EPiC Series in Built Environment

Volume 5, 2024, Pages 541–548

Proceedings of 60th Annual Associated Schools
of Construction International Conference

Comparing Two Artificial Intelligence Language Modeling to Evaluate Construction Schedule Understanding

Tulio Sulbaran, Ph.D.
The University of Texas at San Antonio
San Antonio, Texas

Construction schedules are critical for construction projects. Creating a construction schedule is complex. Because it requires a keen ability to understand construction documents and methods. Furthermore, a single construction project could be completed following multiple construction schedules. The preparation of construction schedules is currently done by humans. However, with the advent of Artificial Intelligence (AI), it is possible to think of a future where humans could be assisted by an AI to develop construction schedules. An important step toward having an AI assisting humans in the creation of construction schedules is for the AI to understand construction schedules. Thus, this paper presents the comparison of two language modeling to evaluate an AI's ability to understand construction schedules. This comparison was done following a quantitative experimental research methodology where the independent variables were two AI language models (the Bidirectional Encoder Representation Transformers and the Masked and Permuted Pre-training for Language Understanding) and the dependent variables were accuracy, precision, recall, and F1 scores of the language modeling to understand construction schedule activities. The results demonstrate the impact of language models on the ability of the AI to understand construction schedules. The Masked and Permuted Pre-training for Language Understanding language model had an overall superior performance in understanding construction scheduling activities. This is important as supports the need to expand research projects as the one presented in his paper to identify the best language modeling and improve it for the construction industry.

Key Words: Natural Language Processing, Language Models, Evaluating, Construction Schedule.

Background

Construction scheduling is a complex process with a lot of considerations for successful project delivery (Okonkwo et al., 2022). Poor scheduling can result in considerable waste (Sulbaran & Ahmed, 2017). Standard scheduling software lacks the ability to execute thousands of potential planning scenarios without direct input from the user (Hatami et al., 2022). Currently, construction schedules are almost always performed manually (with the aid of scheduling software), by experienced practitioners (Amer & Golparvar-Fard, 2021) and the final schedule depends on the expertise of individual schedulers (Le & Jeong, 2020). This current project-specific and scheduler-specific approach to prepare schedules does not benefit from years of experience accumulated by other schedulers in similar situations and, in turn, cannot guarantee that the generated schedules are best (Amer & Golparvar-Fard, 2021).

AI is poised to rapidly transform businesses, particularly the construction industry (Sulbaran, 2023). AI describes the theory and development of computer systems to perform tasks that normally would require human cognition, such as perception, language understanding, reasoning, learning, planning, and problem-solving (Nelson et al., 2020). Researchers have been exploring the use of AI in construction scheduling for almost three decades (Faghihi et al., 2015). Although a considerable amount of engineering data increases unprecedently in construction projects, the adoption of AI techniques in construction still lags behind the process in other industries (Pan & Zhang, 2021). Researchers have tried to develop AI technologies to decrease the dependence level of experts in construction planning and schedule control (Liu et al., 2018) but the industry adoption is very weak.

AI could leverage the data for automated monitoring of site progress, early detection of potential scheduling problems, optimization of construction logistics and scheduling, as well as other purposes (Pan & Zhang, 2021). AI-based scheduling techniques can help minimize the planning delays of a construction project by enabling schedulers to create multiple schedules that are optimized based on a given data set (Hatami et al., 2022). AI-based scheduling offers the potential to drastically reduce construction project delays and costs while improving the efficiency and accuracy of the scheduling process (Faghihi et al., 2015).

One of the fields of AI is Natural Language Processing (NLP) which focuses on the computers' ability to understand and process human language. There are many NLP methods, but the most common are: rule-based methods, statistical methods, and neural network methods. One of the NLP statistical methods is language modeling. Language modeling can be used for a wide range of tasks such as textual entailment, question answering, semantic similarity assessment, and document classification (Radford et al., 2018). Language modeling predicts the next word in a sequence of words by developing long-term dependency, which is the ability to understand the relationship between words that are far apart in a sequence. This can be challenging as ideally it should model both (1) complex characteristics of word use (i.e., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy) (Peters et al., 2018). Thus in 2017 Vaswani et al, introduced the transformer architecture based solely on attention mechanisms, dispensing with recurrence, and convolutions entirely (Vaswani et al., 2017) allowing the development of sentence transformer, which is specifically designed to learn long-range dependencies between words in a sentence. Thus, the objective of this paper is to compare two of those sentence transformer models to evaluate an AI's ability to understand construction schedule activities.

Methodology

A quantitative experimental methodology was used to conduct this research with the hypothesis that language models have an impact on the artificial intelligence (AI) ability to understand construction schedule activities. The independent variables and dependent variables are shown in Table 1.

Table 1. Research dependent and independent variables.

