

ES2Vec: Earth Science Metadata Suggestions and Analogical Reasoning



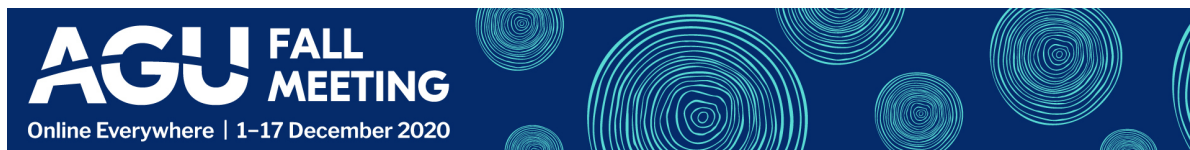
We're sorry...

An error has occurred and our technicians have been informed.
Please return to the previous page and try again in a short while. If the problem persists, contact support.
[Click here to go back to previous page](#)

Carson Davis, Muthukumaran Ramasubramanian, Derek Koehl, Ashish Acharya, Iksha Gurung, Manil Maskey, Rahul Ramachandran, Udaysankar Nair

Manufacturing Technical Solutions, University of Alabama in Huntsville, NASA MSFC

PRESENTED AT:



INTRODUCTION

As the volume of text-based Earth science research grows, it is increasingly possible to discover latent relationships in the literature. However, traditional methodologies are restricted by limited computational capabilities and intractable problem spaces. Advancements in natural language processing (NLP) have allowed us to use an extensive Earth science corpus to create a domain-specific word vector model, Es2Vec, which we have used to surface latent relationships between Earth science concepts and generate improved keyword tags.

Earth science metadata keyword assignment is a challenging problem. Dataset curators select appropriate keywords from the Global Change Master Directory (GCMD) set of keywords. The keywords are an integral part of the search and discovery of these datasets. Hence, the selection of keywords is crucial to increasing the discoverability of datasets. Utilizing machine learning techniques, we provide users with automated keyword suggestions to complement manual selection. We trained a machine learning model that leverages the semantic embedding ability of Word2Vec models to process abstracts and suggest relevant keywords. A user interface tool we built to assist data curators in the assignment of such keywords is also described.

WORD EMBEDDINGS

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 model

Word2Vec

- Two layer neural network to learn word associations from a large corpus of text.
- Validated by the presence of semantic and syntactic patterns that can be reproduced using vector arithmetic. Patterns such as “Man is to Woman as King is to Queen” can be generated through algebraic operations on the vector representations of these words such that the vector representation of “King” - “Man” + “Woman” produces a result which is closest to the vector representation of “Queen”

