ICCEPM 2022
The 9th International Conference on Construction Engineering and Project Management
Jun. 20-23, 2022, Las Vegas, NV, USA

Pilot Experiment for Named Entity Recognition of Construction-related Organizations from Unstructured Text Data

Seungwon Baek[1]*, Seung H. Han[2], Wooyong Jung[3], Yuri Kim[4]

[1] *Department of Civil and Environmental Engineering, Yonsei University, Seoul 03722, Korea,* E-mail address: baeksw@yonsei.ac.kr
[2] *Department of Civil and Environmental Engineering, Yonsei University, Seoul 03722, Korea,* E-mail address: shh6018@yonsei.ac.kr
[3] *Department of Nuclear Power Plant Engineering, KEPCO International Nuclear Graduate School, Ulsan 45014, Korea,* E-mail address: trustjung@gmail.com
[4] *Department of Nuclear Power Plant Engineering, KEPCO International Nuclear Graduate School, Ulsan 45014, Korea,* E-mail address: yurikim@naver.com

Abstract: The aim of this study is to develop a Named Entity Recognition (NER) model to automatically identify construction-related organizations from news articles. This study collected news articles using web crawling technique and construction-related organizations were labeled within a total of 1,000 news articles. The Bidirectional Encoder Representations from Transformers (BERT) model was used to recognize clients, constructors, consultants, engineers, and others. As a pilot experiment of this study, the best average F1 score of NER was 0.692. The result of this study is expected to contribute to the establishment of international business strategies by collecting timely information and analyzing it automatically.

Key words: Named Entity Recognition, Construction-related organizations, Text mining

1. INTRODUCTION

Doing business in the domestic market generally secures stable revenue growth and profit rather than in the international market. However, the growth of the domestic construction market in developed countries has stagnated whereas the international market. Thus, not only large-sized construction companies but also small- and medium-sized construction companies (SMCCs) have expanded their business abroad to grasp the opportunities of the international market [1,2].

Understanding the local business environment is crucial in order for success in international projects. Because an international project is more complicated than a domestic project because of unfamiliarity with the local business environment [3,4]. In addition, decision-making at the planning phase has a higher impact than that at later phases. Thus, construction companies, which seek to engage in international projects, continuously collect market and projects information to make informed decisions. However, there is a little information to refer to at an earlier phase. Thus, construction companies, which seek to engage in international projects, continuously collect market and projects information to establish business strategies. Large construction companies are capable to dispatch employees and hire local agents. And they even launch a local subsidiary if a market is determined as a strategic foothold. However, it is challenging for SMCCs to investigate the local business environment thoroughly due to a lack of experience, resources, and local networks [1,5].

The business environment for construction projects is ever-changing. Therefore, it is important to collect and analyze business environment information timely. Office workers usually browse internet websites and subscribe to professional reports to collect market and project information. And a large portion of business information is in textual data format. Thus, it is time-consuming and requires significant human effort to discover meaningful information by manual review. To address this problem, this study aims to develop an automated process of analyzing the business environment in international construction using Natural Language Processing (NLP). Since this study is in its early stages, this manuscript presents a pilot experiment for Named Entity Recognition (NER) model that extracts construction-related organizations from news articles.

2. RESEARCH BACKGROUND

2.1. Pre-trained language model and BERT

Pre-trained language models (LMs) have been introduced since the late 2010s and they showed outstanding performance at NLP tasks [6]. Like the pre-trained models based on ImageNet in the computer vision domain, recent pre-trained LMs uses a large corpus (e.g., Wikipedia, news and books) to develop universal language representations [7]. This approach enables a machine to learn contextual information of text data based on attention mechanism, which can avoid the vanishing gradient problem of sequence-to-sequence model [8]. After generating contextual word embeddings from pre-trained LMs, fine-tuning is applied to downstream tasks. Most of the recent LMs which show high performance at the benchmark of NLP tasks have utilized the pre-trained language representation models and fine-tuning. There are various pre-trained language representation models such as Embeddings from Language Model (ELMo) [9], Generative Pre-trained Transformer (GPT) [10], and Bidirectional Encoder Representations from Transformers (BERT) [11]. They are mainly different in model architecture; ELMo uses the Bidirectional Long Short-Term Memory (BiLSTM) whereas GPT and BERT are based on the Transformer [12]. A number of previous research have reported that the Transformer-based LM outperforms traditional the state-of-the-art LM for NLP tasks [13,14]. In addition, the GPT is optimized in generation tasks due to unidirectional training based on the Transformer's decoder structure. In contrast, the BERT is a bidirectional LM which understands a single token with considering both right and left contexts based on the Transformer's encoder structure. Therefore, the BERT-based model generally shows better performance in specific downstream NLP tasks such as Named Entity Recognition (NER) through fine-tuning [15]. Against these backdrops, this study employed the BERT model as a base pre-trained LM to develop a NER model.

