Joint BERT Model for Intent Classification and Slot Filling Analysis of Natural Language Instructions in Co-Robotic Field Construction Work

Somin Park[1]; Carol C. Menassa[2]; and Vineet R. Kamat[3]

[1]Ph.D. Candidate, Dept. of Civil and Environmental Engineering, Univ. of Michigan, Ann Arbor, MI. Email: somin@umich.edu
[2]Professor, Dept. of Civil and Environmental Engineering, Univ. of Michigan, Ann Arbor, MI (corresponding author). Email: menassa@umich.edu
[3]Professor, Dept. of Civil and Environmental Engineering, Univ. of Michigan, Ann Arbor, MI. Email: vkamat@umich.edu

ABSTRACT

As the construction industry faces the challenges of a worker shortage and low productivity rate, there is a growing interest in using human-robot collaboration (HRC) in construction. HRC allows for a combination of the accuracy and repeatability of robots with the flexibility and intelligence of human workers. To take advantage of its potential benefits, it is important for construction workers who are not robotic experts to interact easily with robots through intuitive and natural user interfaces. Even though many studies have been performed on HRC using natural language, little research has been conducted on this topic in construction. This paper conducts natural language understanding of language instructions for pick-and-place operations in construction using the language model Joint BERT. Experimental results show high accuracy on intent classification and slot filling tasks, allowing the robot to perform tasks accurately for a given natural language instruction.

INTRODUCTION AND BACKGROUND

Robotics is considered a promising solution to tackle problems related to labor shortages and stagnant productivity growth in construction (Delgado et al. 2019; Cai et al. 2023). However, robots face difficulties in working on construction sites due to the unpredictable and unstructured work environment and varying project conditions (Feng et al. 2015; Pan et al. 2020). To overcome these challenges, Human-Robot Collaboration (HRC) has emerged as a potential strategy with its advantages of higher level of productivity, safety, and flexibility.

Effective communication among team members is critical in construction projects due to the dynamic and unpredictable nature of the work environment, which increases the likelihood of errors (Cupido 2009). Likewise, in the construction industry, where collaborative robots work alongside human workers, the interaction between humans and robots is critical (Delgado et al. 2019). To communicate plans from human workers to robots, user-friendly interfaces are necessary for human operators. However, designing intuitive interfaces is a significant challenge for HRC due to the specialized knowledge required for interacting with robots (Villani et al. 2018). Natural and intuitive interaction allows human operators to easily work with robots and leverage human skills to improve productivity (Maurtua et al. 2017; Villani et al. 2018). Moreover, intuitive interaction can reduce the learning curve for novice operators and minimize fatigue levels.

Natural language-based interaction, which utilizes speech input, has gained attention for its benefits in the field of robotics (Hatori et al. 2018; Ye et al. 2021). Natural language instructions

enable human operators to communicate their requests accurately and efficiently (Liu and Zhang. 2017). Natural language provides a precise way for users to express their intent about actions, tools, workpieces, and location in HRC, without losing information compared to other simplified requests (Liu et al. 2016; Paul et al. 2016). For this reason, language instructions have been utilized to make robots perform pick-and-place operations, which are among the most common tasks of industrial robots. Previous studies have investigated methods to analyze language instructions for the pick-and-place operations, such as extracting information about the final location and identifying everyday workpieces by their color, name, or spatial relationships (Bisk et al. 2016; Hatori et al. 2018; Magassouba et al. 2019; Murray and Cakmak 2022). However, limited research has been conducted on the use of natural language-based interaction in construction.

The objective of this paper is to extract semantic information from natural language instructions for pick-and-place operations in construction tasks, which can allow robots to perform tasks. It is assumed that construction workers have access to building component information, enabling users to describe target objects and destinations based on their IDs, dimensions, and positions. The body of this study builds upon the authors' previous work (Park et al. 2023), which proposed a natural language-enabled HRC system framework. The main differences between the two studies are outlined in Table 1. This study extends a Natural Language Understanding (NLU) module of the previous study by employing Joint BERT (Chen et al. 2019). While Park et al. (2023) focused on slot-filling tasks in the NLU module, this study performs both slot-filling and intent classification. In the experiments, two types of datasets, labeled with different types of intents and tags, are analyzed in this study.

