Joint Triple Extraction for Construction Regulatory Documents using Graph Neural Networks

Sherief Ali
Chair of Computing in Engineering, Ruhr-Universität Bochum, Universitätsstrasse 150
Gebäude IC 6-63, 44801 Bochum, Germany
E-Mail: sherief.ali@ruhr-uni-bochum.de

Abstract: The construction industry depends on regulatory documents, to guarantee the consistency and applicability of their operations, which is crucial for maintaining standards and protocols. Information extraction is a time-consuming and labor-intensive process, which aims to extract the relevant data from documents for definite operation. Extracting information and assigning it to a particular rule requires understanding the semantics of the sentences, essential for ensuring code compliance. One approach to address this concern is to parse the unstructured text into triples by extracting entities and their relation. This process enables the identification of subjects, objects, and relationships that provide a comprehensive understanding of the sentence. This paper proposes a joint entity and relation extraction model based on a graph neural network. The model is trained on construction regulatory documents sourced from Eurocode for design. The dataset undergoes preprocessing and labeling in the form of triples to assess the model's performance. The results demonstrate that the model can effectively predict the sentence triples. The model's prediction of sentence triples indicates the ability to capture complex semantic relationships within textual data, which can be transformed into a comprehensive knowledge representation.

Keywords: Information extraction, Relation extraction, Graph neural networks, Joint entity recognition, Code Compliance

Erschienen in Tagungsband 35. Forum Bauinformatik 2024, Hamburg, Deutschland, DOI: 10.15480/882.13507
© 2024 Das Copyright für diesen Beitrag liegt bei den Autoren. Verwendung erlaubt unter Creative Commons Lizenz Namensnennung 4.0 International.

1 Introduction

The Architecture, Engineering, and Construction (AEC) industry is characterized by continually pursuing innovation across various fronts. Among the evolving topics nowadays is code compliance. Regulatory documents play a role, providing the framework within which construction projects operate, from start to end. Traditionally, rules and regulations have been written and interpreted solely by humans. However, this human-centric approach has occasionally led to instances of

incompleteness, where specific conditions were left unaddressed, or contradictions emerged within regulatory frameworks [1]. The process of code compliance needs multiple iterations until all obligations are inspected, which requires a significant amount of money, time, and labor [2]. Significant efforts are being made to interpret the text through various approaches, including Joint Entity and Relation Extraction (JERE), deep learning techniques, and rule-based systems [3]. JERE involves identifying and extracting subject-predicate-object triples from textual data, providing a structured representation of information [4]. This structured representation has served as a foundation for semantic knowledge data representation, enabling a more comprehensive understanding and analysis of the data. Researchers frequently face challenges in accessing adequate construction regulatory documents and comprehensive datasets, encountering difficulties due to the limited availability of documents for thorough analysis and experimentation. Consequently, in some cases, researchers have utilized synthetic data to support their applications, as demonstrated by Bloch et al., by generating a dataset to facilitate the implementation of their methodology [5].

This study presents an approach for transforming unstructured text within construction regulatory documents into subject-relation-object triples. This transformation is achieved through the training of a Graph Neural Network (GNN), leveraging pre-trained BERT (Bidirectional Encode Representations from Transformers) embeddings for node representation. Our study aims to develop a solution for automating code compliance by deeply understanding the semantics of regulatory text. To the best of our knowledge, this study represents the first application of GNNs for the joint entity and relation extraction task on manually labeled construction regulatory documents.

2 Related studies

The interpretation of regulatory documents for automatic code compliance can be approached using three main methodologies: rule-based methods, deep learning approach, and JERE [3, 6]. Rule-based methods rely on predefined rules and logic to interpret and enforce compliance, which can be efficient but often lack flexibility and are costly to maintain. However, machine learning approaches utilize algorithms to learn from data, enabling more adaptive and scalable compliance checks. These methodologies help streamline the compliance process and reduce manual intervention [7].

Various initiatives are being undertaken to leverage artificial intelligence in enhancing the interpretation of regulations. For instance, Schönfleder and König [8] have presented a study to extract relevant terms from German regulatory documents utilizing a supervised deep learning transformer model BERT. BERT is a deep learning model designed to understand the context of words in a text by analyzing them in both directions [9]. In their study, the authors [8] have utilized a pre-trained BERT model for transfer learning, thereby eliminating the need for a large dataset. Another approach proposed by Zhong et al. [10], is for automatically extracting construction procedural constraints using a hybrid deep neural network. Their method has integrated bidirectional long short-term memory (Bi-LSTM) and conditional random field (CRF), achieving efficient

recognition of named entities and their relationships. The LSTM, a type of recurrent neural network (RNN), is well-known for its ability to effectively capture long-range dependencies in sequential data, and it has shown effective performance in text-mining tasks [11]. CRF is a statistical modeling method used in machine learning for structured prediction tasks and is popular, particularly in sequence labeling tasks, such as named entity recognition and part-of-speech tagging [12].

