Representation Learning in Geology and GilBERT

Zikri Bayraktar, Hedi Driss, Marie Lefranc
Schlumberger-Doll Research Center
Cambridge, MA 02139
zbayraktar@slb.com, hdriss@slb.com, mlefranc@slb.com

Abstract

Geology lays at the foundation of the oil and gas industry and a good understanding of geology in each newly drilled well can make or break an exploration project with a price tag in the millions of dollars. Over the past decades, each drilled well has been extensively analyzed, where geology and other petrophysical properties were interpreted by experts and rigorously documented. As this creates a valuable source of information for future drilling success, most of it is stored in PDF files in knowledge silos of companies. Recent advancements in cloud technologies and machine learning techniques are enabling the future to be open-source and access to these technical documents is providing a broad geological knowledge of the different basins in the world. In this work, we focus on geology reports of wells drilled in the Norwegian Sea with the goal to learn numerical representations for geological descriptions in these fields and utilize these representations to find worldwide geological analogues. The automation of analog identification can improve expert interpretation, exploration success, and save a significant amount of effort and time for oil and gas companies. We will present numerical encoding approaches we took in the pursuit of capturing representations of geological knowledge from files as well as challenges faced during this work and road map towards GilBERT; *Geologically informed language modeling with BERT*, for the use in geology-based NLP applications in oil-and-gas (O&G) industry.

1 Approaches

True knowledge extraction from unstructured text is an extremely challenging task where the definition of knowledge can be open to debate and depends on the domain. In this work, we are targeting geology specifically within the context of O&G exploration. As our goal is to generate numerical representations for geological knowledge, the first approach we took was the shallow-neural network-based word embeddings [1-3, 11]. If one can represent geologically related words in numerical vectors, one can then utilize them to find similarities or exploit them to make recommendations similar to [4,5]. Our initial attempts to utilize pre-trained open-source version of the word2vec and GloVe demonstrated the need for a domain specific training. In the top portion of Figure 1, we can see that the most similar words returned for *'channel'* are completely irrelevant in a geological context, even though these open-source models are trained on large corpus in the range of billions of words. Google's word2vec model, trained on Google News, returns results like *'Cartoon Network'* whereas GloVe model, trained on Wikipedia, returns words related to TV broadcasting as most similar words to *'channel'*. The meaning of *'channel'* in geological context refers to a type of meandering, braided, anastomosing, or straight natural landform filled with fluid (e.g. river channel, submarine fan channel). To overcome this challenge, we leveraged the content of textbooks focusing on sedimentology, subdomain of geology, as well as open-access lecture notes and created a corpus of 4 million words with vocabulary around 30 thousand words. We also leveraged dissertations on geology from universities' open portals, but our geology domain experts quickly realized that such works are almost always specialized projects focusing on a subdomain of geology or a unique region

