Analogy-based Assessment of Domain-specific Word Embeddings

Derek Koehl, Carson Davis, Udaysankur Nair, Rahul Ramachandran

IEEE SoutheastCon 2020
March 14, 2020

IMPACT
Interagency Implementation and Advanced Concepts Team

Analogical Reasoning

Analogical reasoning is fundamental to human cognition and plays an important role in a wide-range of problem solving scenarios.

"Analogy is our best guide in all philosophical investigations; and all discoveries, which were not made by mere accident, have been made by the help of it." – Joseph Priestley

Prediction of new thermoelectric materials

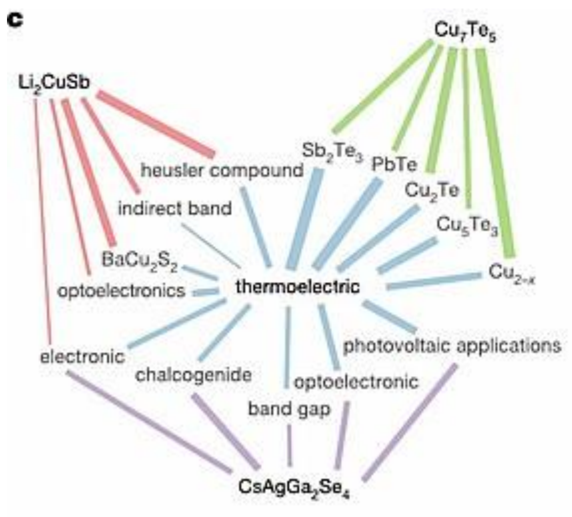


Figure: Tshitoyan et al., 2019.

Characterizing diseases

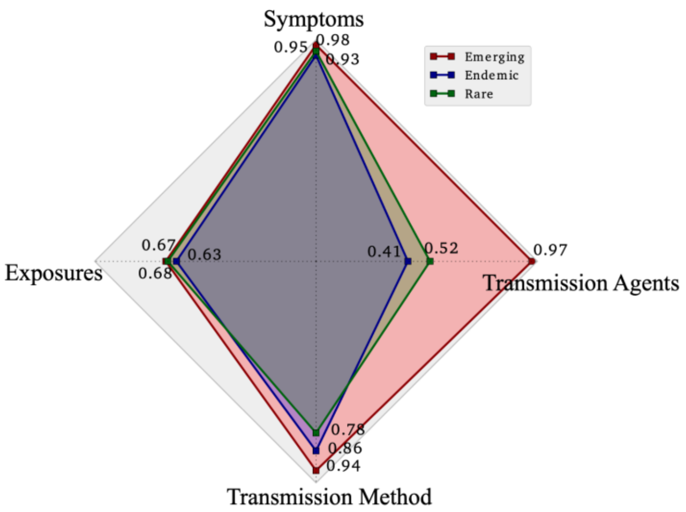


Figure: Ghosh et al., 2016.

Analogy Prediction: Successes and Reality

$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$$

$$\overrightarrow{London} - \overrightarrow{England} + \overrightarrow{France} \approx \overrightarrow{Paris}$$

