ES2Vec: Earth Science Metadata Keyword Assignment using Domain-Specific Word Embeddings

Muthukumaran Ramasubramanian[1], Hassan Muhammad[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

NASA's Earth Science Data Catalog

- 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 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

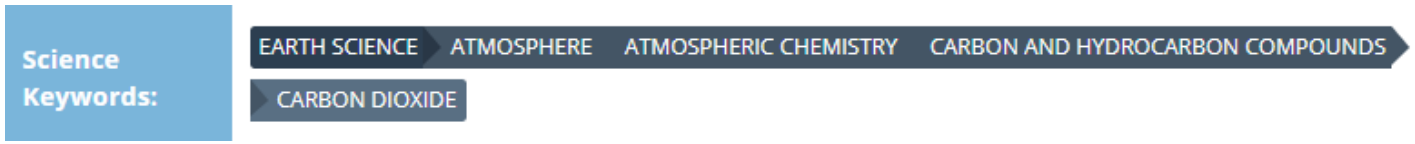
This research utilizes the following pieces of metadata:

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.



Problem Space

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

Research Question

Broader question:

Can machine learning be leveraged to accurately assign science keywords to metadata records in an automated, objective, and consistent manner?

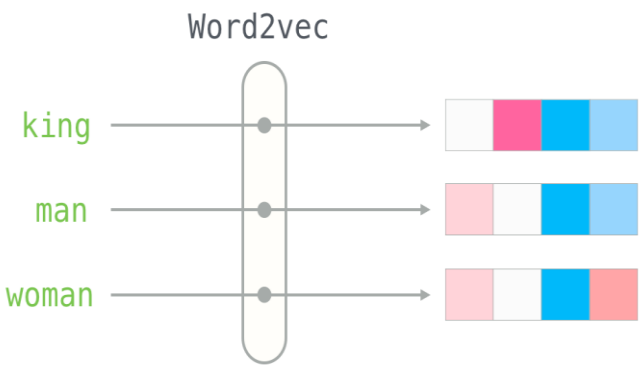
Starting point:

Can we train a machine learning model to accurately assign science keywords based on the dataset abstract?