The BERT model follows two steps; pre-training and fine-tuning. In the pre-training step, the model is trained with unlabeled data and generates contextualized embeddings for each token. Input representation of the BERT model consists of three embeddings, namely, token embeddings, segment embeddings, and position embeddings. The BERT model uses WordPiece embeddings as token embeddings [16]. Segment embedding is used to distinguish two sentences in a sequence; tokens of the first sentence are embedded as 0 while tokens of the second sentence are embedded as 1. Position embeddings indicates absolute positions of each token that has embedding of 0 at the beginning and embedding of the length of tokens at the last in sequence. Output representations of the BERT model is contextualized embeddings of each input token. Then, the contextualized embeddings are handed over to fine-tune the model for downstream NLP tasks. Additional neural networks (e.g., recurrent neural network (RNN)) and traditional machine learning algorithms (e.g., conditional random field (CRF)) are plugged in after the BERT model for fine-tuning.

2.2. Named Entity Recognition

NER is one of NLP tasks that aims to identify and classify terms or phrases to pre-specified categories such as person, location, and organization [17]. NER is a fundamental prior task to other advanced NLP tasks such as relation extraction. Existing NER models for typical entities with open-source datasets have reached near human-level performance [18], however, it is still difficult to apply to domain-specific topics because of its unique characteristics and lack of labeled datasets [19,20]. In the construction domain, several research have proposed NER model to extract entity information from construction documents such as construction specification [21], bridge inspection report [22-24], construction regulatory document [25,26]. And above-mentioned research applied various approach for information extraction such as rule-based approach, ontology-based approach, and neural network-based approach [27]. Some of the previous research showed satisfactory performance on its own purposes. Although it is difficult to compare each other because research design was different depending on the research purpose, models that applied the state-of-the-art technologies showed better performance on their own tasks as the level of NLP technology development improves. Yet, there is a lack of an automated approach to extract construction-related organizations from text data for the purpose of international market analysis.

3. RESEARCH METHODOLOGY

3.1. Data collection

This study collected online news articles by using web crawling technique. Web crawling is a process that automatically navigates websites and collects data defined by a user in advance. It parses Hypertext Markup Language (HTML) structure and extract information based on unique tags of each element [28]. This study attempted to crawl news articles published by MEED which is a representative magazine in the Middle East area. The authors regarded MEED as an appropriate source to get construction market information because the Middle East is a strategically important market of international construction. In the last decade, 15~20% of the revenue of international construction has come from the Middle East region [29]. Especially, Korean firms have monitored the Middle East market and collect construction information since the international projects in the Middle East area have accounted for the largest share of Korea's international construction industry. A total of 44,333 news articles were collected from 2008 to 2018.

3.2. Data labeling

The end of the line of our research is to identify a network of construction-related organizations in the international construction market to support establishing business strategies at the bidding stage. As the first step of our research, this manuscript aims to identify which organizations are active in the international construction market. Main players in any construction project are owner, contractors, suppliers, and consultants. And contractors are divided into engineers and constructors mainly. Accordingly, this study classified entities of construction-related organizations into five categories as client, engineer, constructor, consultant, and other. This study excluded suppliers from independent entities because they are usually not determined at the bidding stage, so they appeared few in our dataset. Instead, the other category covers organizations that are not included in the main categories including not only suppliers but also investors and facility operators. Meanwhile, non-construction-related organizations are not labeled to distinguish them from construction-related organizations.