Table 1. Comparison of differences between the study (Park et al. 2023) and this study.

	Park et al. (2023)	This study
Modules	Natural Language Understanding (NLU), Information Mapping (IM), and Robot Control (RC)	Natural Language Understanding (NLU)
Output of NLU	Slot filling	Slot filling and intent classification
Language models	BiLSTM-CRF (Huang et al. 2015) BERT (Devlin et al. 2018)	Joint BERT (Chen et al. 2019)
Dataset 1– single panel installation	1,584 language instructions	1,074 language instructions
Dataset 2 – single and multi panel installation	-	635 language instructions

METHODOLOGY

In this study, natural language instructions are analyzed for intent classification and slot filling. Intent classification is a classification task in which one label is predicted for each query, while slot filling is a sequence labeling task that assigns a suitable label for every word.

Joint BERT

To conduct a joint intent classification and slot filling, a Joint BERT model (Chen et al. 2019), which exploits the relationship between two tasks, is utilized in this study. Joint BERT is based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018). BERT is a powerful language model based on deep neural networks by jointly conditioning on left and right context in a multilayer transformer structure. To understand the context of words, BERT uses two unsupervised tasks, which are Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In BERT, a special classification embedding 'CLS' is added as the first token of input data as shown in Figure 1. The embedded representation (E) of input data is passed to the next layer, which is Transformer network (TRM) that produces a hidden state.

In Joint BERT, intent and slot tags are predicted using a softmax classifier as follows and the two predicted values are used in a learning objective to jointly train the model.

$$y_I = \text{softmax}\,(w_I h_1 + b_I) \tag{1}$$

where y_I, w_I, h_1 and b_I are a predicted intent, weight matrix, hidden state of the first special token, and bias matrix.

$$y_{S,j} = \text{softmax}\left(w_s h_j + b_s\right), j = 1,2,3, \ldots N \tag{2}$$

where $y_{S,j}, w_s, h_j, b_S$ and N are a predicted j-th slot, weight matrix, the final hidden states, bias matrix, and the number of tokens.

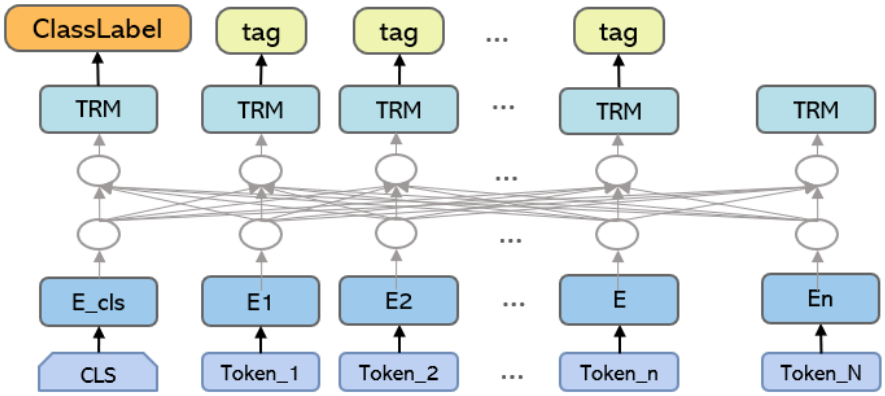
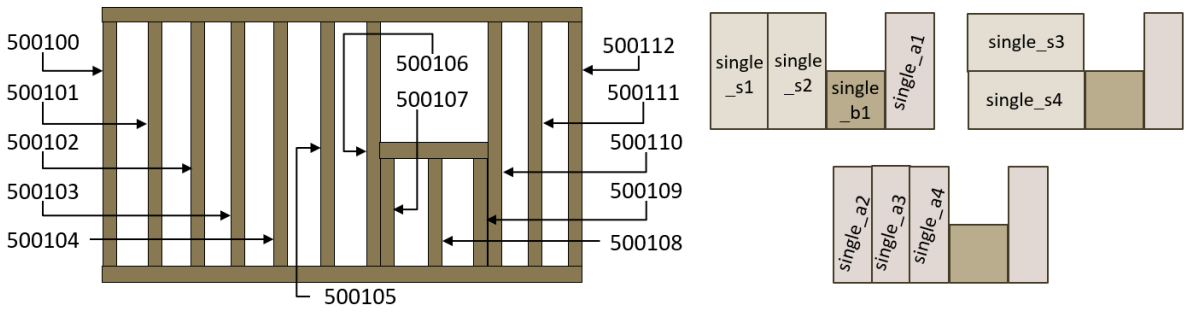


