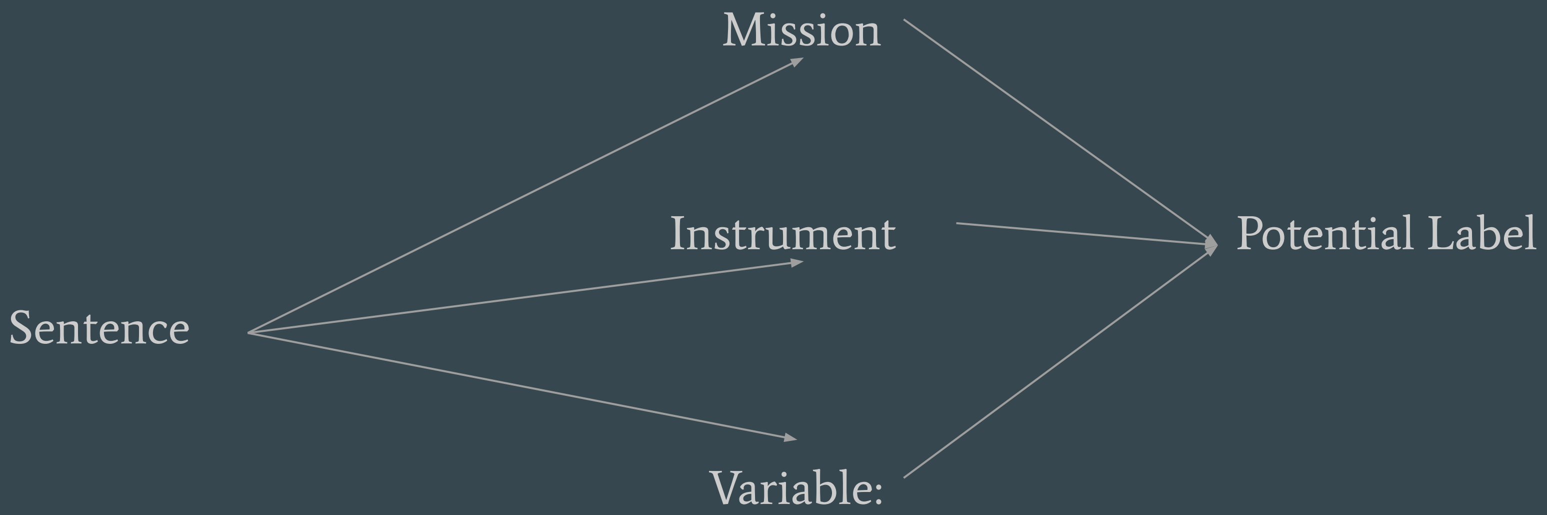
Citations that cannot be found with mission, instrument, variable

Manual Review Process

Labelled Citation

Citation Management System

Diagram demonstrating labelling process



Mission

Instrument

Sentence

Potential Label

Variable:

Example Sentence: “Tropospheric O3 data are obtained from combined observations of two satellite instruments, the Ozone Monitoring Instrument (OMI) and Microwave Limb Sounder (MLS)”

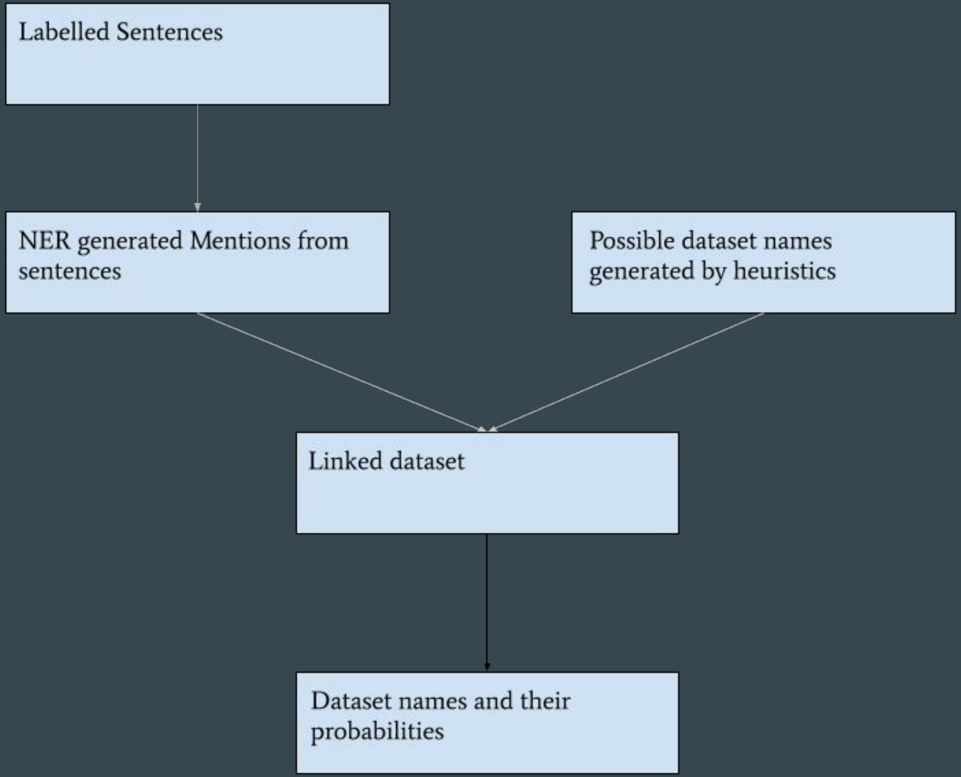
Potential Labels:

- Mission: Aura, Instrument: MLS, Variable: O3, → Label: ML2O3
- Mission: Aura, Instrument: OMI, Variable: O3 → Label is uncertain because the processing level and other dataset specifics are not stated. Label: (Aura/OMI, O3)

Heuristic Search

- Using GCMD ontology check sentence for exact match for mission, instrument, reanalysis model and variable names
- Check if the mission and instrument pair has associated datasets
- Retain only the sentences containing an exact match and provide potential labels

Named Entity Recognition (NER)


Labelled Sentences
NER generated Mentions from sentences
Possible dataset names generated by heuristics
Linked dataset
Dataset names and their probabilities

- The NER model was taken from the Allen Institute for AI's open source collection of NLP Models
- https://github.com/allenai/allennlp

Manual Review

- Each potential dataset label and the mentions are added to the citation management system for manual review
- The extracted mentions and labels are reviewed by domain experts
- The labelled data can then be used as training data in the next iteration

Results

MLS/Aura (heuristic-only approach):

- Precision : .68
- Recall : .86
- F1 : .76
- 110 labelled papers were processed, of those 18 papers were predicted exactly right and 54 additional papers had predictions where the recall was 1

NER results (Hybrid approach):

- Precision: .85
- Recall: .25
- F1: .39

Example of the problem with Precision:

"We used two satellite datasets to evaluate our model results: MIPAS for nitrogen species and the Microwave Limb Sounder for ozone"

Conclusions

- Heuristics-only approach works well to label datasets that are defined by only a mission, instrument, and variable
- NLP approach is needed for datasets that require additional info to be labelled
- Heuristics-only approach provides good recall but mediocre precision
- The NER model has good precision but bad recall, as it doesn’t find every entity but the entities it does find are correct
- The hybrid approach combines heuristics and NLP methods together by feeding the labelled data from the heuristics module into the NER model. This makes NER’s search space mostly relevant.
- The process should iteratively improve with each new set of labelling