There are various encoding schemes to label entities such as IO, BIO, IOU, IOE, BIEO, IEOU, BILOU [30]. Previous studies revealed that performances of each encoding scheme vary depending on language, composition of entities, and dataset [30-32]. This study employed the BIO scheme which is one of the representative encoding schemes. The BIO scheme is a common format for tagging tokens in NER tasks; ‘B’, ‘I’, ‘O’ stands for beginning, inside, and outside, respectively. A token is tagged with ‘B-’ prefix when the token is a beginning of terms or phrases, and ‘I-’ prefix when the token is the second to the end of terms or phrases. If a token does not belong to any pre-specified categories, ‘O’ is tagged.

The labeling task of this study is more complicated than general word-level labeling that can be recognized intuitively without understanding the context. Because an annotator is required to understand the context of the text and determine what kind of role an organization is in charge of. Therefore, three undergraduate students who major in civil engineering participated in the labeling task. A total of 1,000 articles were labeled and used in this study and the distribution of labeled entities is as shown in Table 1.

3.3. BERT-based NER model

This study used BERT model as a base pre-trained LM and then linear classification with the Softmax function is applied for the fine-tuning to classify entities to pre-specified categories. The overall architecture of the BERT-based NER model is as shown in Figure 1.

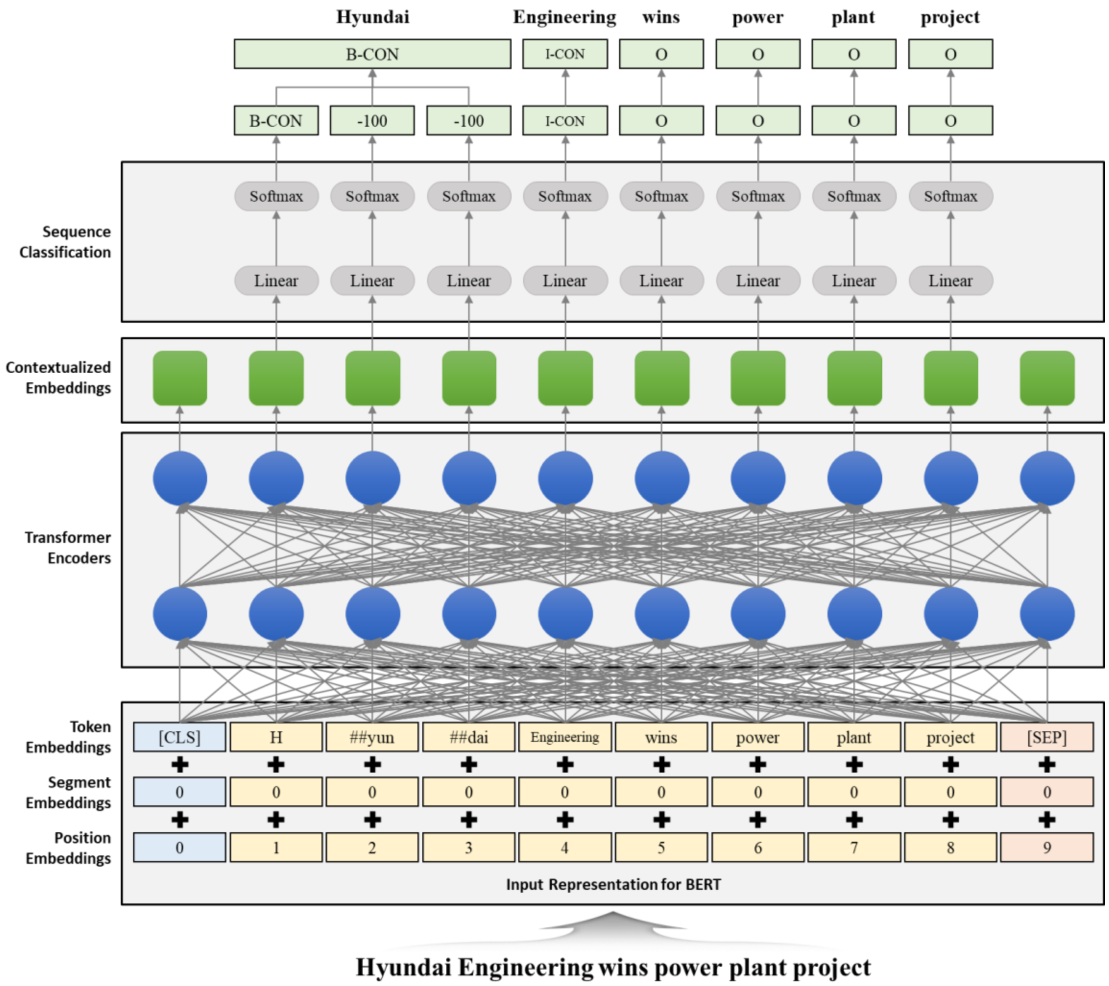

Figure 8. Architecture of the BERT-based NER model

3.4. Evaluation metrics

This study measured the performance of the BERT-based NER model based on precision, recall, and F1 score. Because the volume of labeled entities in each category is unbalanced, this study calculated the metrics for each category, then calculated the micro average of each metric. In the case of NER task, the micro average is commonly used to calculate an overall performance. The micro average calculates metrics globally to take label imbalance into account. This study used *seqeval* which is a python library for sequence labeling evaluation at entity-level of a NER task.

4. EXPERIMENT

4.1. Experimental setup

As a pilot experiment of NER model development, this study aims to investigate the effect of text length. This study performed experiments with two types of dataset; document- and sentence-level datasets. Document-level dataset was created using the original text of news articles, whereas sentence-level dataset was created by segmenting the original text of news articles into sentences so that each sentence becomes a single data point. Moreover, the document-level dataset was split randomly into training dataset, validation dataset, and test dataset at a ratio of 8:1:1. The sentence-level dataset was split before sentence segmentation using the document-level dataset to equalative a composition of entities in test datasets of both document- and sentence-level datasets. The distribution of entities is as shown in Table 1.