Independent Variable	Dependent Variables
1. Bidirectional Encoder Representation Transformers (BERT)	a. Accuracy, b. Precision
2. Masked and Permuted Pre-training for Language Understanding (MPNET)	b. Recall, & d. F1

BERT and MPNET were selected as the independent variables for this research project because BERT is one of the most successful models (Song et al., 2020) and the most popular pre-trained transformer models (Devlin et al., 2019). However, BERT ignores the dependency among the masked (and to be predicted) tokens (Yang et al., 2020). On the other hand, MPNET leverages the dependency among predicted tokens through permuted language modeling and takes auxiliary position information as input to make the model see a full sentence thus reducing the position discrepancy (Song et al., 2020). The accuracy, precision, recall, and F1 score were selected as the dependent variables because they are the most commonly used metrics of Artificial Intelligence performance (Yacouby & Axman, 2020).

As shown in the Figure 1, the experiment was done in four stages. Data collection (first stage) consisted of obtaining a construction project schedule and the Construction Specifications Institute (CSI) Master Format (MF) 50 divisions activity codes with their descriptions. AI training and preparation (second stage) used the BERT and MPNET language modeling to encode both the construction project schedule activities and the CSI Master Format activities as well as the automatic generation of a Qualtrics survey to facilitate gathering information from humans about the schedule. Schedule activity interpretation (third stage) involved the humans, BERT, and MPNET language models to decipher the construct project schedule activities in association with the CSI master format. A Comparison of BERT AND MPNET Construction Schedule Comprehension (fourth stage) evaluated the BERT and MPNET understanding of the construction scheduling activities organized by quartiles and then identified the similarities and differences between dependent variables (accuracy, precision, recall, and F1 Score) for both independent variables (BERT and MPNET) using the confusion matrix.

Data Collection
A.I. Training and Prep
Activity Interpretation
Comparison

Figure 1. Research Experiment Four Stages

Results

Data Collection

The construction project selected for this research was a 2 million United States (U.S.) dollars, 6,500 SF courthouse in a city in the United States. The courthouse was a single-story, structural steel frame, metal stud, and gypsum board partitions, with exterior brick and block. The schedule included 94 activities starting with the notice to proceed construction and ended with final completion. Since the CSI MF is the standard used in the U.S. for organizing specifications, its 7533 individual activities grouped in 50 divisions were also used.

Artificial Intelligence Training and Preparation

Python was used to program the AI because it lets users easily express many complex and high-level tasks concisely, as well as offers a good platform for developing more specialized objects that are directly suited to scientific work (Perez et al., 2011) such as sentence transformer framework to compute semantic similarity and develop the language models (Devika et al., 2021) to answer the research question. Language models (such as BERT and MPNET) represent individual words with semantically fixed-length vectors that make possible natural language processing (NLP) (Greiner-Petter et al., 2020). Thus, the results of this stage were vector representations of all construction activities in the construction schedule as well as in the 7532 activities in CSI MF. This stage also resulted in the automated generated close-end Qualtrics survey questions.

Schedule Activity Interpretation

Using the Qualtrics survey questions, the humans interpreted the construction schedule activities matching them to the standard CSI MF used in the U.S. Given that the ideal survey should have a median length of 10 minutes and a maximum length of 20 minutes (Revilla & Ochoa, 2017) and answering questions for 94 construction activities within that time was not possible, only a randomized sub-set of 18 of the 94 questions were provided to each participant resulting in a total of 316 answers. During, this stage the BERT and MPNET models were also implemented to interpret the construction schedule activities using the CSI MF and the cosine of similarity a to determine the models' interpretation result.

Comparison of BERT and MPNET Construction Schedule Comprehension

There were eighteen participants in the research that interpreted the construction schedule activities. The participants' group was composed of: 77.8% Hispanics, 61.1% between 20 and 24 years old, 77.8% males, and 55.6% with 1 to 5 years of work experience. The participants' answers were considered to be the true interpretation of the construction activities and therefore the BERT and MPNET models' interpretation must match the human interpretation to be considered correct. Furthermore, based on the cosine of similarity, the responses were divided into quartiles as shown in Table 2. The BERT and MPNET language models were able to identify the construction activities on average 50% and 68.6% respectively for the first three quartiles. Likewise, the BERT and MPNET language models were able to identify the construction activities at the lower rate of 25% and 58.3% respectively for the fourth quartile. Thus, these initial results demonstrate that the MPNET model is better at comprehending the construction activities than the BERT model.

Table 2. Quartiles BERT and MPNET Match response based.

