Analogy-based Assessment of Domain-specific Word Embeddings

Derek Koehl
University of Alabama in Huntsville
Huntsville, USA
derek.koehl@uah.edu

Carson Davis
Manufacturing Technical Solutions
Huntsville, USA
carson.davis@nasa.gov

Udaysankar Nair
University of Alabama in Huntsville
Huntsville, USA
nair@nsstc.uah.edu

Rahul Ramachandran
National Aeronautics and Space Administration
Huntsville, USA
rahul.ramachandran@nasa.gov

***Abstract*—The ability of word embeddings to identify shared semantic regularities between word pair categories such as capital–country has led to the use of analogies as a method of validating word embedding models. Further research has shown that relative to the complete breadth of possible analogy categories, there exists a limit to the particular categories accessible, in terms of accuracy, to current analogy equations executed against word embeddings trained on generalized, non domain-specific text corpora. As most, if not all, domain-specific, scientific analogy pairs belong to problematic analogy categories (i.e. the lexicographical and the encyclopedic), we examine the degree to which a domain-specific text corpus and vocabulary positively improve analogy predictions from word embeddings. Our findings demonstrate that in comparison to analogy-based tests performed against general word embeddings, predictions by domain-specific word embeddings outperform in exactly those analogy categories that are both highly problematic and the location of domain knowledge.**

***Index Terms*—analogy test set, word embeddings, domain-specific, Earth science**

I. INTRODUCTION

Analogies are recognized as a method for validating word embeddings [16] [22]. Typically, both the word embeddings and the analogy test sets are built from generalized text corpora and generalized vocabularies. Recent research has examined the performance of word embeddings built from domain-specific text corpora and trained using domain-specific vocabularies [7]. Our research tests the hypothesis that word embeddings built from a domain-specific, Earth science corpus and trained using domain-specific vocabulary will better predict domain-specific, Earth science analogies when compared with the results achieved by tests of non domain-specific analogies against word embeddings produced by generalized corpora. Further, we tested the hypothesis that the improvement in predictions would occur in the categories of analogical relationships in which most, if not all, domain knowledge is to be found.

978-1-7281-6861-6/20/$31.00 ©2020 IEEE

The corpus from which we created a word embedding space consists of over 21,000 Earth science journal articles. We incorporate domain-specific vocabularies using two recognized Earth science ontologies: the Semantic Web for Earth and Environmental Technology ontology [20] and NASA's Global Change Master Directory (GCMD) [15]. To evaluate the ability of the word embeddings to predict Earth science domain analogies, we built an analogy prediction tool utilizing the linear analogy prediction equation set forth by Mikolov et al. [13].

Our results exceeded the outcomes produced by a comprehensive test set of analogy questions performed by Gladkova et al. against a general text corpus [8]. In addition, our results suggest remedies for a gap observed by Gladkova et al. with respect to the capabilities of current analogy prediction algorithms to accurately predict the full range of semantic relationship types, particularly in the lexicographical and the encyclopedic categories of analogical relationships. In summary, our investigation demonstrates that aligning the content of the corpus underlying the word embeddings to users' domain knowledge objectives and leveraging domain-specific vocabulary significantly improves the prediction of domain-specific analogies in precisely those categories where domain knowledge resides.

II. RELATED WORKS

A fundamental challenge of surfacing domain-specific knowledge from a text corpus based on keyword searches is that either the syntactic characteristics or the semantic denotation (or both) of any keyword is dependent to a greater or lesser extent on the larger context of the circumstance and manner in which a user utilizes that word as a material structure within a complex and abstract world [5]. Standardized glossaries of technical keywords (e.g. NASA's GCMD) mitigate this challenge, but do not entirely eliminate variance in syntactic use and denotation. Boden formalized the distinction between the localized use of words (termed P-creativity) and the globalized use of words (termed H-creativity) [4], and Wiggins argues that the computational

extraction of information from conceptual spaces must account for the syntactic and semantic differences that exist between the globalized and the localized (i.e. domain-specific) use of language [25].

Recent research has explored the application of domain-specific vocabulary to the formation of word vectors. Ghosh et al. demonstrated that training a Word2Vec model from a domain-specific corpus and utilizing a domain-specific vocabulary improved the prediction of disease attribute relationships [7]. In data-scarce domains, leveraging domain-specific vocabulary has been shown to improve the quality of generated word vectors [19]. Khatua et al. concluded that domain-specific input resulted in better predictions of meaningful semantic relationships from word vectors than did models trained on generalized texts [10].