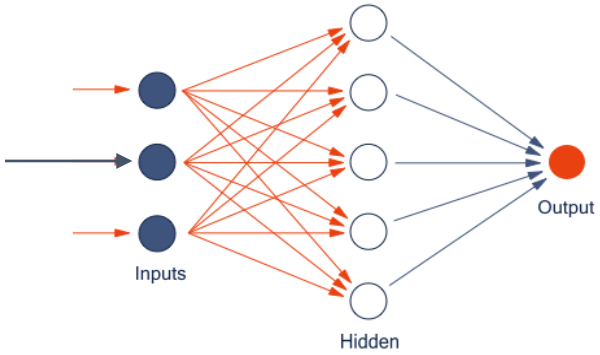
Approach

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



Word Embeddings



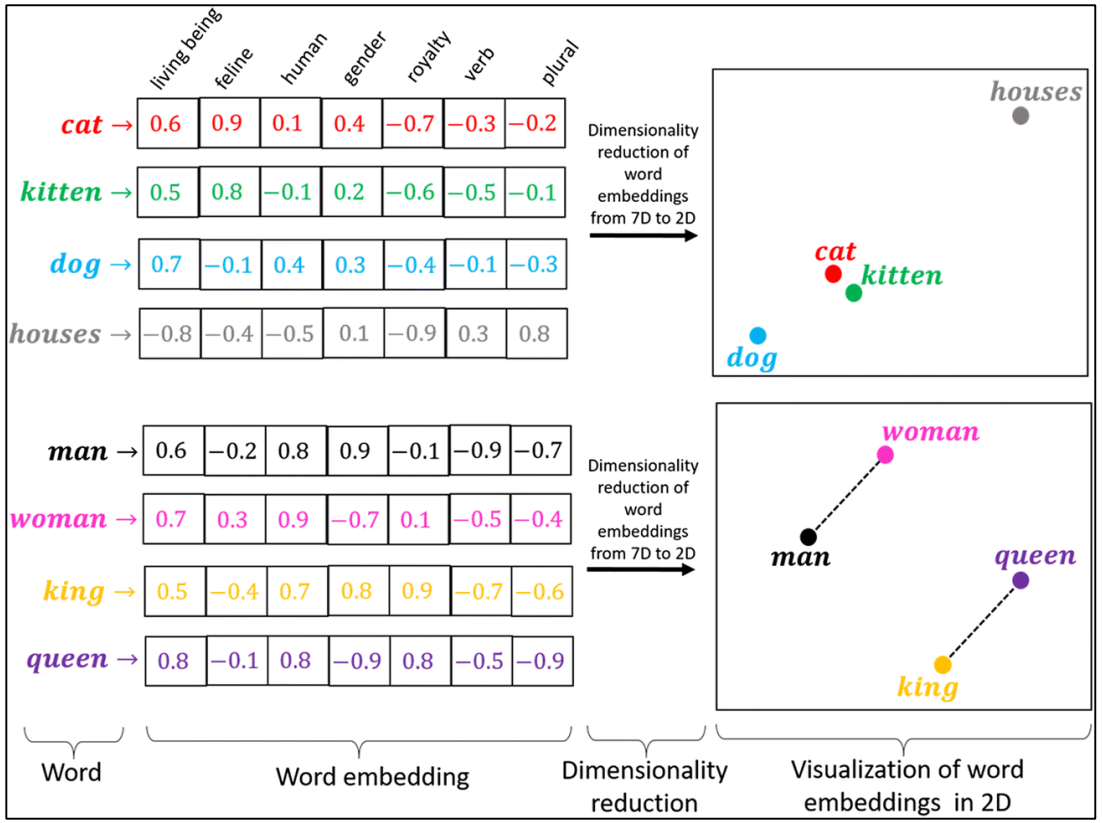
Classifier

Predicted Keyword	Score
carbon dioxide	0.4513424
land use/land cover classification	0.3825603
terrain elevation	0.1924277
barometric altitude	0.18085223
carbon and hydrocarbon compounds	0.07634798

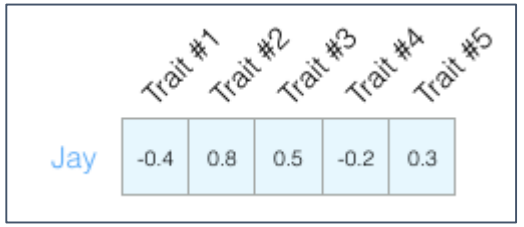
Suggested Keywords

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
- **ES2Vec: Domain Specific Word2Vec, trained on a corpus of ~23,000 Earth science journal articles (AGU)**
- **Preprocessing on Corpus: Stemming, Lemmatization, Stopword removal, Phrase Retention**



An example of word embeddings [1]



In this example each trait is one of the 'Big Five' personality traits (openness to experience, Conscientiousness, extraversion, agreeableness, neuroticism)

