GCMD Keyword Classification using Neural Networks

Hassan Muhammad$^1$, Jeanne le Roux$^1$ *, Muthukumaran Ramasubramanian$^1$, Iksha Gurung$^1$, Manil Maskey$^2$, and Rahul Ramachandran$^2$

(1) University of Alabama in Huntsville, Huntsville, AL, United States
(2) NASA Marshall Space Flight Center, Huntsville, AL, United States

13 December 2019
AGU Fall Meeting
NASA Earth Science

- NASA has an extensive collection of Earth science data obtained from various sources including satellites, aircraft, field measurements, and model derived data.

- NASA’s Earth Observing System Data and Information System (EOSDIS) provides end-to-end capabilities for managing NASA Earth science data.

- Started in the 1990’s, EOSDIS today has 10,000+ data collections (or datasets).
NASA’s growing collection of Earth science datasets are described by metadata records stored in a database called the Common Metadata Repository (CMR).

Users can search the CMR directly via an API to find and access Earth science data.

The CMR also serves as the backend for search interfaces such as the Earthdata Search Client.

What makes finding data possible?

- Metadata -

How?
- Metadata acts as a proxy for data
- Limits & focuses attention to the relevant information about a dataset
- Contains information indexed for search
Metadata of Focus

This research utilizes the following pieces of metadata:

1. **Title**: formal title of the dataset.

2. **Abstract**: a brief but comprehensive description of the dataset. Comparable to a journal article abstract.

3. **Science Keywords**: controlled vocabulary field containing keywords relevant to the scientific purpose/content of the dataset. Science keywords aid in data discovery. For example, the Earthdata Search Client uses science keywords for:
   - faceted search
   - search relevancy rankings

The CMR leverages the Global Change Master Directory (GCMD) science keyword taxonomy, which is a hierarchical set of controlled science keywords.
Assigning science keywords is currently a manual process, which is prone to human error and inconsistencies:

- Metadata managed across a network of multiple data centers (i.e. keywords not assigned by a central entity)
- Keywords may be assigned by non-subject matter experts (SMEs)
- SMEs assigning keywords may not be familiar with GCMD keywords or how keywords are used in search engines
- Science keywords are meant for data discovery across a broad range of users and may not encompass highly specific scientific variables
Broader question:
Can machine learning be leveraged to accurately assign science keywords to metadata records in an automated, objective, and consistent manner?

Starting point:
Can we train a machine learning model to accurately assign science keywords based on the dataset abstract?
Version 7.3 is the current version of the data set. Version 3.5 is no longer available and has been superseded by Version 7.3. This data set is currently provided by the OCO (Orbiting Carbon Observatory) Project. In expectation of the OCO-2 launch, the algorithm was developed by the Atmospheric CO2 Observations from Space (ACOS) Task as a preparatory project, using GOSAT TANSO-FTS spectra. After the OCO-2 launch, “ACOS” data are still produced and improved, using approaches applied to the OCO-2 spectra. The “ACOS” data set contains Carbon Dioxide (CO2) column averaged dry air mole fraction for all soundings for which retrieval was attempted. These are the highest-level products made available by the OCO Project, using TANSO-FTS spectral radiances, and algorithm build version 7.3. The GOSAT team at JAXA produces GOSAT TANSO-FTS Level 1B (L1B) data products for internal

<table>
<thead>
<tr>
<th>Predicted Keyword</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>carbon dioxide</td>
<td>0.4513424</td>
</tr>
<tr>
<td>land use/land cover classification</td>
<td>0.3825603</td>
</tr>
<tr>
<td>terrain elevation</td>
<td>0.1924277</td>
</tr>
<tr>
<td>barometric altitude</td>
<td>0.18085223</td>
</tr>
<tr>
<td>carbon and hydrocarbon compounds</td>
<td>0.07634798</td>
</tr>
</tbody>
</table>
Word Embeddings