Several studies have employed a hybrid approach, combining multiple techniques, to maximize performance in information extraction tasks. Thus, these studies have combined JERE with deep learning, to privilege the performance of both techniques for a comprehensive understanding of natural language. Fu et al. [13] have proposed an end-to-end relation extraction model for joint entity and relation extraction. The GraphRel model leverages graph convolutional networks (GCNs) to jointly extract named entities and their relationships from unstructured text, providing a comprehensive understanding of the text. GCNs are a type of neural network designed to operate on graph-structured data, capable of encoding both graph structure and node features. In GCNs, each node in the graph represents an entity, and the edge between nodes represents relationships [14]. In a parallel endeavor, Zhao et al. [15] have proposed an approach called Representation Iterative Fusion based on Heterogeneous GNNs for Relation Extraction (RIFRE), to extract entities and relations as relational triples. The GNNs are designed to operate directly on graph-structured data, to model and process relationships between entities represented as nodes in a graph [16, 17]. In their study, the authors [15] have leveraged a graph-based representation, considering relations and words as nodes and iteratively fusing semantic information.

3 Methodology

In this section, the detailed methodology outlines the steps and processes involved in structuring our model. Additionally, a comprehensive explanation of how the joint entity and relation extraction model is designed. To test this methodology, an experiment is applied in Section **4** to evaluate the results and assess the performance of the approach.

3.1 Data preprocessing

Regulatory documents comprise various structures of information, and identifying the informative sentences is crucial for our joint entity and relation extraction model. Data preprocessing begins with identifying the document level of digital standards [18]. The documents need a comprehensive reading to capture the full context and content. The next step involves extracting sentences from the documents, which is crucial for maintaining the coherence and structure necessary for accurate entity and relation detection. Each sentence is then saved in a plain text (.txt) format. The extracted sentences are then cleaned by removing hyphens, as well as checking for proper punctuation, such as full stops, to ensure the text is clear and relevant for analysis.

3.2 Triplets tagging

The objective is to label the sentences for utilization in our joint model, aiming to extract triplets based on the subject-relation-object structure. To achieve this, the sentences undergo tagging using Label Studio [19], a tool tailored for efficient data labeling. Multiple relations could be in different sentences, therefore a thorough examination of the data extracted is needed to create a guideline for the relations. Given that multiple relations may exist across different sentences, a thorough examination of the extracted data is essential to establish a guideline for identifying these relations. Furthermore, both the subject and object within a sentence may consist of multiple words, which will be tested to assess the efficiency of our model in accurately determining these complex subject-object relationships.

3.3 Vector representation

To utilize the tagged sentences in the model, the data need to be transformed into numerical representations. This involves creating vector representations for each entity and relation involved. In our study, we adopt a two-node representation approach as conducted before [15], the first node represents words, utilizing embeddings from the pre-trained BERT model [20], particularly the token embeddings in the last hidden layer of BERT are utilized as word nodes. The second node represents relations, with pre-defined nodes assigned for each relation.

3.4 Joint entity and relation extraction

3.4.1 Graph Neural Network (GNN)

GNNs are instrumental in effectively capturing and modeling the complex dependencies between words and their relationships [16, 17]. The GNN operates on a heterogeneous graph where nodes represent both words and relationships, and edges denote the connections between them based on syntactic and semantic dependencies. The process begins with the initialization of node representations, where each word and relation node is assigned an initial embedding based on its features. Similar to the procedure applied in [15], the GNN then updates these node representations through the message-passing mechanism. In this mechanism, each node aggregates information from its neighbors to update its representation. The process enhances the model's ability to extract information and dependencies, enabling it to accurately extract triplets.

3.4.2 Joint triple extraction

To extract the information in the form of triplets, the procedure to apply is similar to studies [15, 21], which have the subject tagger and object tagger. The subject tagger detects all possible subjects in the word nodes. Specifically, two binary classifiers are employed to identify the beginning and end positions of the subject within the word nodes. The object tagger is utilized to complete the extraction of triples. This combined extraction method uses enhanced node representations to capture essential contextual and relational information, which is vital for accurate entity and relation extraction.