Figure 1. BERT for joint intent classification and slot filling

Dataset 1

Dataset 1 is comprised of 1,074 language instructions, which are a subset of the dataset created by Park et al. (2023). The instructions relate to drywall installation on a single stud wall depicted in Figure 2(a). In each instruction, a stud is referred to the destination of a pick-and-place operation, and is described by its ID or position. A drywall panel, a target of the operation, is described by its ID, size, or position. Figure 2(b) illustrates the available layouts of the drywall panels, which include three different types of panels. For slot filling, this study employs 13 tags

related to the target, destination, and placement orientation, which were labeled by Park et al. (2023). In this study, an intent labeling task was carried out to classify each instruction as one of the possible panel placements shown in Figure 2(b). For example, the instruction to install a full-size drywall on the leftmost stud 500100, as shown in Figure (c), corresponds to the arrangement of single_s1 in Figure (b), and thus the intent of the instruction is labeled as single_s1.



(a) (b)

Language command	please install the full size drywall to the top part of the stud 500100.
Output of NLU Intent & slot	**Intent** single_s1 (single standard – arrangement 1) **Slot** full size = dimension top part = horizontal_top 500100 = stud id

(c)

Figure 2. An experimental design for Dataset 1 (a) single stud wall (b) layouts (c) examples of prediction.

Dataset 2

Dataset 2, similar to the first dataset, involves the installation of drywall panels on a single stud wall. Figure 3(a) illustrates the single stud wall, while Figure 3(b) shows the layout of panels that can be installed on it. In this dataset, we consider a scenario where two different panels are available. Unlike the first dataset, which only provides instruction for installing a single panel, the language instructions in Dataset 2 contain information on installing up to three panels in a single sentence. Intent of the language instruction is determined by the number of panels to be installed, including single, multi_two, and multi_three. Slot filling tags are identified by numbers that indicate the installation order. For examples, corresponding words to represent target objects for the first and second installation are labeled as target1 and target2, respectively. Words indicating multiple targets are marked as target12, target23, or target123. For instance, as shown in Figure 3(c), the intent of the language command for installing two standard panels is multi_two. The words 'standard wall panels' are for the first and second installations, so that they are labeled as target12, while '102' and '104' are tagged as destination1 and destination2, respectively.

Model parameters

This study utilized the uncased BERT-base model from Hugging face transformers library, which provides a pre-trained model for lowercased English language. The specific parameters

used to train both datasets were as follows: 12 transformer layers, 12 attention-heads, 768 hidden states, 16 batch size, and the dropout is 0.1, and a learning rate 5e-5 using Adam. Dataset 1 was split into 967 training samples and 107 testing samples while Dataset 2 was split into 563 training samples and 62 testing samples.

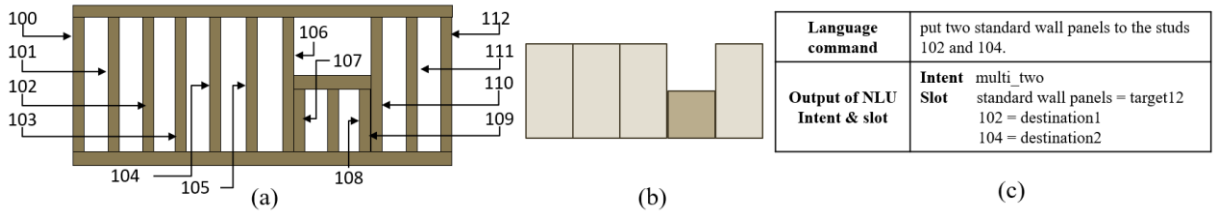


Figure 3. An experimental design for Dataset 2 (a) single stud wall (b) layouts (c) examples of prediction.

