Natural Resources Canada | Ressources naturelles Canada

GEOLOGICAL SURVEY OF CANADA
OPEN FILE 8848

Datasets to support geoscience language models

S. Raimondo, T. Chen, A. Zakharov, L. Brin, D. Kur, J. Hui,
S.L. Burgoyne, G. Newton, and C.J.M. Lawley

2022

Canada

GEOLOGICAL SURVEY OF CANADA
OPEN FILE 8848

Datasets to support geoscience language models

S. Raimondo[1], T. Chen[1], A. Zakharov[1], L. Brin[1], D. Kur[1], J. Hui[1], S.L. Burgoyne[2], G. Newton[2], and C.J.M. Lawley[2]

[1]ServiceNow, 296 Richmond Street West, Toronto, Ontario
[2]Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario

2022

Permanent link: https://doi.org/10.4095/329265

This publication is available for free download through GEOSCAN (https://geoscan.nrcan.gc.ca/).

Recommended citation
Raimondo, S., Chen, T., Zakharov, A., Brin, L., Kur, D., Hui, J., Burgoyne, S.L., Newton, G., and Lawley, C.J.M., 2022. Datasets to support geoscience language models; Geological Survey of Canada, Open File 8848, 1 .zip file. https://doi.org/10.4095/329265

Publications in this series have not been edited; they are released as submitted by the author.

Datasets to support geoscience language models

Introduction

Language models are the foundation for the predictive text tools that billions of people use in their everyday lives. Although these language models are often trained on vast digital corpora, they are often missing the specialized vocabulary and underlying concepts that are important to specific scientific sub-domains. Herein we report two new language models that were updated using geoscientific text to address that knowledge gap. The raw and processed text from the GEOSCAN publications database, which were used to generate these new language models are also reported. Language model performance and validation are discussed separately in Lawley et al. (in press). The supporting datasets and geoscientific language models can be used and expanded on in the future to support a range of down-stream natural language processing tasks (e.g., keyword prediction, document similarity, and recommender systems).

Geoscientific text

Language models are based, in part, on a variety of geoscientific publications sourced from the Natural Resources Canada (NRCan) GEOSCAN publications database (n = 27,081 documents). Figures, maps, tables, references, irregularly formatted text, and other large sections of documents from poor-quality scans were excluded from further analysis (i.e., the total GEOSCAN database contains approximately 83k documents; however <32% were readily available for use as part of the current study). The "pdfminer" library (https://github.com/pdfminer/pdfminer.six) was used to extract text from the remaining pdf documents prior to a number of pre-processing steps, including removing punctuation, replacing upper casing, removing French text, removing specific forms of alpha-numeric data (e.g., DOIs, URLs, emails, and phone numbers), converting all non-ascii characters to their ascii equivalent, filtering text boxes that contain an insufficient percentage of detectable words, and merging all of the extracted text for each document. Raw and pre-processed text data from the GEOSCAN publications database are freely available herein as "GEOSCAN_Text_Raw.zip" and "GEOSCAN_Text_Processed.zip", respectively. Additional geoscientific publications that were used to re-train language models were sourced from provincial government publication databases (e.g., Ontario Geological Survey, Alberta Geological Survey, and British Columbia Geological Survey; n = 13,898 documents) and a subset of open access journals (e.g., Materials, Solid Earth, Geosciences, Geochemical Perspective Letters, and Quaternary) available through the Directory of Open Access Journals (DOAJ; n = 3,998 documents). The code to reproduce data processing and train language modelling is available

for free (https://github.com/NRCan/geoscience_language_models). Testing and validation of the language models is described in Lawley et al. (in press).

GloVe Model

The Global Vectors for Word Representation (GloVe) method (Pennington et al., 2014) was used to map each word in the training corpus to a set of numerical vectors in N-dimensional space and was originally trained using billions of words, or sub-words, from the Wikipedia (2014) and the 5th Edition of English Gigaword (Parker et al., 2011). This original GloVe model was then re-trained as part of the current study using the smaller, but domain-specific corpora to improve model performance (i.e., the preferred GloVe model). This preferred GloVe model was trained using the AdaGrad algorithm with the most abundant tokens (i.e., minimum frequency of 5), considering a context window of size 15 for 15 iterations, fixed weighing functions (x_max = 10 and alpha = 0.75) and is based on the 300-dimensional vectors as described by Pennington et al. (2014).

BERT Model

Contextual language models, including the Bidirectional Encoder Representations from Transformers (BERT) method (Devlin et al., 2019), consider words and their neighbours for a more complete representation of their meaning. The original BERT model was pre-trained on the Books Corpus (Zhu et al., 2015) and English Wikipedia, comprising billions of words. More recently, the DistilBERT method (Sanh et al., 2019) was proposed to simplify the training process for smaller datasets, produce language models that are less susceptible to overfitting, and yield model performance that are comparable to the original BERT method. The first step for all BERT models is to convert pre-processed text to tokens, which may include words, sub-words or punctuation. Sub-word tokenization limits the number of out-of-vocabulary words, which allows BERT models trained on general corpora to be applied to specific sub-domains. A geology specific tokenizer was created as part of the current study by adding geology tokens prior to continued pre-training using the geoscientific corpora. This preferred BERT (i.e., using the geo-tokenizer and geoscientific corpora) model was generated using the "HuggingFace" machine learning library (https://github.com/huggingface) with the same combination of hyper-parameters described in the original Devlin et al. (2019) method (e.g.,

learning rate = 5e-5 and 2.5e-5; batch size = 48; max steps = 1 and 3 million; warm-up steps: 0, 100k, 300k).

Acknowledgments

This work was completed as part of the Targeted Geoscience Initiative program. Thanks to Boyan Brodaric for providing constructive comments that improved the research.

References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., 2019, Bert: Pre-training of deep bidirectional transformers for language understanding: arXiv, v. arXiv:1810, p. 16.

Lawley, C.J.M., Raimondo, S., Chen, T., Brin, L., Zakharov, A., Kur, D., Hui, J., Newton, G., Burgoyne, S.L., and Marquis, G., in press, Geoscience language models and their intrinsic evaluation: Applied Computing and Geosciences.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K., 2011, English Gigaword Fifth Edition: English Gigaword Fifth Edition LDC2011T07.

Pennington, J., Socher, R., and Manning, C., 2014, GloVe: Global vectors for word representation: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), p. 1532–1543.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T., 2019, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter: arXiv, v. arXiv:1910, p. 5.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S., 2015, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books: arXiv, v. arXiv:1506, p. 23.