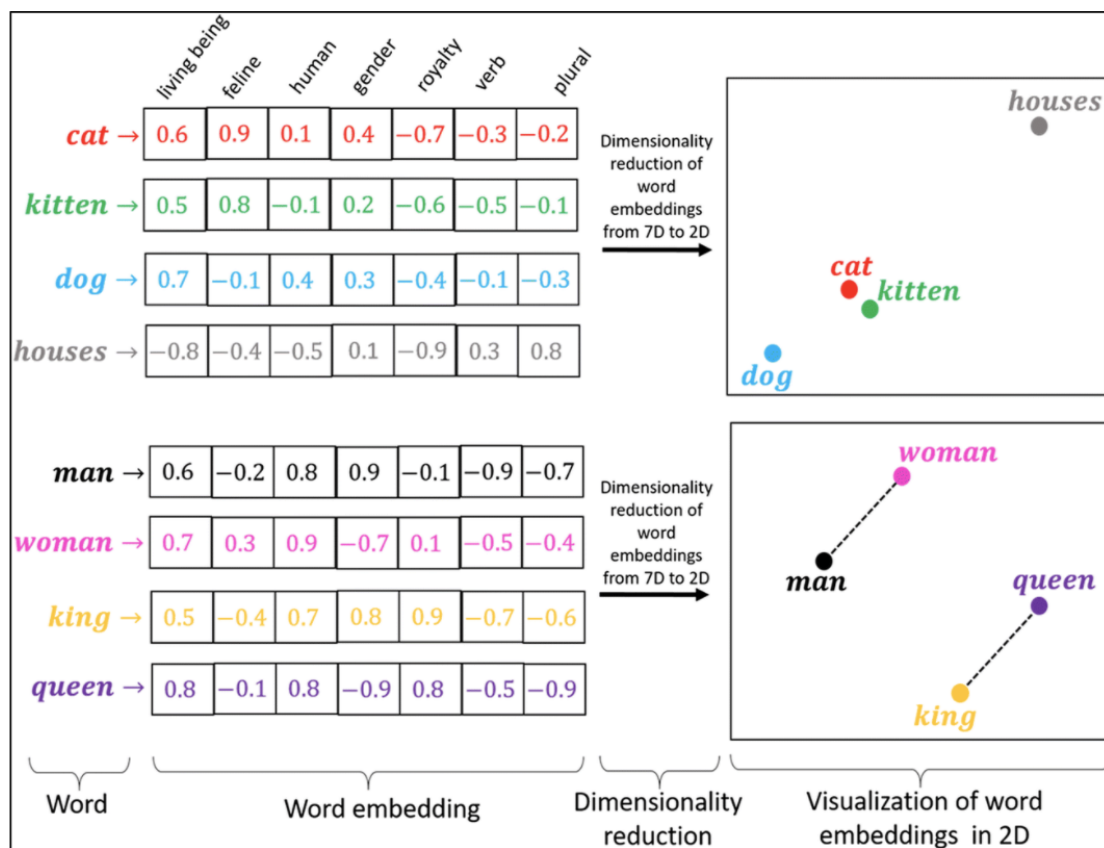


Fig. 1. Word embeddings and word pairs

ES2VEC: EARTH SCIENCE ANALOGICAL REASONING

Es2Vec: Domain Specific Word Embeddings

- Built using a large corpus of domain-specific text documents.
- Our approach utilized a corpus of ~75,000 Earth science journal articles from AGU.
- Trained two models :
 - ES_1 (150 embedding size)
 - ES_3: (300 embedding size)

Measuring Effectiveness of Word Embeddings

- Effectiveness of domain-specific word embeddings can be measured using presence and quality of analogy pairs surfacable by the embeddings

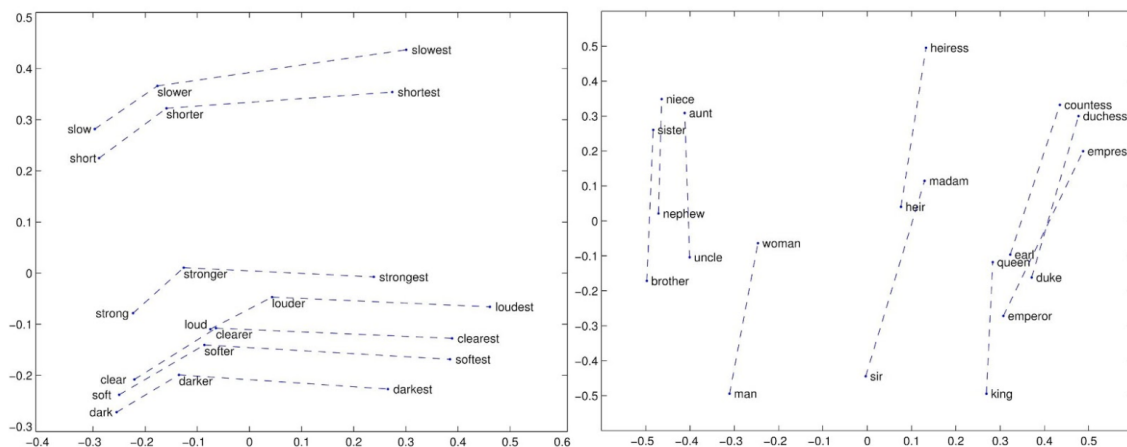


Fig. 2. Word Pairs

- The exploratory set of Earth science analogies categorized using a classification framework similar to the one employed by A. Gladkova et al., 2016.
- Laying the groundwork for an Earth science analogy test set

Analogy Category	Number
Lexicographical	12
Meronym (<i>graupel:ice</i>)	4
Member (<i>potential temperature:temperature</i>)	3
Hypernym/Hyponym (<i>stratus:low-level cloud</i>)	3
Part-whole (<i>cyclostrophic:centrifugal</i>)	1
Gradable (<i>blue:ultraviolet</i>)	1
Encyclopedic	16
Phenomenon-effect (<i>anticyclonic:divergence</i>)	9
Compound-effect (<i>sulfate:scattering</i>)	4
Property-instrument (<i>temperature:thermometer</i>)	2
Compound-property (<i>carbon dioxide:acidic</i>)	1

Analogy Category	Accuracy
ES_3: lexicographical and encyclopedic	0.43
ES_1: lexicographical and encyclopedic	0.29
GloVe general: lexicographical and encyclopedic	0.21
SVD general: lexicographical and encyclopedic	0.17
ES_3: lexicographical	0.50
ES_1: lexicographical	0.25
GloVe general: lexicographical	0.11
SVD general: lexicographical	0.11
ES_3: encyclopedic	0.38
ES_1: encyclopedic	0.32
GloVe general: encyclopedic	0.32
SVD general: encyclopedic	0.20

Fig. 3. Analogy Categories and Accuracy of models in surfacing the word pairs

- ES_3 and ES_1 were shown to be capture the most Earth science analogies.