RESULTS

Figure 4 displays the training accuracy of Joint BERT for Dataset 1 and Dataset 2. Both graphs show a high accuracy converging towards 1, with Dataset 2 converging much faster than Dataset 1. Table 2 provides the performance of Joint BERT on the test dataset, highlighting the excellent results achieved for intent classification and slot filing. For the test set of Dataset 1, the intent classification achieved an accuracy of 0.9720, with only three false predictions. These were single_s1 being incorrectly predicted as single_a2, single_a2 being incorrectly predicted as single_s1, and single_a3 being mispredicted as single_a2. The first two errors may be due to the similarity in language instructions for installing the panel on the most left stud. For example, two instructions for the errors were 'I want you to take the drywall sheet 500300 and position it to the stud 500100 vertically' and 'I want you to take the drywall piece 500310 put it vertically into the stud the stud is placed on the left to the stud 500101.' For the slot filling tasks, both precision and recall for Dataset 1 were 0.9959, with only two errors in the entire dataset. In a phrase 'the length of the panel is 4 and its width is 4', the values corresponding to length and width were predicted in reverse.

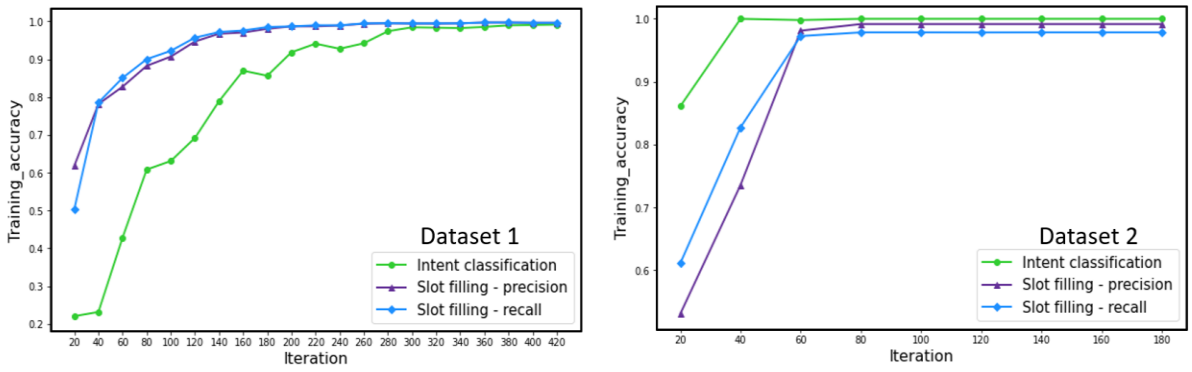


Figure 4: Training accuracy of the Joint BERT for Dataset 1 and Dataset 2

Table 2 demonstrates that all 62 test datasets of Dataset 2 were accurately predicted. Unlike Dataset 1, which included cases where both studs and panels were described in relative positions,

Dataset 2 did not include descriptions of position in its instructions, making prediction less challenging. In addition, Dataset 2 had fewer intent labels and slot tags than Dataset 1, potentially contributing to the rapid increase in training accuracy shown in Figure 4.

Table 3 shows examples of successful intent classification results. For Dataset 1, accurate intent prediction was achieved for examples where the target and destination were described by id, size, and position. The predicted intent represents one of the possible ways to place the panel, thus conveying information about the target type, stud id, and placement orientation. For example, if single_s3 is predicted as an intent, it is interpreted as installing the standard panel horizontally on the top row of the most left stud. Combining the results of intent prediction and slot filling in Dataset 2 enables easy extraction of pairs of target objects and studs. For the sentence predicted as multi_two in Table 3, two pairs of target panels and studs can be extracted. 100 and 102 were predicted as destination1 and destination2, respectively, and 'standard panels' were predicted as target12 as a result of slot filling. As a result, pairs of (standard - 100) and (standard - 102) can be obtained.

Table 2. NLU Performance on test dataset of Dataset 1 and Dataset 2.

Dataset	Intent classification -accuracy	Slot filling -precision	Slot filling -recall
Dataset1	0.9720	0.9959	0.9959
Dataset2	1.00	1.00	1.00

Table 3. Examples of intent prediction on test dataset of Dataset 1 and Dataset 2.