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

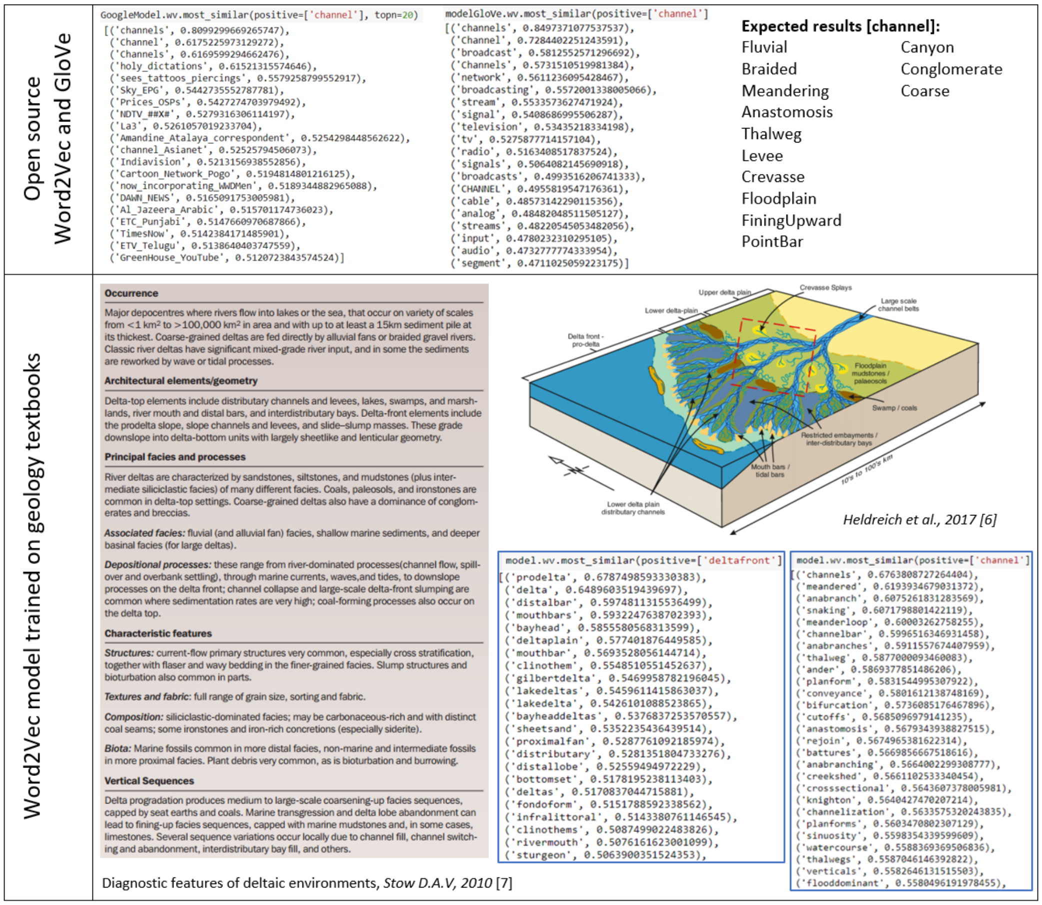

Figure 1: This figure compares the open-source word2vec and GloVe models against the geology based word2vec model that we trained using only geology corpus.

causing models to be biased, whereas knowledge from textbooks is more universal. All our corpus was in PDF format and parsing them presented many problems like multi-column text, textboxes over images, summary text in tables, irrelevant information like references and front/back matters as well as the non-standard nature of PDFs themselves. Once the data ingestion pipeline was built, we trained word2vec models. In the lower part of Figure 1, we can see that the most similar words to *'channel'* are now geologically relevant. Similarly, the word *'deltafront'* has high similarity scores to the physically closer landforms like *'deltaplain'*, or *'mouthbar'*, and other semantically meaningful words like *'gilbertdelta'* and *'lakedelta'* are different types of deltas. In addition, model learned the analogies like *'pointbar'* + *'inner'* - *'cutbank'* = *'outer'*, where *'pointbar'* represents a low crescentic shoal on the convex side (*'inner'* bend) of a river bend and *'cutbank'* represents the *'outer'* bend.

While our trained word2vec model is significantly better in geology compared to the open-source versions, as inspected by geology domain experts, it lacks the understanding of the context even within the domain. Understanding the context is a step towards overcoming the word sense disambiguation, where the promise of the deep-learning based models like ELMo [8] and ULMFiT [9] lays. These methods utilize LSTM based architectures where internal states can encode words in sentences leading the way towards context awareness. In our second approach, we utilized U.S.E. (Universal Sentence Encoder) [10], a transformer-based model capable of handling multiple sentences or paragraphs, to demonstrate that we can encode multi-sentence expert descriptions of geological formations into numerical vectors and then query it to find the desired formation. A formation is defined as the fundamental unit with specific features that will distinguish one rock formation from another and may show similarities to formations in other parts of the world. In the example in Figure 2, we extracted formation descriptions written by the Norwegian Petroleum Directorate (NPD) and encoded each description into numerical representations with U.S.E. We then queried with a sentence describing the

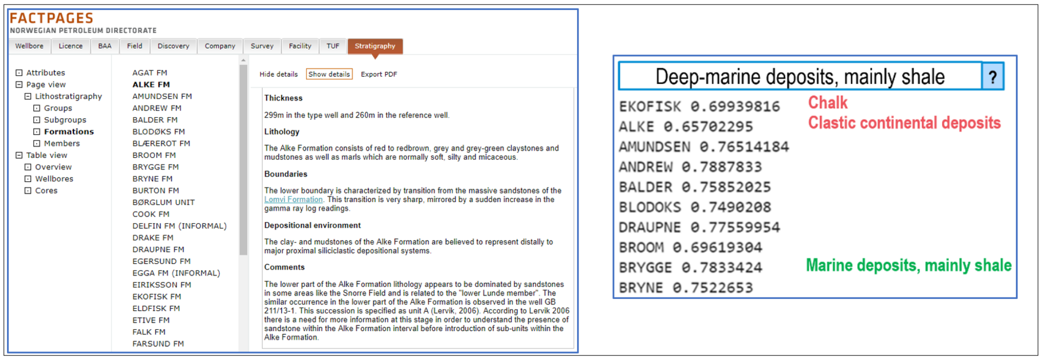

Figure 2: Norwegian Petroleum Directorate (NPD) factpages [12] stratigraphy formations under Lithostratigraphy tab, description of *'Alke'* formation is shown (left). Similarity score of selected 10 formations against the user-defined query sentence vector encoded with U.S.E. is shown (right).

desired geological formation such as: *'deep-marine deposits, mainly shale'*. Similarity scores to the query vector is high for formations with similar geological features such as in the *'Brygge'* formation but much lower for formations with different geological properties like *'Ekofisk'* or *'Alke'* formations. This is very useful for experts to quickly identify analogous formations - whose properties such as porosity, permeability, depositional environments, production and drilling history - will be used to improve the interpretation in specific areas with less data.