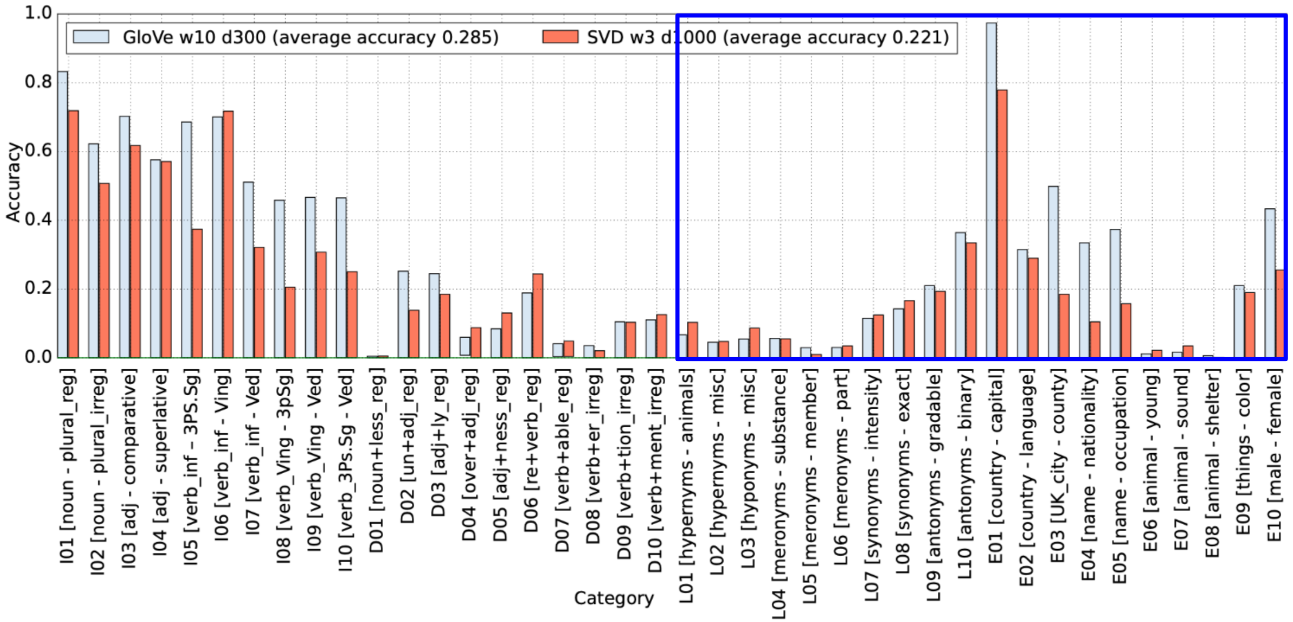
Reported prediction accuracies exceeding 80%

(e.g. Pennington et al., 2016)

Bigger Analogy Test Set

(Gladkova et al., 2016)

- Balanced across four semantic categories
- Average lexicographic accuracy ≈ 0.11
- Average encyclopedic accuracy ≈ 0.26



Underlying chart: Gladkova et al., 2016

Domain knowledge resides in the lexicographic and encyclopedic domains.

Domain Analogy Prediction: An Exploration

Can word embeddings trained on a domain-specific corpus utilizing domain-specific vocabulary outperform in domain-specific analogy prediction as compared to word embeddings trained on a non domain-specific corpus predicting non domain-specific analogies?

carbon dioxide:*acidic* :: *ammonium*:*alkaline*

vs.

horse:*stable* :: *chicken*:*coop*

Domain Analogy Prediction: An Exploration

Domain corpus: 21,380 Earth science articles comprising 81 million words

Domain vocabulary:

- Global Change Master Directory (GCMD) keywords
- Semantic Web for Earth and Environmental Technology (SWEET) ontology

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

Domain Analogy Prediction: Analogy Test Set

The exploratory set of Earth science analogies categorized using a classification framework similar to the one employed by the Bigger Analogy Test Set.

- Allows category-level comparison
- Laying the groundwork for an Earth science analogy test set

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

Domain Analogy Prediction: Accuracies

Accuracy tiers:

- **ES_1** – The fourth term of the analogy is the top prediction
- **ES_3** – The fourth term appears in the top three predictions

Comparison accuracies:

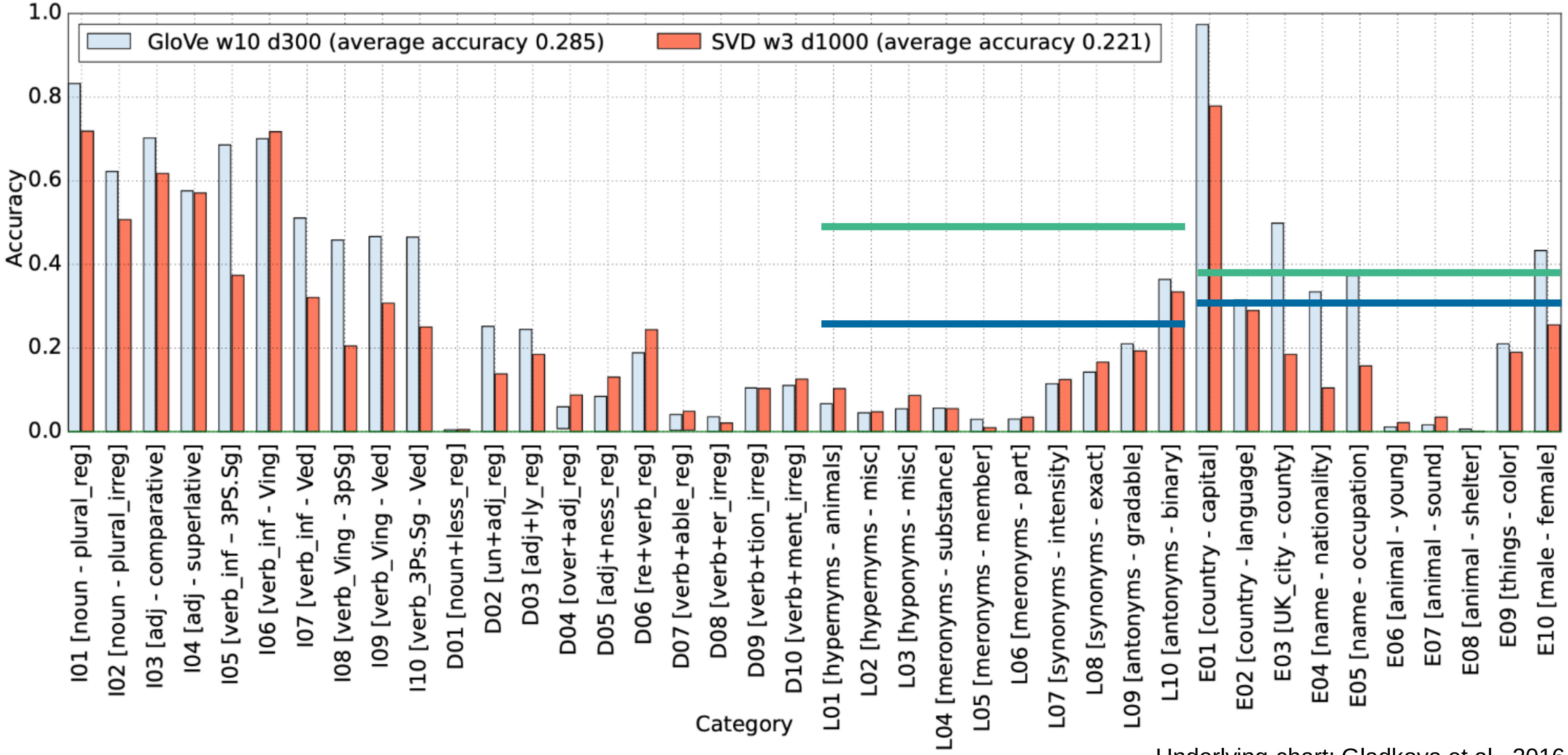
- **GloVe general** – reported by Gladkova et al. using the GloVe model
- **SVD general** – reported by Gladkova et al. using the Singular Value Decomposition model

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

Domain Analogy Prediction: Comparisons

ES_3: lexicographical – 0.50; encyclopedic – 0.38

ES_1: lexicographical – 0.25; encyclopedic – 0.32



Underlying chart: Gladkova et al., 2016

Domain Analogy Prediction: Next Steps

Expansion of the domain corpus

- Increase the total number of documents to a target of 100,000
- Expand the scope of American Geophysical Union publications
- Ingest the documents into a database for domain subsetting

Develop a comprehensive Earth science analogy test set

- Lexicographical: minimum 10 subcategories
- Encyclopedic: minimum 20 subcategories
- Database-driven to allow for domain subsetting

Domain Analogy Prediction: Contact and Questions

Derek Koehl

derek.koehl@uah.edu

Questions?

Citations:

S. Ghosh et al., “Characterizing diseases from unstructured text: A vocabulary driven word2vec approach,” In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016.

A. Gladkova et al., “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*, 2016.

V. Tshitoyan et al., “Unsupervised word embeddings capture latent knowledge from materials science literature,” *Nature*, 2019.