Quartile	Number of Activities	Match Number and Percent		No Match Number and Percent		Match Number and Percent		No Match Number and Percent	
		BERT	MPNET	BERT	MPNET	BERT	MPNET	BERT	MPNET
1st 25%	24	14 (58.3%)	19 (79.2%)	10 (41.7%)	5 (20.8%)	35 (50.0%)	48 (68.6%)	35 (50.0%)	22 (31.4%)
2nd 25%	23	10 (43.5%)	15 (65.2%)	13 (56.5%)	8 (34.8%)				
3rd 25%	23	11 (47.8%)	14 (60.9%)	12 (52.2%)	9 (39.1%)				
4th 25%	24	6 (25.0%)	14 (58.3%)	18 (75.0%)	10 (41.7%)	6 (25.0%)	14 (58.3%)	18 (75.0%)	10 (41.7%)

The comparison of the BERT and MPNET language model's ability to comprehend construction scheduling activities was done using a confusion matrix. A confusion matrix was used because it represents the prediction summary in a matrix form (Tiwari, 2022) and it is a tool used to determine the performance of the AI useful to identify areas where improvement is needed. The confusion matrix is useful because shows how many predictions are correct (true) and incorrect (false) per class (Tiwari, 2022). In this research, the two classes were: 1- AI expected to identify the construction activities (top three quartiles), and 2- AI was not expected to identify the construction activities (bottom quartile). The values used to prepare the confusion matrix for the BERT and MPNET language model are summarized in Table 2.

As shown in Figure 2, a confusion matrix was prepared for the BERT and MPNET language models (independent variable). Each confusion matrix was composed of four possible outcomes (true positive, false negative, false positive, and true negative) organized in four quadrants:

- Top left quadrant: number of true positives corresponding to the A.I. correctly identifying the construction activities when it was expected to identify them.
- Top right quadrant: number of false positives (also known as type I error), which are cases where the A.I. was expected to identify the construction activities, but it provided the wrong construction activity interpretation.
- Bottom left quadrant: number of false negatives (also known as type II error), which are cases where the A.I. implementation was not expected to identify the activity but was, in fact, able to identify it.
- Bottom right quadrant: number of true negatives, which are cases where the A.I. was not expected to identify the construction activity and as expected it was not able to identify the construction activity.

		Actual Activity			
		BERT		MPNET	
		ID	NID	ID	NID
AI Activity Interpret	ID	TP (1) 35	FP (2) 35	TP (1) 48	FP (2) 22
	NID	FN (3) 6	TN (1) 18	FN (3) 14	TN (1) 10

Legend:

ID = Identified NID = Not Identified

TP = True Positive , FP = False Positive , FN = False Negative , TN = True Negatives

(1) Correct predictions (2) Type I Error (3) Type II Error

Figure 2. AI Interpretation of Construction Activities Confusion Matrix.

As shown in Table 3, the confusion matrix information was used to calculate the dependent variables (accuracy, precision, recall, and F1 Score) for both independent variables (BERT and MPNET). The results indicate that the MPNET had a higher accuracy than the BERT meaning that MPNET is better at correctly understanding the construction schedule activities. The precision of the MPNET was higher than BERT meaning that the MPNET is able more often to correctly understand the activities when is expected to comprehend activities. The recall of the BERT was higher than the MPNET meaning that the BERT more often correctly identifies an activity whether or not is expected to identify it. The F1 Score value of the MPNET was higher. The F1 score is a more balanced measure than precision or recall alone and therefore provides a more complete picture indicating that the MPNET language model is better overall than the BERT language model in understanding construction scheduling activities.

Table 3. Results Construction Schedule Comprehension BERT vs MPNET.

Dependent Variable	Equations	Independent Variables BERT	MPNET
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	0.56	0.63
Precision	$\frac{TP}{TP + FP}$	0.50	0.69
Recall	$\frac{TP}{TP + FN}$	0.85	0.77
F1 Score	$\frac{2 * \ Precision * Recall}{Precision + Recall}$	0.63	0.73

Summary and Future Work

Given the importance of the schedules for the construction industry and the advent of Artificial Intelligence, the research project presented in this paper is both important and timely. This paper presented the comparison of two language models to evaluate an AI's ability to understand construction schedules. The two language models compared were the Bidirectional Encoder Representation Transformers (BERT) and the Masked and Permuted Pre-training for Language Understanding (MPNET). The comparison was done using A.I. performance benchmarks accuracy, precision, recall, and F1 score. The results indicated that the overall MPNET language model is superior (than BERT) in comprehending construction scheduling activities. MPNET also had higher accuracy and precision while BERT had a better recall.

The results of this paper are important not only by providing specific metrics regarding the two language models compared, but also by opening a window of opportunity for further AI research for construction scheduling. The researcher is particularly interested in partnering with other researchers to 1- Expand the number of construction projects comprehended by the AI, 2- Use other AI performance metrics such as Area Under the Curve (AUC) of Receiver Operating Characteristics (ROC), 3 – Compare other language model performance, 4- Explore the possibility of using Term Frequency – Inverse Document Frequency (TF -IDF) as the metric to determine the model interpretation result, 5- Evaluate the impact of modifying the quartiles used in this paper for other percentiles, just to mention a few of the possible future research.