Language commands of the Dataset 1	Intent	Language commands of the Dataset 2	Intent
Move the wall panel 500320 horizontally into the stud 500100 place it to the top row.	single_s3	Once you finish installation of the standard panel, please install a standard panel to the stud 110.	single
please grab the 4 foot by 4 foot piece and hang it in the stud 500107 vertically.	single_b1	can you pick up standard panels and move the panels in the studs 100 and 102	multi_two
Install the drywall vertically in the stud that is left to the stud 500111. The panel is to the right of the full sized panel.	single_a1	take standard drywall panels and place them on the studs 100 and 110 and 102	multi_three

CONCLUSIONS

In this study, we demonstrated the effectiveness of using intent detection and slot filling to analyze natural language instructions for drywall installation. Our results achieved high accuracy, over 99%, for analyzing instructions related to single or multiple panel installation, indicating the potential for utilizing natural language instructions in collaboration with construction robots. Using the first dataset, we confirmed that placement as an intent was accurately predicted for indicating which target object to install on which stud and how to install it. Through the second dataset, we confirmed that pairs of target objects and destinations could

be obtained through the results of intent detection and slot filling. However, this study was limited to training and testing on a dataset targeting a single stud wall. Future research can consider generating language instruction datasets for various types of construction structures and complex structures, leading to a more generalized command analysis system.

ACKNOWLEDGMENTS

The work presented in this paper was supported financially by two United States National Science Foundation (NSF) Awards: 2025805 and 2128623. The support of the NSF is gratefully acknowledged.

REFERENCES

Cai, J., Du, A., Liang, X., and Li, S. (2023). "Prediction-Based Path Planning for Safe and Efficient Human–Robot Collaboration in Construction via Deep Reinforcement Learning." *Journal of Computing in Civil Engineering*, 37(1), 04022046.

Chen, Q., Zhuo, Z., and Wang, W. (2019). "Bert for joint intent classification and slot filling." arXiv preprint arXiv:1902.10909.

Cupido, G. (2009). "The role of production and teamwork practices in construction safety: A cognitive model and an empirical case study." *Joural of Safety Research*, 40(4), 265-275.

Delgado, J. M. D., Oyedele, L., Ajayi, A., Akanbi, L., Akinade, O., Bilal, M., and Owolabi, H. (2019). "Robotics and automated systems in construction: Understanding industry-specific challenges for adoption." *Journal of Building Engineering*, 26, 100868.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

Feng, C., Xiao, Y., Willette, A., McGee, W., and Kamat, V. R. (2015). "Vision guided autonomous robotic assembly and as-built scanning on unstructured construction sites." *Automation in Construction*, 59, 128-138.

Hatori, J., Kikuchi, Y., Kobayashi, S., Takahashi, K., Tsuboi, Y., Unno, Y., and Tan, J. (2018). "Interactively picking real-world objects with unconstrained spoken language instructions." In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 3774-3781.

Huang, Z., Xu, W., and Yu, K. (2015). "Bidirectional LSTM-CRF models for sequence tagging." arXiv preprint arXiv:1508.01991.

Magassouba, A., Sugiura, K., Quoc, A. T., and Kawai, H. (2019). "Understanding natural language instructions for fetching daily objects using gan-based multimodal target–source classification." *IEEE Robotics and Automation Letters*, 4(4), 3884-3891.

Maurtua, I., Fernandez, I., Tellaeche, A., Kildal, J., Susperregi, L., Ibarguren, A., and Sierra, B. (2017). "Natural multimodal communication for human–robot collaboration." *International Journal of Advanced Robotic Systems*, 14(4), 1729881417716043.

Murray, M., and Cakmak, M. (2022). "Following natural language instructions for household tasks with landmark guided search and reinforced pose adjustment." *IEEE Robotics and Automation Letters*, 7(3), 6870-6877.

Park, S., Wang, X., Menassa, C. C., Kamat, V. R., and Chai, J. Y. (2023). "Natural Language Instructions for Intuitive Human Interaction with Robotic Assistants for Pick and Place Construction Operations." *Journal of Computing in Civil Engineering*, (Tentatively Accepted; In-Re-Review).

Pan, M., Linner, T., Pan, W., Cheng, H. M., and Bock, T. (2020). "Influencing factors of the future utilisation of construction robots for buildings: A Hong Kong perspective." *Journal of Building Engineering*, 30, 101220.

Villani, V., Pini, F., Leali, F., and Secchi, C. (2018). "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications." *Mechatronics*, 55, 248-266.