The use of analogies as a validation method for word embedding models has been established as a recognized approach in natural language processing (NLP) research [9] [12] [21] [22]. Since its introduction by Mikolov et al. [13] the linear equation commonly represented by the expression:

$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen} \quad (1)$$

has become a standard for analogy prediction [8] [16] [24]. Over the previous decade, a range of analogy test sets have emerged built from generalized text corpora [9] [13] [14] [23]. The introduction of the Bigger Analogy Test Set (BATS), a semantically fine-grained analogy test set, by Gladkova et al. [8] revealed a wide gap in the ability of word embedding models derived from general text corpora to predict analogous relationships in the lexicographical and the encyclopedic semantic categories. We posit that it is precisely those semantic spaces in which domain knowledge resides.

III. METHODOLOGY

A. Compilation of the Training Corpus

We compiled a corpus of Earth science journal articles which consisted of the full text of 21,380 research papers downloaded from several scientific journals focused on Earth science (see Table I). Fourteen journals are represented in the corpus, with the majority sourced from just two of the journals: 36% coming from the *Geophysical Research Letters* and 29% from the *Journal of Geophysical Research: Atmospheres*.

Articles were identified using the open-source CrossRef API [2], and then downloaded and scraped using BeautifulSoup [18]. When possible, the abstract, full-text, keywords, and other important metadata were extracted. The resulting data was compiled into the final corpus, which, after processing, contained a total of 81 million words.

B. Generation of Domain Vocabularies

Many scientific concepts and instruments are uniquely represented by compound words that lose their meaning when deconstructed. In order to ensure a final embedding space that included these compound words, we complied a domain-specific vocabulary set by combining two key vocabularies: the Global Change Master Directory (GCMD) keywords[1], and the Semantic Web for Earth and Environmental Technology (SWEET) ontology[2]. Each was processed using the methods discussed in the next section, and compound phrases were identified and tokenized using underscores. In future work, these vocabularies can provide important filtering when performing tasks such as cosine similarity searches, in order to provide targeted searches for instruments, scientific keywords, or other sub-vocabularies provided by GCMD and SWEET.

TABLE I
CORPUS ARTICLES BY JOURNAL

Journal	Number	Percent of Total
Atmospheric Science Letters	273	1.28
Geophysical Research Letters	7,664	36.06
J. of Geophys. Res.: Atmospheres	6,161	28.99
J. of Geophys. Res.: Biogeosciences	539	2.54
J. of Geophys. Res.: Earth Surface	415	1.95
J. of Geophys. Res.: Oceans	1,392	6.55
J. of Geophys. Res.: Planets	483	2.27
J. of Geophys. Res.: Solid Earth	1,474	6.94
J. of Geophys. Res.: Space Physics	2,336	10.99
Meteorological Applications	255	1.20
Q. J. Royal Meteorological Society	8	0.04
Review of Geophysics	168	0.79
Water and Environment Journal	21	0.10
Wiley Interdisciplinary Reviews: Water	63	0.30

C. Text Processing

We utilized a standard general preprocessing step [3] in order to standardize and tokenize the raw text data. This included converting everything to lowercase, replacing contractions, removing punctuation, removing stopwords, handling scientific units, vectorizing based on word separation, and lemmatizing the remaining word tokens. A log of word lemmatization was kept so that processed versions of a word could be converted back to the most likely unprocessed version. Finally, compound words from the domain-specific vocabularies were processed in an identical manner, and then compared with the tokens from the corpus to identify compound phrases. These compounds were joined with underscores such that they would function as individual tokens with their own distinct word embeddings.

D. Generation of Word Embeddings

The processed corpus was used to generate word embeddings using the Gensim [17] implementation of Word2Vec in Python. A total embedding size of 150 was used, with a

[1]https://earthdata.nasa.gov/earth-observation-data/find-data/gcmd/gcmd-keywords

[2]https://github.com/ESIPFed/sweet

window of 10 tokens, and a min count of 3 tokens, using the entire text as an input, without breaking it up by sentences. Training was conducted for 300 epochs in order to generate the final word embeddings.