References

Amer, F., & Golparvar-Fard, M. (2021). Modeling dynamic construction work template from existing scheduling records via sequential machine learning. Advanced Engineering Informatics, 47, 101198. https://doi.org/10.1016/j.aei.2020.101198

Devika, R., Vairavasundaram, S., Mahenthar, C. S. J., Varadarajan, V., & Kotecha, K. (2021). A Deep Learning Model Based on BERT and Sentence Transformer for Semantic Keyphrase Extraction on Big Social Data. IEEE Access, 9, 165252–165261. https://doi.org/10.1109/ACCESS.2021.3133651

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423

Faghihi, V., Nejat, A., Reinschmidt, K. F., & Kang, J. H. (2015). Automation in construction scheduling: A review of the literature. The International Journal of Advanced Manufacturing Technology, 81(9), 1845–1856. https://doi.org/10.1007/s00170-015-7339-0

Greiner-Petter, A., Youssef, A., Ruas, T., Miller, B., Schubotz, M., Aizawa, A., & Gipp, B. (2020). Math-word embedding in math search and semantic extraction. Scientometrics. https://doi.org/10.1007/s11192-020-03502-9

Hatami, M., Franz, B., Paneru, S., & Flood, I. (2022). Using Deep Learning Artificial Intelligence to Improve Foresight Method in the Optimization of Planning and Scheduling of Construction Processes. 1171–1178. https://doi.org/10.1061/9780784483893.143

Le, C., & Jeong, H. D. (2020). Artificial Intelligence Framework for Developing a Critical Path Schedule Using Historical Daily Work Report Data. 565–573. https://doi.org/10.1061/9780784482889.059

Liu, N., Kang, B. G., & Zheng, Y. (2018). Current Trend in Planning and Scheduling of Construction Projects Using Artificial Intelligence. 11-. https://doi.org/10.1049/cp.2018.1731

Nelson, S. D., Walsh, C. G., Olsen, C. A., McLaughlin, A. J., LeGrand, J. R., Schutz, N., & Lasko, T. A. (2020). Demystifying artificial intelligence in pharmacy. American Journal of Health-System Pharmacy: AJHP: Official Journal of the American Society of Health-System Pharmacists, 77(19), 1556–1570. https://doi.org/10.1093/ajhp/zxaa218

Okonkwo, C., Garza, R., Sulbaran, T., & Awolusi, I. (2022). A Review of Genetic Algorithm as a Decision-Making Optimization Tool in Project Management. EPiC Series in Built Environment, 3, 254–262. https://doi.org/10.29007/jzcl

Pan, Y., & Zhang, L. (2021). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Automation in Construction*, *122*, 103517. https://doi.org/10.1016/j.autcon.2020.103517

Perez, F., Granger, B. E., & Hunter, J. D. (2011). Python: An Ecosystem for Scientific Computing. Computing in Science & Engineering, 13(2), 13–21. https://doi.org/10.1109/MCSE.2010.119

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. North American Chapter of the Association for Computational Linguistics.

Radford, A., Narasimhan, K., Salimans, Tim, & Sutskever, Ilya. (2018). Improving Language Understanding by Generative Pre-Training. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf
Revilla, M., & Ochoa, C. (2017). Ideal and maximum length for a web survey. International Journal of Market Research, 59(5), 557–565. https://doi.org/10.2501/IJMR-2017-039

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. Advances in Neural Information Processing Systems, 33, 16857–16867. https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html

Sulbaran, T. (2023). Evaluating the Comprehension of Construction Schedules of an Artificial Intelligence. 23° International Conference on Construction Applications of Virtual Reality.

Sulbaran, T., & Ahmed, F. (2017). Expert System for Construction Scheduling Decision Support Based on Travelling Salesman Problem. Proceedings of the 53rd ASC Annual International Conference Proceedings, Seattle, WA, USA, 545–556.

Tiwari, A. (2022). From theory to applications—Chapter 2—Supervised learning: In R. Pandey, S. K. Khatri, N. kumar Singh, & P. Verma (Eds.), Artificial Intelligence and Machine Learning for EDGE Computing (pp. 23–32). Academic Press. https://doi.org/10.1016/B978-0-12-824054-0.00026-5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Yacouby, R., & Axman, D. (2020). Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, 79–91. https://doi.org/10.18653/v1/2020.eval4nlp-1.9

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). XLNet: Generalized Autoregressive Pretraining for Language Understanding (arXiv:1906.08237). arXiv. https://doi.org/10.48550/arXiv.1906.08237