Current state-of-the-art in language modeling is BERT [13], which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. It is designed with a pre-training step of bidirectional encoders on unlabeled data to create context-based representations without supervision. It has been shown in [14] that BERT without fine-tuning can contain relational knowledge competitive with traditional NLP methods. It can also be used as embedding step in other NLP tasks with a slight modification which makes it an ideal candidate for our purposes as well. The benefits of pre-training of BERT on large-scale domain-specific corpora as in biomedical domain has been demonstrated with BioBERT [15]. Similarly, we propose GilBERT; *Geologically informed language modeling with BERT*, to capture and encode context-aware geological understanding from technical expert reports as well as to be the founding step for geology-based domain-specific language model in NLP applications in O&G industry. Our initial corpus constituted the content we leveraged from textbooks with expert reports written for wells documented in NPD database. Initial attempts to fine-tune open-sourced BERT using our geology corpus was less than desirable. To increase our success, we supplement our corpus with more open-sourced geological text and train GilBERT only on our corpus. As training is ongoing, results will be presented in details at the workshop.

2 Challenges

During this work, we encountered many problems at the data digestion stage mainly categorized in two data groups. The first one represents the data from PDF textbooks and the second one represents the data from PDFs in database. PDF textbooks were produced over many decades by different publishers and unfortunately not standardized. Processing them required multiple different open-source PDF readers to extract most of the text. Multi-column format, different embedding styles, many illustrative images with short text overlaid, front and back matters of each book, bibliography sections in each chapter, text in tables, equations, unique characters were few of the challenges diluting the quality of the extracted text. Well-based reports from NPD were also in PDF format which had to be OCRed. These reports were spanning over 50 years of extensive documentation of each drilled well and recorded. Older documents contained handwritten notes, tables, measurements and sketches. These were the least accurate OCRed documents. More recent reports were more structured but often without any common templates between different companies. Moreover, diverse expert styles in documentation, different company documentation standards, as well as sections written in different languages, i.e. English and Norwegian descriptions in NPD, increased the difficulty by many folds.

References

[1] Mikolov, T., Chen, K., Corrado, G, Dean, J., Efficient Estimation of Word Representations in Vector Space", ArXiv preprint arXiv: 1301.3781 [cs.CL], 2013.

[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., Distributed Representations of Words and Phrases and their Compositionality, ArXiv preprint arXiv:1310.4546 [cs.CL], 2013.

[3] Pennington, J., Socher, R., Manning, C. D., GloVe: Global Vectors for Word Representation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), p. 1532-1543, 2014.

[4] Grbovic, M., Cheng, H., Real-time Personalization using Embeddings for Search Ranking at Airbnb, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, p. 311-320, 2018.

[5] Wang, J., Huang, P., Zhao, H., Zhang, Z., Zhao, B., Lee, D. L., Billion-scale Commodity Embedding for E-commerce Recommendation in Alibaba, ArXiv preprint arXiv:1803.02349 [cs.IR], 2018.

[6] Heldreich, G., Redfern, J., Legler, B., Gerdes, K., Williams, P. J. B., Challenges in characterizing subsurface paralic reservoir geometries: a detailed case study of the Mungaroo Formation, North West Shelf, Australia. Geological Society, London, Special Publications. 444. SP444.13. 10.1144/SP444.13.

[7] Stow, D. A.V., Sedimentary rocks in the field - A Colour Guide. Fifth impression. Manson Publishing Ltd. 320 pp, 2010.

[8] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., Deep contextualized word representations", ArXiv preprint arXiv:1802.05365 [cs.CL], 2018.

[9] Howard, J., Ruder, S., Universal Language Model Fine-tuning for Text Classification, ArXiv preprint arXiv:1801.06146 [cs.CL], 2018.

[10] Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., Kurzweil, R., Universal Sentence Encoder", ArXiv preprint arXiv:1803.11175 [cs.CL], 2018.

[11] Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., Jain, A., Unsupervised word embeddings capture latent knowledge from materials science literature", Nature, v. 571, p. 95 - 98, 2019.

[12] Norwegian Petroleum Directorate (NPD) factpages, http://factpages.npd.no/factpages/

[13] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ArXiv preprint arXiv:1810.04805 [cs.CL], 2018.

[14] Petroni, F., Rocktaschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., Riedel, S., Language Models as Knowledge Bases?, ArXiv preprint arXiv:1909.01066 [cs.CL], 2019.

[15] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., Kang, J., BioBERT: a pre-trained biomedical language representation model for biomedical text mining", ArXiv preprint arXiv:1901.08746 [cs.CL], 2019.