E. Analogy Prediction Tool

In order to allow easy access to the embedding space for researchers, an interactive website was created with Django in order to conduct cosine similarity searches on words in the corpus. In the Gensim implementation, the equation below was used to search the embedding space for embeddings b that minimize the cosine similarity to the given embedding a.

$$sim(a, b) = \cos(\theta) = \frac{a \cdot b}{\|a\|\|b\|} \tag{2}$$

This interface allowed for the entry of positive and negative contributions to the embedding a, as well as the selection of specific embedding model versions and the use of domain vocabularies as a filter for the results. By entering parts of an uncompleted analogy into the positive and negative fields, a researcher could produce a list of words in the embedding space with cosine similarities most similar to the analogy predicted by the basic analogy equation used in Mikolov et al. With this equation, given the analogy $a : b :: c : d$, the fourth term d can be predicted as follows:

$$argmax_{d \in V}(sim(d, c - a + b)) \tag{3}$$

F. Evaluation Process

An Earth science subject matter expert generated thirty-nine analogy word pairs which were combined into a list of twenty-eight Earth science-specific analogy statements.[3] Utilizing a classification framework similar to the one employed by Gladkova et al. [8], the twenty-eight analogy statements were categorized and then grouped as shown in Table II.

TABLE II
ANALOGY BY RELATIONSHIP TYPES

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

[3]The twenty-eight analogy statements are provided in Table IV in the Appendix.

The first three terms (a, b, c) of each analogy statement were entered into the analogy prediction tool which returned the top predictions for the fourth term, d. We established two accuracy tiers. The first tier (ES_1) defined accuracy as the fourth term of the analogy statement being the top prediction produced from the similarity search of the word embeddings. The second tier (ES_3) defined accuracy as the fourth term of the analogy statement being in the top three predictions produced from the similarity search.

IV. RESULTS

For the Earth science analogy test set, taken as a whole, the word embeddings produced an average accuracy of 0.29 at the ES_1 tier and 0.43 at the ES_3 tier. The average accuracies for the two primary analogy category divisions were:

- Lexicographical accuracy: 0.25 (ES_1) and 0.50 (ES_3)
- Encyclopedic accuracy: 0.32 (ES_1) and 0.38 (ES_3)

A. Contextualization

When developing their BATS, Gladkova et al. noted that many analogy test sets are unbalanced in that their analogy statements tend to privilege particular category types (e.g.morphology) and particular encyclopedic subcategories (e.g. capital–country) [8]. As is true for any domain knowledge area, Earth science analogy categories cluster exclusively in the lexicographical and the encyclopedic domains and exclude the inflectional and derivational morphological categories (e.g. singular–plural) entirely. Accordingly, we compared our results against the detailed results reported by Gladkova et al. when they used the BATS (which is balanced across four main analogy categories) to test general word embeddings. Table III presents our accuracies alongside the accuracies from the comparable category types reported by Gladkova et al. The table shows the analogy category accuracies for three word embedding models:

- Our domain-specific word embeddings built from a corpus of Earth science journal articles and domain-specific vocabularies (ES_1 and ES_3),
- General word embeddings built by Gladkova et al. using the GloVe model and a general text corpus (GloVe general), and
- General word embeddings built by Gladkova et al. using the Singular Value Decomposition model and a general text corpus (SVD general).

B. Discussion

Overall, the accuracy of domain-specific analogy predictions against a domain-specific word embedding space exceeded the best comparable accuracy of the predictions of general analogies performed against a general corpus word embeddings (i.e. GloVe general) by 8 and 22 percentage points for the ES_1 and ES_3 accuracy tiers respectively. When considering only analogy questions that fell within the lexicographical category domain, the improvement in accuracy was 14 and 39 percentage points for the ES_1 and ES_3 accuracy tiers respectively. Considering only the analogy questions

TABLE III
RESULTS COMPARISON

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

that fell within the encyclopedic domain, the improvement in accuracy was 6 percentage points for the ES_3 accuracy tier and comparable results for the ES_1 accuracy tier. If the outlier of the encyclopedic category capital–country[4] is excluded from the GloVe general and SVD general results, the improvement in accuracy in the encyclopedic domain was 8 and 14 percentage points and the improvement in overall accuracy was 12 and 26 percentage points for the ES_1 and ES_3 accuracy tiers respectively.

V. CONCLUSIONS

Our results demonstrate that the accuracy of domain-specific word embeddings in predicting domain-specific analogy questions outperforms the ability of general corpus word embeddings to predict general analogy questions in comparable analogy categories. This result was expected, as we anticipated that a domain-specific corpus would necessarily constrain the semantic relationships between words and, as a result, eliminate ambiguous relationships that emerge when words are used in a more generalized social context. We were encouraged, however, at the degree to which the accuracies increased compared to the predictions of general word embeddings, particularly in the lexicographical domain. This is due to the fact that in comparison to the inflectional morphology category which drives many current successes in analogy testing experiments, Gladkova et al. demonstrated that the lexicographical category is the most problematic in terms of prediction accuracies (e.g. 0.11) produced by general word embeddings. [8]

Research into the natural language processing utility of analogical reasoning has shown that when applied to word embeddings produced from general text corpora, an analogical approach can be utilized for word sense disambiguation [6] and broad-range detection of semantic features [11]. The improved results demonstrated by this baseline querying of domain-specific analogy statements against domain-specific word embeddings suggest that more semantically advanced applications of analogical logic in the context of such word embeddings could produce equally promising results.