USING ES2VEC IN EARTH SCIENCE METADATA KEYWORD SUGGESTIONS

Task: Given descriptions of datasets, use Es2Vec and neural networks to provide appropriate keyword suggestions.

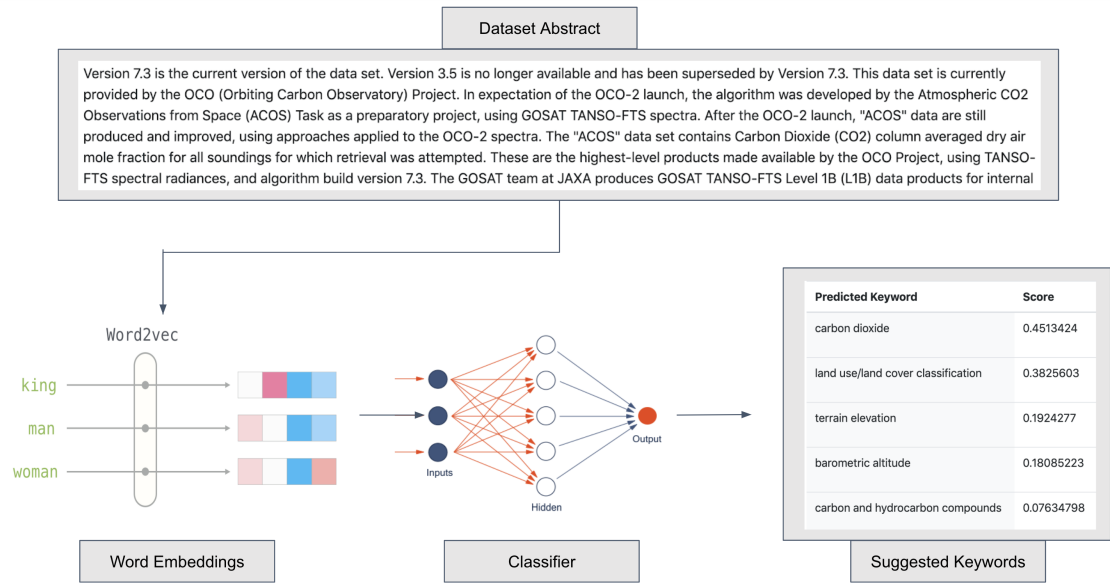


Fig. 4. Architecture

- ES2Vec is used to extract word embeddings from the abstracts.
- ES2Vec is trained separately and used as a black box during both the training and inference cycles.
- The activations (embeddings) from word2vec are sent to a trainable neural network classifier to predict keywords
- Predicted keywords in combination with the extracted keywords are used to calculate the error. The error is then used to update the weights between the layers in the classifier.