Table 7. Distribution of entities

Dataset	None	Client	Constructor	Engineer	Consultant	Other
Training	179,221 (93.1%)	5,692 (2.96%)	4,157 (2.16%)	593 (0.31%)	1,026 (0.53%)	1,758 (0.91%)
Test	19,255 (93.1%)	607 (2.94%)	415 (2.01%)	79 (0.38%)	128 (0.62%)	191 (0.92%)
Total	198,476 (93.1%)	6,299 (2.96%)	4,572 (2.15%)	672 (0.32%)	1,154 (0.54%)	1,949 (0.91%)

An experiment was conducted in Ubuntu 18.04 and Python version 3.7.10 using PyTorch-1.4.0, transformers-4.4.2. Python library *Transformers* provides various pre-trained deep learning models including BERT [33]. This study employed BERT-Base model to develop a NER model. The hyperparameters of the model were configured with respect to the recommendation of [11] as follows; a learning rate of 2e-5; a dropout probability of 0.1. However, a batch size was set to 4 due to the limitation of GPU memory. In addition, this study applied early stopping in training process to prevent overfitting. The training was stopped if F1 score had not improved more than five times sequentially because Devlin et al. [11] recommend the number of epochs less than five. The model was trained using k-fold cross-validation with the k of 10.

4.2. Experimental results

The performance of the NER model tested in this study is as shown in Table 2. The average F1 scores with document-level and sentence-level input data were 0.625 and 0.692 respectively. In detail, the NER model predicted 'Client' and 'Constructor', and 'Other' entities better with sentence-level dataset whereas 'Consultant' and 'Engineer' were better with document-level dataset. The reasons for this result are as follows. First, a company may play a different role in each project. For examples, an engineering company plays as a consultant such as construction management time to time in a construction project. A construction company may perform either engineering or construction depending on the occasion. These ambiguities of the role of construction-related organizations affected labeling and the performance of the NER model. Second, the model with document-level input data understands the overall context well because it was trained using a full article. Let us assume a sentence, ‘S company awarded Tabreed cooling plants contract.’ This sentence does not contain any information about what kind of contract ‘S company’ has made. A next sentence, ‘S company has received a letter of award for a contract to build two district-cooling plants at the Al-Dhafra airbase in Abu Dhabi,’ supports to determine the actual role of ‘S company’. However, this context of a full text has not been considered when the NER model was trained using the sentence-level input data. Third, the overall balance of entities is imbalanced. Only 6.9% of tokens were related to the entities in this study. Moreover, 74% of labeled entities were either ‘Client’ or ‘Constructor’. There were only 0.54% and 0.32% of ‘Consultant’ and ‘Engineer’ labels respectively in the dataset of this research.