4 Experiment

4.1 Data preparation

To conduct our experiment on the proposed methodology, the focus is to experiment on construction regulatory documents because this area needs further studies. Therefore, the selected dataset is the Eurocode for the design [22]. The EN Eurocodes are a series of 10 European Standards, EN 1990 - EN 1999, but for our experiment, the choice is to initially test our methodology on the documents detailed in **Table 1**, to assess their effectiveness and suitability. The Eurocode documents exhibit a diverse text structure, including unstructured text written in normal paragraphs, as well as structured text presented in tables and figures. Our extraction process focuses on both the unstructured and structured text, aiming to effectively capture the relevant information from these different formats.

Table 1: Experimental Dataset

Standard	Eurocode name
EN 1990	Basis of structural design
EN 1991	Actions on structures
EN 1991	Densities, self-weight, imposed loads for buildings

To extract the text from the dataset, a comprehensive reading and understanding are conducted to have a clear insight. Consequently, the purpose is to identify and extract informative sentences from the code and gather them into a structured text file. The labeling process, as detailed in Section **3.2** proceeds by identifying the subject, relation, and object. Within our dataset, three distinct relations were extracted. The total number of tagged sentences is 300, encompassing 600 entities. As shown in **Figure 1**, sentences may have subjects or objects comprising multiple words. This aspect is integral to our model testing phase, to evaluate the model's proficiency in accurately extracting triplets.

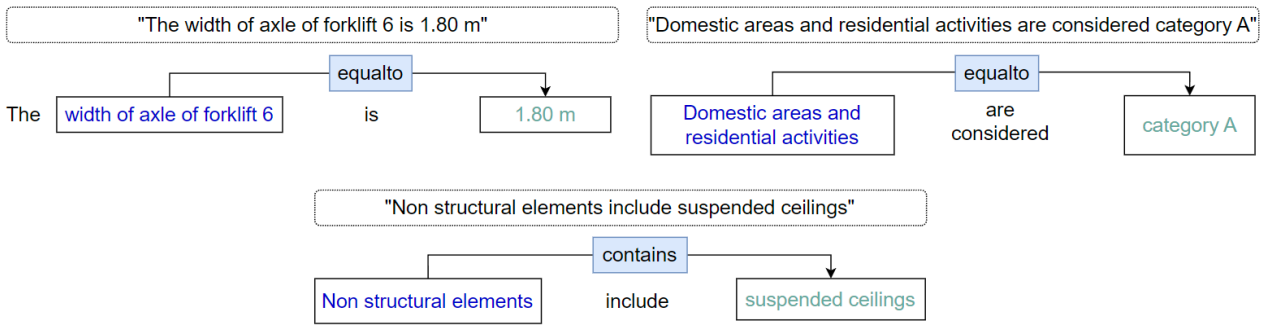

Figure 1: Dataset labeling

4.2 Model implementation

The model is trained by implementing the pre-trained BERT model (base-cased) as a foundation and adding the additional layers and components from vector representation, graph layer, and joint extraction on top of the pre-trained transformer. The dataset has been divided into three distinct subsets to optimize model performance: 70% of the data is allocated for training, 20% is reserved for validation, and the remaining 10% is for testing. Various hyperparameters are fine-tuned to optimize the model's performance. A regularizer with a value of 0.01 is added to address instability during training, improving the model's stability. Early stopping is implemented to prevent overfitting and a learning rate scheduler with patience of 10 epochs is used to adjust the learning rate, ultimately stabilizing at 0.1. The SGD optimizer is chosen for training the model, and the batch size is optimized to 16. Given the presence of long sentences with crucial information in our dataset, the maximum number of words in a sentence was set to 120. The model is trained for 60 epochs to ensure thorough learning.

4.3 Results and Discussion

Our model's performance is evaluated using precision, recall, and F1-score metrics, which are crucial for assessing the accuracy and robustness of relational triple extraction. The model demonstrates outstanding results as outlined in Table 2, achieving a precision of 86.4%, a recall of 88.2%, and an F1-score of 87.2%. The model results show that it accurately predicts the subject, relation, and object within sentences, correctly identifying the start and end points of both the subject and object and the relation between them. The F1 score varies during training, with the best score achieved at 91.3%. The results of this study cannot be directly benchmarked against several research due to the use of a newly developed dataset. This is a promising outcome, suggesting that our model is well-suited for information extraction tasks in the form of triples. This capability is particularly beneficial for the construction industry, which requires automated extraction of information from regulations to support digitization efforts.