Model Diagram

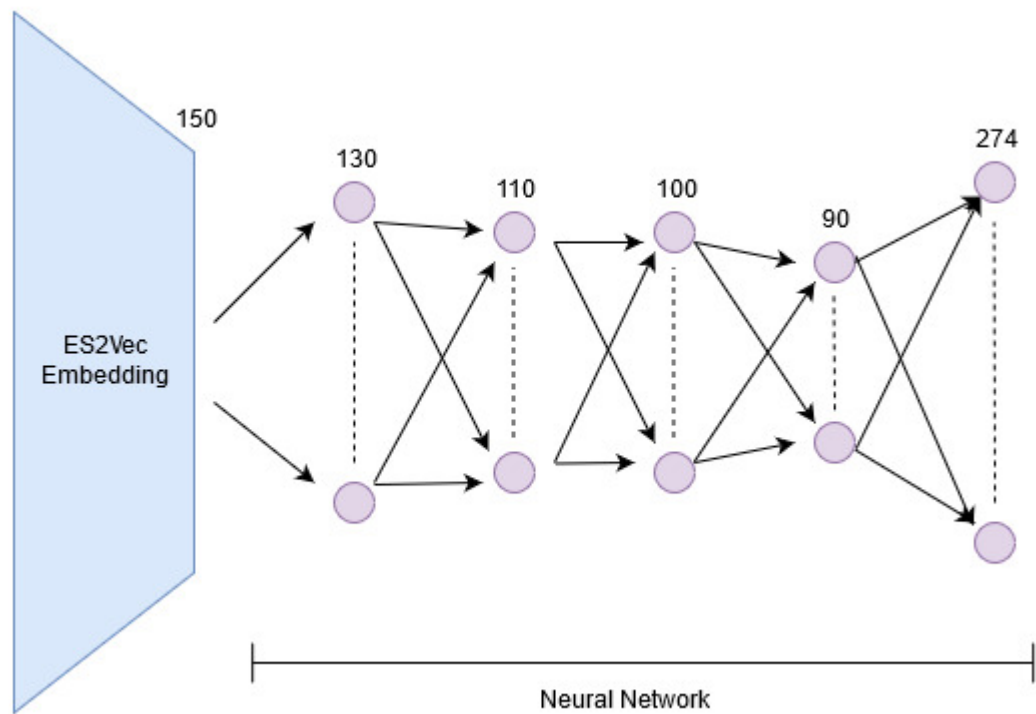


Fig. 5. Model Architecture

- Training split: 2078, validation split: 520
- Total number of unique science keywords: 274

Tool:

- **Workflow**

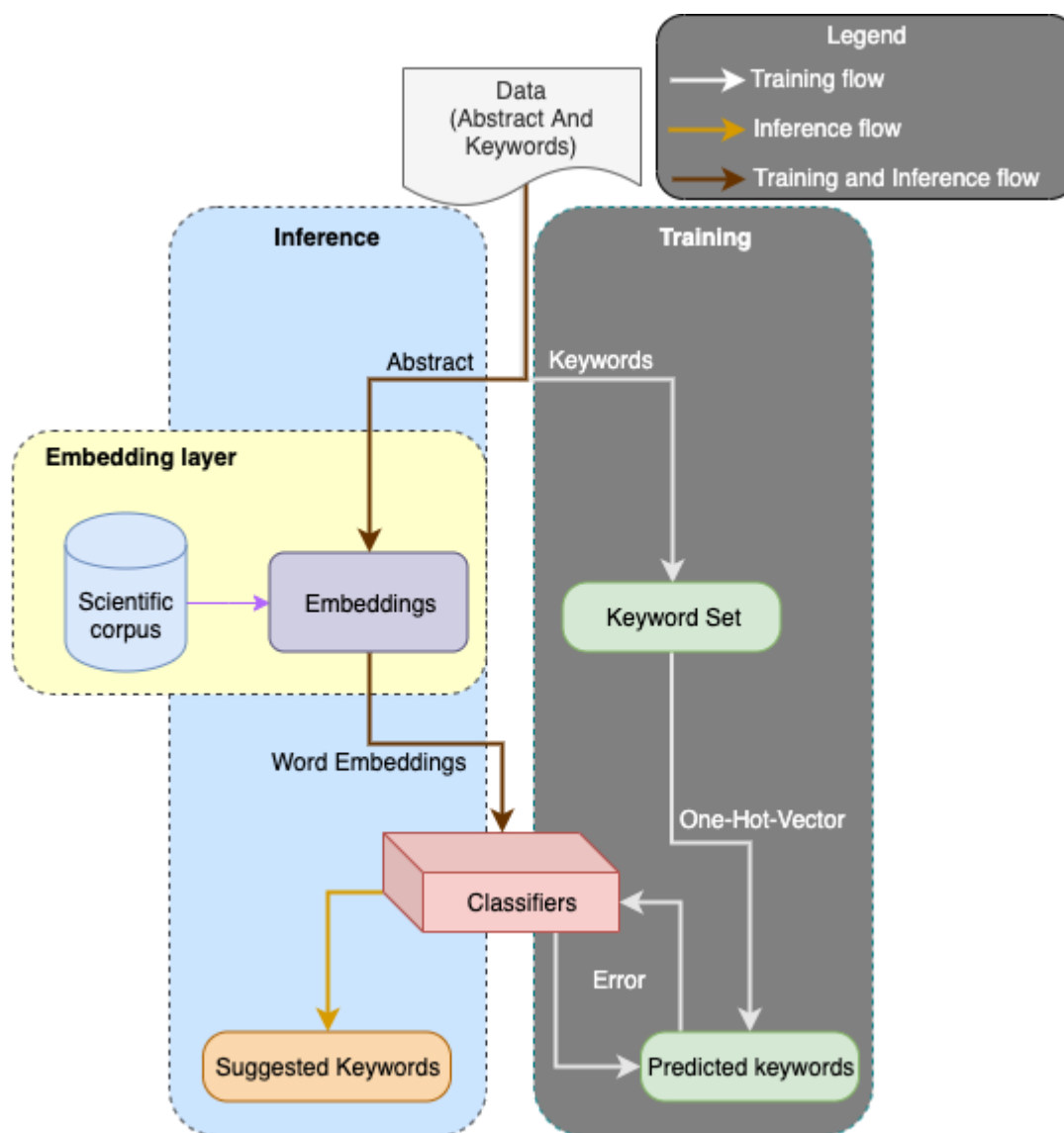


Fig. 6. Tool Workflow

- **Web Tool**

<https://gcmd.nasa-impact.net>

Description

The objective of the International Satellite Land Surface Climatology Project (ISLSCP II) study that produced this data set, ISLSCP II University of Maryland Global Land Cover Classifications 1992-1993, was to create a land cover map derived from 1 kilometer Advanced Very High Resolution Radiometer (AVHRR) data using all available bands. During this re-processing, the original University of Maryland (UMD) land cover type and fraction maps were adjusted to match the water/land fraction of the ISLSCP II land/water mask. These maps were generated for use by modelers of global biogeochemical cycles and others in need of an internally consistent, global depiction of land cover. This product describes the geographic distributions of 13 classes of vegetation cover (plus water and unclassified classes) based on a modified International Geosphere-Biosphere Programme (IGBP) legend.

Predicted Keyword	Score	Is Correct?	Add Keywords Add
land use/land cover classification	0.71453536	Yes No	True Keywords land use classes land use/land cover classification vegetation cover
agricultural lands	0.5260576	Yes No	
vegetation index	0.45733252	Yes No	
vegetation cover	0.41815156	Yes No	
land use classes	0.3206329	Yes No	

Fig. 7. Web Tool

ES2VEC PERFORMANCE COMPARISON

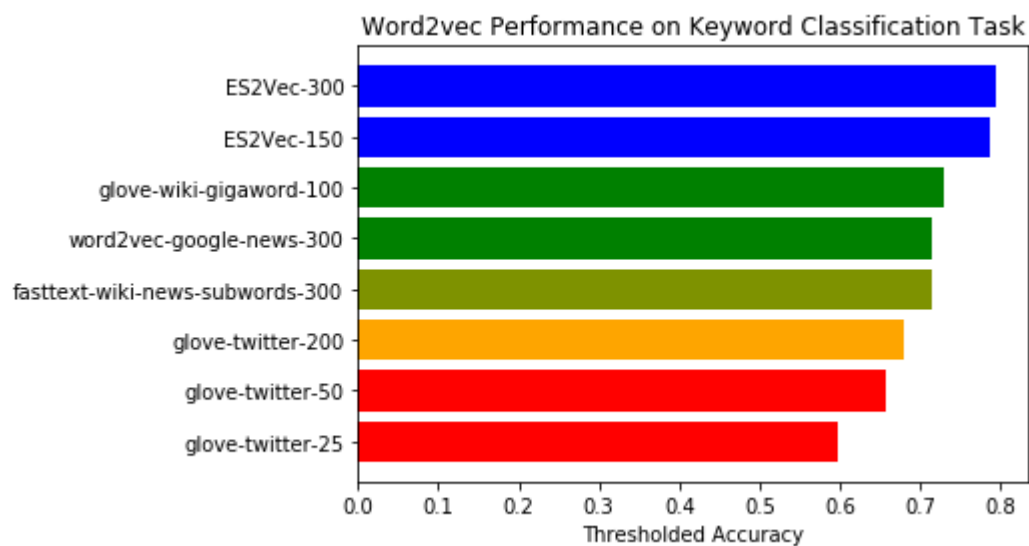


Fig. 8. ES2Vec performance vs generic Word2vecs

Some Notes on the Performance:

- Both variations of ES2Vec are fairly close: ~79% accuracy; next best model is ~73% accurate
- Smaller vocabulary size vs next best performing generic word2vec (500k vs 3M)
- Fewer tokens (115M vs 6B)
- Could be improved further by increasing the corpus size
- Shows the importance of developing domain-specific embeddings for domain-related downstream tasks

CONTACT AND ACKNOWLEDGEMENTS

Contact:

teamimpact@uah.edu

Acknowledgement:

This research is supported by NASA Grant NNM11AA01A as part of the Interagency Implementation and Advanced Concepts Team (IMPACT) program