As this research experiment was largely exploratory in nature, the analogy test set it employed was limited in size and category scope. The better-than-expected results validate the utility and importance of taking the next step and developing a comprehensive Earth science analogy test set that covers a full breadth of lexicographical and encyclopedic subcategories and contains a deeper set of domain-specific word pairs within each subcategory. We predict that such an undertaking would further demonstrate the efficacy of domain-specific analogy test sets for natural language processing experiments undertaken within various academic and professional disciplines.

In addition, the results support continued research into the development of domain-specific text corpora and the application of domain-specific vocabularies, such as the American Meteorological Society Glossary [1], to the training of word embeddings. In particular, analysis of the test results demonstrated the potential of domain-specific vocabulary filters to enhance prediction accuracy. For example, when the analogy statement *carbon dioxide* : *acidic* :: *ammonium* : *?* was tested against the baseline domain-specific word embeddings, the fourth term *alkaline* did not appear in the top twenty predictions. However if the user choose the filtering vocabulary 'All SWEET Words' which is more closely aligned with the category of the analogy statement (i.e. encyclopedic: compound-property), *alkaline* appeared in the top four predictions. If the user choose the filtering vocabulary 'Property', *alkaline* appeared in the top two predictions. In domain fields of knowledge, the selection of a filter vocabulary could be attempted programmatically by categorizing the second term of the analogy statement into a category—either manually or by its presence in or its cosine similarity to a category—and using that as the filter for the fourth term. The ability of domain-specific vocabulary to provide prediction contextualization is an avenue of research suggested by our findings.

A final set of conclusions that presented themselves as we reviewed the output of the experiment address limitations observed in current analogy prediction methods, particularly when comparing prediction accuracies that fall within the lexicographical categories with prediction accuracies that fall within the inflectional and derivational morphological categories. [8]. Whereas in the results obtained by Gladkova et al., the lexicographical category posted the lowest accuracies (i.e. 0.11) of the four analogy categories, our preliminary results (i.e. 0.25 and 0.50) suggest that domain-specific word embeddings have the promise of addressing this area of weakness in analogical domain knowledge prediction.

In addition, as we analyzed the test results we discovered that the word order of analogy statements affected the results produced by the linear equation currently the standard for

[4]The capital–country category scored accuracies of 0.97 (GloVe general) and 0.77 (SVD general), and based on accuracy results such as these, is often over-represented in general analogy test sets (e.g. it constitutes 57 percent of the semantic questions within the Google analogy test set) [14].

analogy prediction. For example, *stratus* : *cooling* :: *cirrus* : *?* correctly predicted the fourth term *warming*; whereas *cooling* : *stratus* :: *warming* : *?* did not correctly predict *cirrus* as the fourth term. Likewise for the analogy question *sulfate* : *scattering* :: *carbon* : *?*, the correct term *absorption* did not appear in the top twenty predictions; however when the analogy terms were reordered as *carbon* : *scattering* :: *sulfate* : *?*, *absorption* was the top prediction. These results confirm that there is still research to be done to improve analogy prediction heuristics themselves, especially as it touches upon their ability to predict analogous relationships within the lexicographical and encyclopedic categories.

REFERENCES

[1] American Meteorological Society, "AMS Glossary," 2019. [Online]. Available: http://glossary.ametsoc.org/wiki/Special:AllPages.

[2] Bilder, Geoffrey, "CrossRef API," 2019. [Online]. Available:https://github.com/CrossRef/rest-api-doc.

[3] S. Bird, E. Loper, and E. Klein, *Natural Language Processing with Python, Sebastopol*, CA: O'Reilly Media, 2009.

[4] M. A. Boden, *The creative mind: myths and mechanisms*, 2nd ed. London: Routledge, 2004.

[5] A. Clark, "Language, embodiment, and the cognitive niche," Trends in Cognitive Sciences, vol. 10, no. 8, pp. 370–374, Aug. 2006.

[6] S. Federici, S. Montemagni, and V Pirrelli, "Inferring semantic similarity from distributional evidence: an analogy-based approach to word sense disambiguation," in *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997.