Table 8. Performance of the NER model

Category	**Document-level**			**Sentence-level**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Client	56.9%	77.2%	65.5%	67.6%	83.4%	**74.6**%
Constructor	69.8%	74.3%	72.0%	72.0%	77.6%	**74.7**%
Engineer	65.1%	51.9%	**57.7**%	43.9%	54.7%	48.7%
Consultant	66.2%	65.6%	**65.9**%	55.5%	56.0%	55.7%
Other	34.3%	36.0%	35.1%	52.0%	58.5%	**55.0**%
Average	58.3%	67.4%	62.5%	64.8%	74.3%	**69.2**%

5. CONCLUSION

This study presented the BERT-based NER model to extract construction-related organizations from news articles. The results of this experiment revealed that a pre-trained LM is applicable to text analytics in the construction domain, especially for the business environment analysis based on relatively long text data. Because the news data used in this study is less domain-specific than other construction documents such as specifications and technical reports. Nevertheless, since this study is in its early phase, the performance of the presented model was not enough satisfactory for practical use. To improve the NER model, the authors will complement the limitations discussed in the result section in future research; using different type of pre-trained LM, adding more layers to the presented model such as BiLSTM and CRF, filtering unnecessary documents before NER, considering relations between entities. As the first step of text-based business environment analysis, this research is expected to support marketing, bid/no-bid decisions, and the strategy development of the international construction business.

ACKNOWLEGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2020R1A2C1012739 and NRF-2022R1A2C1012018).

REFERENCES

[1] W. Jung, S.H. Han, C. Park, C. Lee, S. Baek, "Three-phased risk-management benchmark for internationalization of small and medium-sized construction companies", KSCE Journal of Civil Engineering, 2021.

[2] J.K. Lee, S.H. Han, W. Jang, W. Jung, "Win-win strategy for sustainable relationship between general contractors and subcontractors in international construction projects", KSCE Journal of Civil Engineering, vol. 22, no. 2, pp. 428-439. 2017.

[3] K.-W. Lee, S.H. Han, H. Park, H.D. Jeong, "Empirical analysis of host-country effects in the international construction market: An industry-level approach", Journal of Construction Engineering and Management, vol. 142, no. 3, pp. 04015092. 2015.

[4] H. Park, S.H. Han, E.M. Rojas, J. Son, W. Jung, "Social Network Analysis of Collaborative Ventures for Overseas Construction Projects", Journal of Construction Engineering and Management, vol. 137, no. 5, pp. 344-355. 2011.

[5] A.H. Abu Bakar, M.N. Yusof, M.A. Tufail, W. Virgiyanti, "Effect of knowledge management on growth performance in construction industry", Management Decision, vol. 54, no. 3, 2016.

[6] S. Edunov, A. Baevski, M. Auli, "Pre-trained language model representations for language generation", arXiv preprint arXiv:1903.09722, 2019.

[7] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, "Pre-trained models for natural language processing: A survey", Science China Technological Sciences, vol. 63, no. 10, pp. 1872-1897. 2020.

[8] Z. Huang, Z. Fang, "An Entity-Level Sentiment Analysis of Financial Text Based on Pre-Trained Language Model", 2020 IEEE 18th International Conference on Industrial Informatics (INDIN), Vol. 1, 2020, pp. 391-396.

[9] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, "Deep Contextualized Word Representations", Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018.

[10] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, "Improving language understanding by generative pre-training", 2018.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805, 2018.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need", Advances in neural information processing systems, vol. 30, 2017.

