An Automated Approach to Labelling Datasets in Earth Science Publications

ESIP Summer 2021
July 19-23, 2021
ejahoda@gatech.edu
Irina.Gerasimov@nasa.gov

https://disc.gsfc.nasa.gov/ NASA/Goddard Earth Sciences Data and Information Services Center (GES DISC)

Edward Jahoda[1], Irina Gerasimov[1,2], Mohammed Khayat[1,2], and Jennifer Wei[1]
[1]Code 610.2, NASA Goddard Space Flight Center, Greenbelt, MD, USA [2]ADNET Systems Inc., Lanham, MD, USA

Importance of Labelling Datasets

- NASA Earth Science data archives contain over 8,000 publicly available data collections which are consistently used in novel research applications.
- These datasets, especially in older publications, are not cited despite having a unique DOI.
- Similar to linking journal articles to other journal articles, linking datasets to the publications they are used in yields numerous benefits:
  - Dataset science impact metrics.
  - Credit to the dataset creators.
  - Dataset usage-based discovery.
  - Dataset recommendation.
  - Reproducibility.
- Manually reviewing publications to identify the datasets is time consuming and requires a subject matter expert and does not generalize to unseen publications

Challenges for Automated Dataset Detection

- How many datasets are in the publication is unknown (there might not even be any)
- Authors often spend a lot of time discussing datasets they did not use
- “Semantic Gymnastics” and interwoven ideas among sentences/paragraphs; Ideas are incomplete without all of the sentences

Automated Extraction of Science Keywords from Research Publications

Keyword Extraction Process

1. Convert Journal Article to text file using Cermine NLP Package
2. Remove sections from Journal Article that are likely to introduce a lot of noise (ie: Introduction, References)
3. Extract sentences with Science Keywords (platform, instrument, variables, authors, processing levels, spatial resolution, temporal resolutions)

Detecting Science Keywords (Simple Cases)

- GES DISC dataset metadata includes the short name and long name for the platform, instrument, and variables of each data
- Simple string matching approach for Short and Long Names

Detecting Science Keywords (Complex Cases)

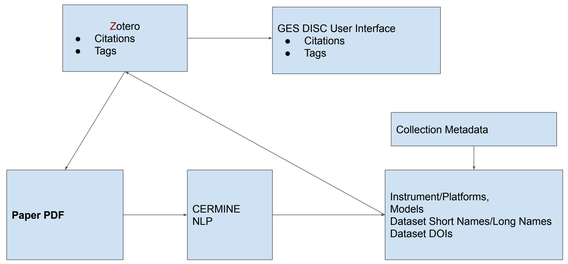
- Some science keywords like spatial resolution are referenced in many different ways (“vertical resolution *5 km*”, 'horizontal resolution (*40 km x 320 km*)”)
- Regular Expression patterns crafted based on observed occurrences detected many instances of spatial resolutions.

An Automated Pipeline for Detecting and Linking Explicit Dataset Mentions

Locating Explicit Dataset Mentions

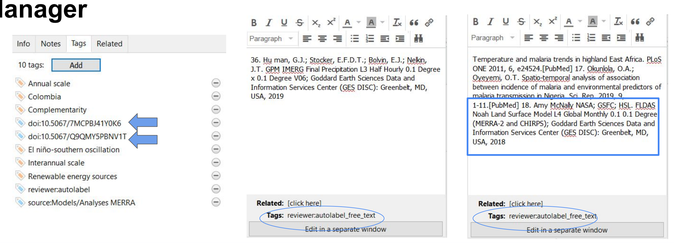
- Looking for instances of
  - Explicit Short Name
  - Explicit Long Name
  - Explicit DOI
  - References to ‘GES DISC’ or the GES DISC Website

Automated Feature Extraction Pipeline



- Automated Pipeline Results from a Newer Collection of Papers collected from Google Scholar
  - 209 Papers from 2020-2021 (197 successfully converted to text files)
  - 59% had an explicit citation:
    - 20% contained an explicit DOIs
    - 50% contained either an explicit DOI or an explicit ShortName
    - 14% contained references to ‘GES DISC’ or the GES DISC Website
  - Dataset Source (Platform/Instrument or Model) tagging
    - 73% tagged with at least one source
    - In the 197 Papers
      - 154 Platform/Instrument Couples
      - 180 Models/Analyses

Automatically Entering Explicit Dataset Citations into Citation Manager

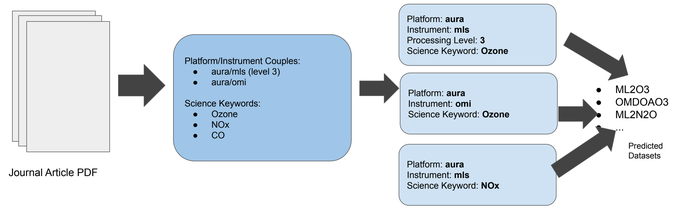


Future Work

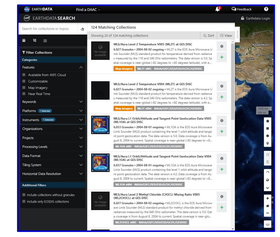
- Using the extracted science keywords, build a Knowledge Graph connecting the journal articles and their science keywords.
- Use this graph to make predictions about the datasets used in the journal articles.

Implicit Datasets: A CMR Based Approach

- https://search.earthdata.nasa.gov/search



Aura/Microwave Limb Sounder Collection (363 Papers)	First Result Returned from CMR
Datasets Correctly Cited	509
False Negatives	565
False Positives	603



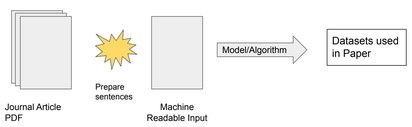
Earth Data Search Website

Analysis of CMR Results

- CMR changes, so results and overall effectiveness may change over time
- CMR returns datasets based on ease of use and popularity, so some datasets much more likely to be near the top of the results
- Cannot use all the features that we extract
  - Ie: GES DISC spatial resolutions are not included inside of CMR
  - Temporal resolutions in CMR require start/end dates, so just knowing the period is insufficient
  - CMR does not have a facet for dataset creator last name

Implicit Datasets: A Supervised Machine Learning Approach

- Extract all sentences that mention a platform, instrument, and science keyword
- Prepare the sentences for machine learning (NLP Techniques)
  - Bag of Words
  - TF-IDF
  - Doc2Vec (extension of Word2Vec)
- Run Machine Learning Models and Evaluate the Results
  - Decision Trees/Random Forests/Neural Networks



ML Lessons

- Multiclass classification with limited amount of training data is difficult to learn relationships from
- Sentence Structure
  - Complex structure of sentences
    - We use p1 with i1 and p2 with i1 to study sk1 and sk2. Additionally, we also looked at sk3 from the same source.
  - Lots of noise
    - We used p1 and [lots of words unrelated to datasets] to study sk1 and [more words] i1
- Imbalance of datasets led to some models to just defaulting to predicting the most common datasets or No datasets at all