[7] S. Ghosh, P. Chakraborty, E. Cohn, J. S. Brownstein, and N. Ramakrishnan, "Characterizing Diseases from Unstructured Text: A Vocabulary Driven Word2vec Approach," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16*, Indianapolis, Indiana, USA, 2016, pp. 1129–1138.

[8] A. Gladkova, A. Drozd, and S. Matsuoka, "Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't," in *Proceedings of the NAACL Student Research Workshop*, San Diego, California, 2016, pp. 8–15.

[9] D. A. Jurgens, S. M. Mohammad, P. D. Turney, and K. J. Holyoak, "SemEval-2012 Task 2: Measuring Degrees of Relational Similarity," in *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, Montréal, Canada, 2012, pp. 356–364.

[10] A. Khatua, A. Khatua, and E. Cambria, "A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks," Information Processing & Management, vol. 56, no. 1, pp. 247–257, Jan. 2019.

[11] Y. Lepage, C. L. Goh, "Towards automatic acquisition of linguistic features." in *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, 2009, pp. 118-125.

[12] S. McGregor, M. Purver, and G. Wiggins, "Words, Concepts, and the Geometry of Analogy," in *Proceedings of the 2016 Workshop on Semantic Spaces at the Intersection of NLP, Physics and Cognitive Science*, Glasgow, Scotland, 2016, pp. 39-48.

[13] T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 746-751.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2013.

[15] National Aeronautics and Space Administration, "Global Change Master Directory," 2019. [Online]. Available: https://gcmd.nasa.gov/index.html.

[16] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543.

[17] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010. pp. 45-50.

[18] Richardson, Leonard, "BeautifulSoup," 2019. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/.

[19] J. C. Ross, R. Murthy, K. K. Ganguli, and P. Bhattacharyya, "Identifying Raga Similarity in Hindustani Classical Music through Distributed Representation of Raga Names," in *Proceedings of the 13th International Symposium on CMMR*, Matosinhos, Portugal, 2017.

[20] R. G. Raskin, "SWEET 2.1 Ontologies," in *AGU Fall Meeting Abstracts*, 2010.

[21] P. D. Turney and M. L. Littman, "Corpus-based Learning of Analogies and Semantic Relations," *Machine Learning*, vol. 60, pp. 251–278, Sep. 2005.

[22] P.D. Turney, "Similarity of Semantic Relations," *Computational Linguistics*, vol. 32, no. 3, pp. 379-416, Sep. 2006.

[23] P.D. Turney, "The latent relation mapping engine: algorithm and experiments," *Journal of Artificial Intelligence Research*, vol. 33, pp. 615-655, 2008.

[24] E. Vylomova, L. Rimell, T. Cohn, and T. Baldwin, "Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning," arXiv:1509.01692v4, Aug. 2016.

[25] G. A. Wiggins, "A preliminary framework for description, analysis and comparison of creative systems," *Knowledge-Based Systems*, vol. 19, no. 7, pp. 449–458, Nov. 2006.

TABLE IV
APPENDIX: EARTH SCIENCE ANALOGY TEST SET

Analogy Test Questions		
carbon dioxide : acidic	::	ammonium : alkaline
divergence : anticyclonic	::	convergence: cyclonic
warm : sensible heat flux	::	wet : latent heat flux
longwave : planetary vorticity	::	shortwave : relative vorticity
water vapor : longwave	::	ozone : shortwave
ozone : shortwave	::	water vapor : longwave
gravity wave : buoyancy	::	rossby wave : potential vorticity
temperature : thermometer	::	pressure : pressure sensor
temperature : thermometer	::	humidity : humidity sensor
blue : ultraviolet	::	red : infrared
cyclostrophic : centrifugal	::	geostrophic : coriolis
temperature : potential temperature	::	vorticity : potential vorticity
extra tropical cyclone : baroclinic	::	tropical cyclone : barotropic
cyclone : weather	::	enso : climate
cyclone : weather	::	north atlantic oscillation : climate
front : temperature	::	dryline : humidity
pressure : millibar	::	temperature : degree
graupel : ice	::	rain : water
rain : water	::	graupel : ice
stratus : low-level cloud	::	cirrus : high-level cloud
low-level cloud : stratus	::	high-level cloud : cirrus
cirrus : warming	::	stratus : cooling
stratus : cooling	::	cirrus : warming
cooling : stratus	::	warming : cirrus
sulfate : scattering	::	carbon : absorption
sulfate : scattering	::	dust : absorption
scattering : sulfate	::	absorption : carbon
absorption : carbon	::	scattering : sulfate