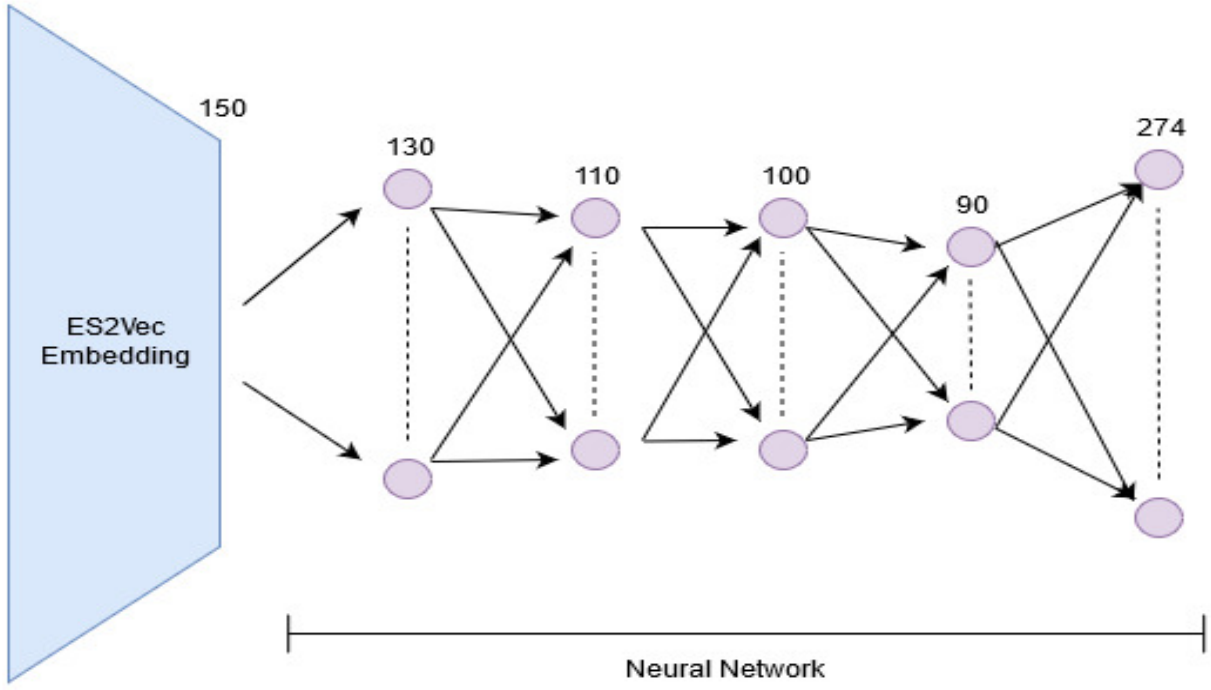
Classifier

Training data Labels: Each dataset has human assigned earth science keywords, which is converted to one hot vectors

	Keywords		
Abstract 1	Forest	Biomass	Vegetation
Abstract 2	Ozone	Radiation	Atmosphere
Abstract 3	Aerosols	Atmosphere	

	One-hot Encoding							
Keyword	Aerosols	Ozone	Forest	Biomass	...	Radiation	Vegetation	Atmosphere
Abstract 1	0	0	1	1	...	0	1	0
Abstract 2	0	1	0	0	...	1	0	1
Abstract 3	1	0	0	0	...	0	0	1

Classifier: A Neural Network 4 Fully connected layers with word embeddings as input and one hot keywords as output



Classifier Training & Accuracy Metric

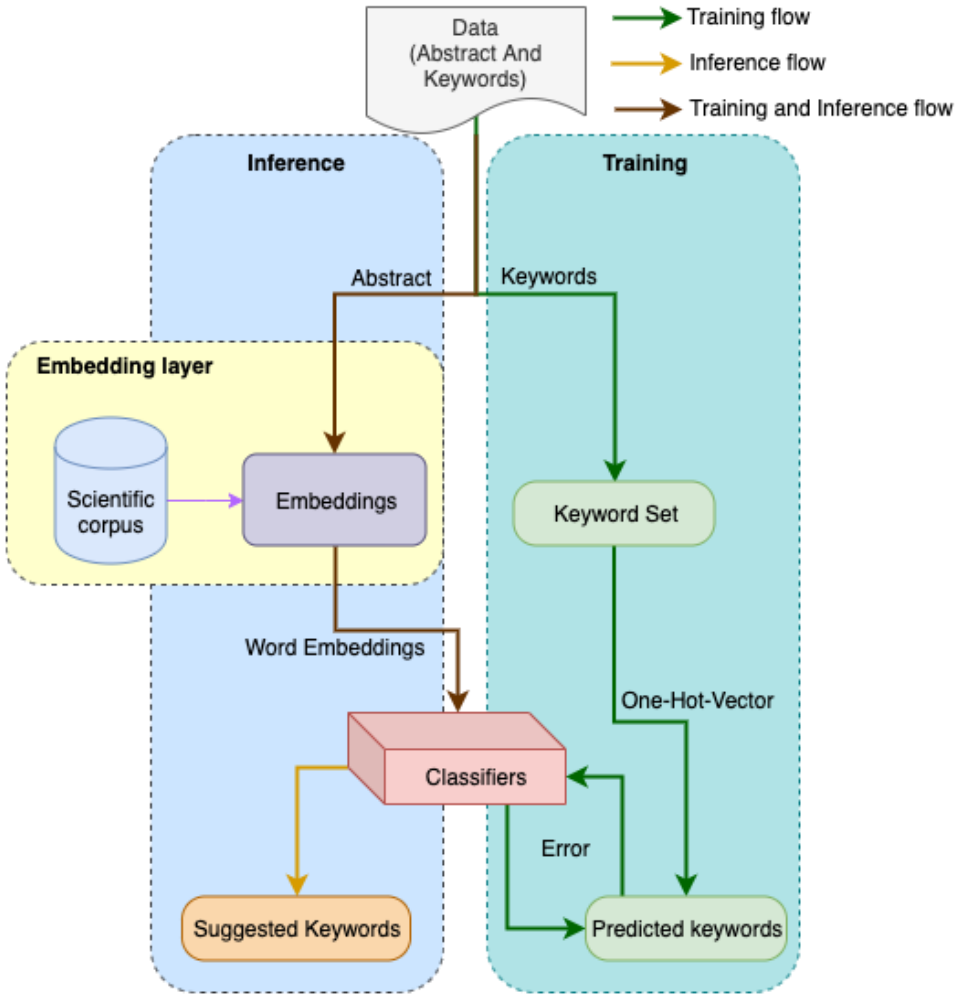
- Trained on a total of 2,598 abstracts and keywords from the following NASA data centers:
  - GHRC
  - ORNL
  - GODDARD
- Training split: 2078, validation split: 520
- Total number of unique science keywords: 274