[13] M. Zaib, Q.Z. Sheng, W. Emma Zhang, "A short survey of pre-trained language models for conversational ai-a new age in nlp", Proceedings of the Australasian Computer Science Week Multiconference, 2020, pp. 1-4.

[14] A. Gillioz, J. Casas, E. Mugellini, O. Abou Khaled, "Overview of the Transformer-based Models for NLP Tasks", 2020 15th Conference on Computer Science and Information Systems (FedCSIS), IEEE, 2020, pp. 179-183.

[15] M. Chen, F. Du, G. Lan, V.S. Lobanov, "Using Pre-trained Transformer Deep Learning Models to Identify Named Entities and Syntactic Relations for Clinical Protocol Analysis", AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1), 2020, pp. 1-8.

[16] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, "Google's neural machine translation system: Bridging the gap between human and machine translation", arXiv preprint arXiv:1609.08144, 2016.

[17] C. Rao, V.N. Gudivada, Computational analysis and understanding of natural languages: principles, methods and applications, Elsevier, 2018.

[18] Papers with Code, "Named Entity Recognition on CoNLL 2003 (English)", https://paperswithcode.com/sota/named-entity-recognition-ner-on-conll-2003 (accessed at February 10, 2022).

[19] H. Cho, H. Lee, "Biomedical named entity recognition using deep neural networks with contextual information", BMC bioinformatics, vol. 20, no. 1, pp. 1-11. 2019.

[20] P. Lison, A. Hubin, J. Barnes, S. Touileb, "Named entity recognition without labelled data: A weak supervision approach", arXiv preprint arXiv:2004.14723, 2020.

[21] S. Moon, G. Lee, S. Chi, H. Oh, "Automated Construction Specification Review with Named Entity Recognition Using Natural Language Processing", Journal of Construction Engineering and Management, vol. 147, no. 1, pp. 04020147. 2021.

[22] K. Liu, N. El-Gohary, "Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports", Automation in Construction, vol. 81, pp. 313-327. 2017.

[23] R. Li, T. Mo, J. Yang, D. Li, S. Jiang, D. Wang, "Bridge inspection named entity recognition via BERT and lexicon augmented machine reading comprehension neural model", Advanced Engineering Informatics, vol. 50, pp. 101416. 2021.

[24] S. Moon, S. Chung, S. Chi, "Bridge Damage Recognition from Inspection Reports Using NER Based on Recurrent Neural Network with Active Learning", Journal of Performance of Constructed Facilities, vol. 34, no. 6, pp. 04020119. 2020.

[25] J. Zhang, M. El-Gohary Nora, "Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking", Journal of Computing in Civil Engineering, vol. 30, no. 2, pp. 04015014. 2016.

[26] R. Zhang, N. El-Gohary, "A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking", Automation in Construction, vol. 132, pp. 103834. 2021.

[27] S. Baek, W. Jung, S.H. Han, "A critical review of text-based research in construction: Data source, analysis method, and implications", Automation in Construction, vol. 132, pp. 103915. 2021.

[28] S. Moon, Y. Shin, B.G. Hwang, S. Chi, "Document Management System Using Text Mining for Information Acquisition of International Construction", KSCE Journal of Civil Engineering, vol. 22, no. 12, pp. 4791-4798. 2018.

[29] ENR, Top 250 International Contractors, Engineering News-Record, 2012-2021.

[30] M. Konkol, M. Konopík, "Segment representations in named entity recognition", International Conference on Text, Speech, and Dialogue, Springer, 2015, pp. 61-70.

[31] D.O.F. do Amaral, M. Buffet, R. Vieira, "Comparative analysis between notations to classify named entities using conditional random fields", Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology, 2015, pp. 27-31.

[32] M. Chen, X. Luo, H. Shen, Z. Huang, Q. Peng, "A Novel Named Entity Recognition Scheme for Steel E-Commerce Platforms Using a Lite BERT", CMES-Computer Modeling in Egineering & Science, vol. 129, no. 1, pp. 47-63. 2021.

[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, "Huggingface's transformers: State-of-the-art natural language processing", arXiv preprint arXiv:1910.03771, 2019.