Table 2: Evaluation of the model

Evaluation metrics	Performance
Precision	86.4
Recall	88.2
F1-Score	87.2

5 Conclusion and future work

The study has presented a model for joint entity and relation extraction utilizing GNN. The proposed model is trained on our own labeled dataset of Eurocode for design, demonstrating its capability to identify and predict sentence triples. Our findings has shown that the model not only predicts sentence triples accurately but also captures complex semantic relationships within the textual data. However, there are limitations to our study, particularly the size of the dataset. Increasing the amount of labeled data for training is essential for enhancing the model's accuracy and robustness. Regarding future work, we plan to label more data to expand the dataset, which we expect will improve the model's performance. Additionally, we aim to represent the extracted triplets in a comprehensive knowledge representation framework. This will further facilitate the transformation of textual information into structured, meaningful knowledge, advancing the utility of our approach in the construction regulatory domain and beyond.

Acknowledgments

The author expresses his sincere appreciation to the Chair of Computing in Engineering for generously funding this research paper. Their support has played a pivotal role in facilitating the study and contributing to advancements in the field.

References

[1] C. Eastman, J. Lee, Y. Jeong, and J. Lee, “Automatic rule-based checking of building designs,” Automation in Construction, vol. 18, no. 8, pp. 1011–1033, 2009, doi: 10.1016/j.autcon.2009.07.002.

[2] Borrmann and preidel, “BIM-based Code Compliance Checking,” 2018.

[3] Daniel Jurafsky, James H. Martin, “Speech and Language Processing,” 2023.

[4] Anthony Fader, Stephen Soderland, and Oren Etzioni, “Identifying Relations for Open Information Extraction,” 2011.

[5] Tanya Bloch, Andre Borrmann, Pieter Pauwels, “An alternative approach to Automated Code Checking - Application of Graph Neural Networks trained on synthetic data for an accessibility check case study,” 2023.

[6] S. Fuchs and R. Amor, “Natural Language Processing for Building Code Interpretation: A Systematic Literature Review,” *Proceedings of the 38th International Conference of CIB W78*, 2021.

[7] Nawari, “A Generalized Adaptive Framework (GAF) for Automating Code Compliance Checking,” *Buildings*, vol. 9, no. 4, p. 86, 2019, doi: 10.3390/buildings9040086.

[8] P. Schönfelder and M. König, “Deep Learning-Based Entity Recognition in Construction Regulatory Documents,” *Proceedings 38th International Symposium on Automation and Robotics in Construction (ISARC 2021)*, 2021, doi: 10.22260/ISARC2021/0054.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2019, doi: 10.48550/arXiv.1810.04805.

[10] B. Zhong *et al.,* “Deep learning-based extraction of construction procedural constraints from construction regulations,” *Advanced Engineering Informatics*, vol. 43, p. 101003, 2020, doi: 10.1016/j.aei.2019.101003.

[11] J. S. Sepp Hochreiter, “Long Short-Term Memory,” 1997, doi: 10.1162/neco.1997.9.8.1735.

[12] C. Sutton and A. McCallum, “An Introduction to Conditional Random Fields,”

[13] Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma, “GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction,” 2019, doi: 10.18653/v1/P19-1136.

[14] T. N. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks,” 2017.

[15] K. Zhao, H. Xu, Y. Cheng, X. Li, and K. Gao, “Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction,” *Knowledge-Based Systems*, vol. 219, p. 106888, 2021, doi: 10.1016/j.knosys.2021.106888.

[16] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph Attention Networks,”

[17] V. Garcia and J. Bruna, “Few-Shot Learning with Graph Neural Networks,”

[18] Damian A. Czarny, Gilles Bülow, Michael Noll, Dietmar Lochner, Damian A. Czarny, Jens Gayko, Jui Nahid Pervin, Christian Diedrich, Johannes Diemer, “Scenarios for Digitizing Standardization and Standards,” 2021.

[19] Tkachenko, Maxim, Malyuk, Mikhail, Holmanyuk, Andrey, & Liubimov, Nikolai, “Label Studio: Data labeling software,” 2020. [Online]. Available: https://github.com/heartexlabs/label-studio

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,”

[21] F. Li, Y. Song, and Y. Shan, “Joint Extraction of Multiple Relations and Entities from Building Code Clauses,” *Applied Sciences*, vol. 10, no. 20, p. 7103, 2020, doi: 10.3390/app10207103.

[22] British Standards Institution, “Eurocode: EN 1990 - EN 1999,”