Javier Huamani Huamani2@illinois.edu

Applications of Natural Language Processing in the Oil and Gas Industry

The oil and gas industry is one of the oldest industries of the world economy. In 1858 the first oil well became operational and since then the demand for energy and oil production have simultaneously increased, with some variability, for the last two centuries. Developing technologies have been improving the means to locate and extract oil and gas, and therefore meet the energy demand. However, most of these improvements focused on refining engineering, chemical and geological processes. In contrast, research in Artificial Intelligence (AI) and its subset Machine Learning (ML) have grown immensely over the last 50 years. In the past two decades alone, some of the more prominent focus of AI is Deep Learning and Big Data. ML Algorithms have been developed, or are being developed, for the retrieval and mining of structured and unstructured data. However, algorithms for unstructured text based data are still maturing in the form of Natural Language Processing (NLP). The oil and gas industry is entering a digital age in which multiple sources of structured and unstructured data need to be mined for optimization of costs, operations and risk reduction. Specifically, two case studies will be discussed on applications of NLP in the oil and gas industry: Sequence Mining and Pattern Analysis in Drilling Reports with Deep Natural Language Processing and Representation Learning in Geology and GilBERT [1][2].

Textual data is captured at every stage of the hydrocarbon production lifecycle. Some examples sources of text data are incident reports, maintenance logs, daily service reports, patents, etc. The first case study focused on text data from drilling reports in which field operators interpret sensor measurements. The interpretations include descriptions of the state of a drilling job and, in particular, failures. Analyzing these drilling reports could potentially bring insight into operation patterns, risk mitigation activities and to analyze environmental hazards.

The first case study applied different deep learning NLP methods via TensorFlow and Keras for classifying sentences within each drilling report to different labels [1]. The drilling reports were categorized as Productive Time (PT) and Non-Productive Time (NPT) Reports. The NPT reports are created whenever an action is taken to mitigate a failure or incident. A drilling expert labels each sentence within an NPT report as an EVENT, SYMPTOM or ACTION. EVENTs are major accidents or failure during a drilling operation. SYMPTOMs are lesser faults which if not treated could become EVENTs. ACTIONs are taken after EVENTs occur to mitigate the failure or after a SYMPTOM is witnessed to prevent a major failure. These labels are used to train the models outlined in the paper.

The workflow for the first case study begins by pre-processing and cleaning the drilling reports. Text was concatenated in chronological order and python expressions were used to remove meaningless symbols. The corpus's token size is T = 810,375 and vocabulary size is V = 17,623. A Mikolov et. al. methodology is then followed in order to create word embeddings from the corpus. The skip-gram model was used to predict the context of words given a center word. Skip-gram was used over continuous bag-of-words (CBOW), the inverse of skip-gram, due to the relatively small size of the corpus and its many infrequent words. The training technique used was negative sampling which intends to maximize the probability of word pairs within the same context and minimizes word pairs for k randomly selected words from the vocabulary. For the study, the window size for the corpus was m = 3 and each word in the vocabulary was assigned two random vectors in which their values were $[-1,1]^d$ with the embedding dimension d = 300. A k = 64 words were sampled randomly from the distribution of each word in order to fulfill the negative sampling technique via an objective function. The k value was set high due to the relatively smaller size of the dataset. The word vectors were updated iteratively via stochastic gradient descent and the model's hyper-parameters were tuned by trial and error.

The output vectors represent weights in a deep neural network trained for the task of classifying each sentence of a drilling report to one of the labels (EVENT, SYMPTOM and ACTION). The architecture of the neural network would be a major factor in the classification accuracy due to the dependence of a network's ability to memorize context information. The tested neural networks architectures were a simple network with arithmetic averaging, a convolutional neural network (CNN) and a long short-term memory network (LSTM). Each neural network was trained using classified sentences from NPT reports. Of the NPT reports, 80% were used for training and the remaining 20% were used for testing.

The results from the study showed that the LSTM network was the most accurate with a mean accuracy of 82.7%. LSTM was superior due to its capability to memorize contexts from text. Most misclassifications occurred for EVENT and SYMPTOM labels. This was due to the imbalance of labeling within the NPT reports since sentences were labeled 28% as EVENTs, 15% as SYMPTOMs and 57% as ACTIONS. With the classification results from the neural networks, various kinds of queries can be enabled by operators of the corresponding wells. Some possible queries are determining the most problematic wells, retrieving all wells with a specified sequence of SYMPTOMs and ACTIONs and the success rate of remediation ACTIONs.

Another important source of textual data regarding oil and gas are geology reports for existing wells. The second case study outlines multiple attempts to find geological analogues via word embeddings [2]. The intention for determining these analogues, is to improve expert interpretation of wells and improve exploration success. The first attempts at creating word embeddings for the geology reports involved using the open source pre-trained word2vec and GloVe. Both word2vec and GloVe are trained on Google News and Wikipedia, respectively. The training datasets are an issue since a word like “channel” in a geology report could be similar to “cartoon network” via word2vec but similar to “television” via GloVe. A “channel” in geological context refers to a natural formation in land that is filled with fluid. A domain specific approach was needed to create appropriate word embeddings for the geology reports. Textbooks regarding the appropriate subdomain of geology were leveraged. However, attempts to leverage geology dissertations were abandoned since the content of these dissertations are project and region specific, thereby being a source of bias for the embeddings. The textbook corpus was then used to train word2vec models which resulted in better word embeddings than the pre-trained word2vec and GloVe models.

The trained word2vec models were a significant improvement. However, according to geology domain experts, the results lacked understanding of the domain’s context. A second approach was undertaken using a Universal Sentence Encoder (USE) to encode multiple sentences/paragraphs. USE was utilized to encode multiple sentences describing a geological rock formation. This approach was an aspect of information retrieval in which a user could query for analogous rock formations.

The second case study also explores the possibility of using the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT). The idea would be to create GilBERT: Geologically informed language modeling with BERT. GilBERT would encode the context from geology text sources in order to develop a geology domain language model that can be used in the Oil and Gas industry. Unfortunately, GilBERT is still in development and attempts to pre-train BERT with a geology corpus have been unsuccessful due to the relatively small size of the corpus.

State of the art deep learning methods and encodings are currently being explored and utilized on text data sources of the oil and gas industry. Results from the first case study were successful at creating word embeddings and the USE approach within the second case study were successful at creating multi-sentence encodings. However, it’s evident that the major limitation within both these studies is the relatively small size of each corpus. A collaborative approach should be considered by oil and gas operators in order to increase the size of each corpus and therefore improve the accuracy of either model.

[1] Hoffimann, J., Mao, Y., Wesley, A., & Taylor, A. (2018, September 24). Sequence Mining and Pattern Analysis in Drilling Reports with Deep Natural Language Processing. Society of Petroleum Engineers. doi:10.2118/191505-MS

[2] Bayraktar, Z., Driss, H., & Lefranc, M. (2019). Representation Learning in Geology and GilBERT.