Accuracy score is defined as the percentage of the number of keywords correctly predicted over the total number of true keywords. Threshold = 0.15, Empirically chosen

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

$$Score = \sum(X^{'} * X) / \sum X$$

GCMD Classifier Tool



Tool Pipeline

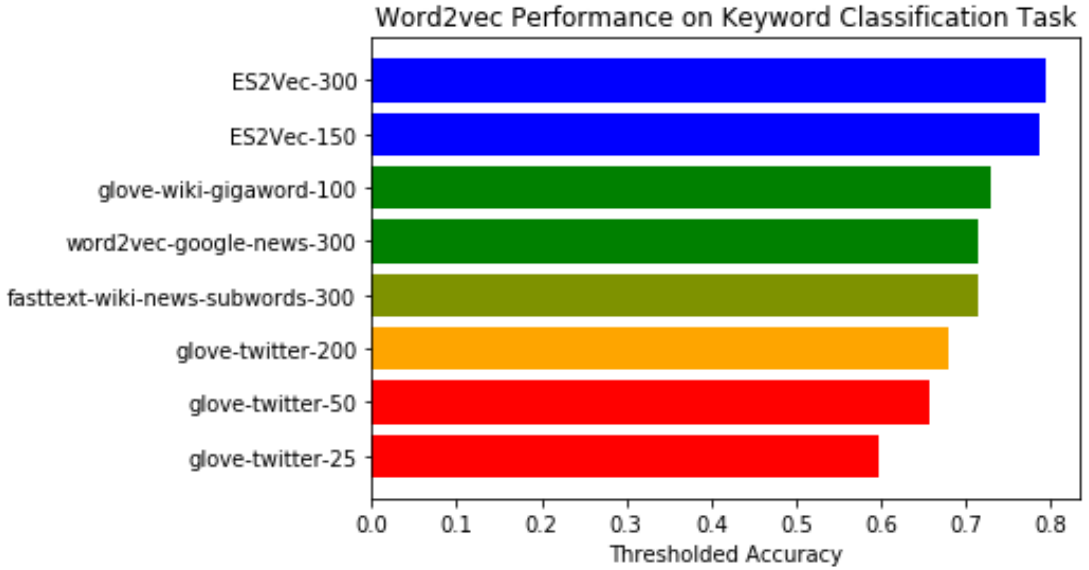
- Tool input options:
  - 1. User selects metadata record from CMR (abstract is automatically extracted from the record)
  - 2. User provides abstract text
- Abstract is passed through the word2vec to get word embeddings
- Embeddings are passed to the classifier to get keyword predictions
- Any keyword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

ES2Vec Performance

Word2Vec model	Embedding size	Vocabulary size	No. of Tokens	Accuracy
glove-twitter-25	25	1.2M	27B	0.596
glove-twitter-50	50	1.2M	27B	0.657
glove-twitter-200	200	1.2M	27B	0.679
fasttext-wiki-news-subwords-300	300	1M	16B	0.713
Word2Vec-google-news-300	300	3M	100B	0.714
glove-wiki-gigaword-100	100	400k	6B	0.728
ES2Vec-150	**150**	**500k**	**115M**	**0.786**
ES2Vec-300	**300**	**500k**	**115M**	**0.794**

TABLE I
WORD2VEC MODEL PERFORMANCE FOR KEYWORD CLASSIFICATION TASK

ES2Vec better performing compared to other generic models in Earth Science keyword classification task



Relatively less vocabulary size and number of tokens, Could be improved further

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

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:
Muthukumaran R. (mr0051@uah.edu)
Iksha Gurung (ig0004@uah.edu)