- Numerical representation of text
- Maps words or phrases from the vocabulary to vectors of real numbers
- Can be used for:
  - compact and machine readable representation of words in a corpus
  - performing statistical analysis on corpus
  - embedding text as input into ML models
- Types of word embeddings:
  - Count Vectors (CV)
  - Tf-idf Vectors (TF)
  - Word2Vec (W2V)
  - Latent Dirichlet Allocation (LDA)
- Utilized Word2Vec: trained on a corpus of ~23,000 Earth science journal articles
Classifier: Neural Network

- Computing systems inspired by biological neural networks
- Learns features or patterns from provided examples without specifying the rules explicitly
- Learns by propagating back errors from predictions and updates the weights between layers
- Used as classifiers, regression models, etc.
• Trained on a total of 2,598 abstracts and keywords from the following NASA data centers:
  ○ GHRC
  ○ ORNL
  ○ GODDARD
• Training split: 2078, validation split: 520
• Total number of unique science keywords: 274
• Validation accuracy: 75.62%

\[ X' = [x_i > \text{threshold}], i: 1 \ldots n \]

\[ \text{Score} = \frac{\sum (X' * X)}{\sum X} \]

Note: Accuracy score is defined as the percentage of the number of keywords correctly predicted over the total number of true keywords.
GCMD Classifier Tool

- Tool input options:
  1. User selects metadata record from CMR (abstract is automatically extracted from the record)
  2. User provides abstract text
- Abstract is passed through the word2vec to get word embeddings
- Embeddings are passed to the classifier to get keyword predictions
- Any keyword predictions with score higher than 0.15 is displayed as recommendation for the given abstract
Tool Demo: https://gcmd.nasa-impact.net/
Discussion & Future Work

Limitations:
• Results depend on the quality of the abstracts and word embeddings.
• Limited training set (i.e. model is tuned to work well only with the metadata from select data centers)
• Assumes abstracts in training set are of high quality and that science keywords in training set are accurately assigned

Future Work:
• Include metadata from all data centers into the training set for a more generic model
• Investigate other classifiers
• Improve Word2Vec implementation with increased corpus size
• Automate algorithm training using feedback provided from the user interface
References


Questions?

- Website is publicly accessible: https://gcmd.nasa-impact.net/
- Plans to open source through NASA in 2020

Thank you!

For further inquiries please contact:
Iksha Gurung (ig0004@uah.edu)
Muthukumaran R. (mr0051@uah.edu)
Appendix 1: Text Preprocessing Steps

Preprocessing Steps
1. Stemming: Root word extraction “Stemming to Stem”
2. Stop word removal: remove words such as “a, the, an …”
3. Lemmatization: Contextual root word extraction “Lemmatization to Lemma”

Word2Vec Training
Trained on a corpus of ~23,000 Earth science journals
Embedding size: 150
Vocab size (after pre-processing): ~183,000
Appendix 2: Embedders and Classifiers Tested

- **Embedders**
  - Count Vectors from Descriptions (CV)
  - Tf-idf Vectors from Descriptions (TF)
  - Word2Vec built from science corpus (W2V)
  - Latent Dirichlet Allocation from corpus (LDA)

- **Classifiers**
  - Multi-class Logistic regression (LR)
  - Random Forest Classifier (RF)
  - Fully Connected Neural Network (FCNN)
  - LSTM Neural Network (LSTM)
  - 1D Convolutional Neural Network (1DC)
  - Labelled LDA (LLDA)
Appendix 3: Classifier Comparison
Global Change Master Directory (GCMD) Keywords

- Hierarchical set of controlled Earth science keywords
- Help ensure Earth science data, services, and variables are described in a consistent and comprehensive manner
- Facilitate data search and discovery
- Periodically analyzed and expanded/updated for relevancy in response to user needs
- Science keywords are manually assigned to metadata records

The CMR leverages the GCMD science keyword taxonomy.