Standard-Driven Chinese Knowledge Extraction in Highway Domains Using Machine Learning and NLP Approach

Ching W. Loh[1]; Xiao-Bing Zhang[2]; Jia-Rui Lin[3]; and Zhen-Zhong Hu[4]

[1]Tsinghua Shenzhen International Graduate School, University Town of Shenzhen, PR China. Email: luo-zh22@mails.tsinghua.edu.cn
[2]Tsinghua Shenzhen International Graduate School, University Town of Shenzhen, PR China. Email: zxb21@mails.tsinghua.edu.cn
[3]Associate Professor, Dept. of Civil Engineering, Tsinghua Univ., Beijing, PR China. Email: lin611@tsinghua.edu.cn
[4]Associate Professor, Tsinghua Shenzhen International Graduate School, University Town of Shenzhen, PR China. Email: huzhenzhong@tsinghua.edu.cn

ABSTRACT

Advancements in informatics have led to numerous data-driven strategies for improving transportation efficiency. However, data exchange between transportation systems is often hindered by complexity, ambiguous definitions, and varied data sets. To tackle these issues, this article introduces a method for creating an automatic knowledge extraction model for highway standards. Machine learning models like BiLSTM-CRF, TextCNN-BiLSTM-CRF, BERT, and BERT-CRF are utilized on small training sets for tasks such as Named Entity Recognition using ISO 12006-3 as upper-level ontology and relationship classification. Additionally, post-prediction and manual corrections refine training data sets for iterative learning. The best results are formatted into graphs and saved as OWL ontologies. This approach yielded 158 graphs from Chinese standards, linked via ISO 12006-3 referenced classes, and the outcome links highway domain concepts, enhancing data management and project collaboration. This method shows promise for better data management and interdisciplinary collaboration in highway projects, furthering data-driven progress in the field.

INTRODUCTION

The digital transformation of urban infrastructure has been championed as a pivotal strategy by prestigious institutions and agencies such as the National Highways UK (2021), American Road & Transportation Builders Association (2023). In detail, information management has been identified as a cornerstone within the realms of digitization, digitalization, and digital transformation in the highway sector. As Bolton, Butler et al. (2018) indicated, technologies that rely heavily on data and models, such as Digital Twin and Building Information Modelling (BIM) necessitate robust information management as an enabler to facilitate smart infrastructure and green economic performance (Ma and Lin 2023). While the highway sector, as a component of urban infrastructure, may also stand to benefit from these technologies. Nevertheless, challenges, that includes data fragmentation, weak data governance, large knowledge and experience gaps, have hindered the pace of digital transforming across different construction projects (Zhang, Ye et al. 2023).

Given the prevalent challenge of fragmented information within unstructured or semi-structured sources, machine learning-based knowledge extraction has emerged as an efficient solution. Data can be stored in structured, semi-structured and unstructured formats, nevertheless, however, the more effectively the data is organized, the more easily the information can be processed (Hodapp and Hanelt 2022), shared (Liepert, Stary et al. 2023), and reused (Petrasch and Petrasch 2022). Furthermore, knowledge extraction is one of the way to compile scattered data into integrated knowledge, and as a result, knowledge graphs may enhance data-driven applications in the construction industry, such as BIM (Hu, Leng et al. 2022), smart buildings (Pauwels, Costin and Rasmussen 2022), digital twins (Ramonell, Chacón and Posada 2023), etc. Specifically, knowledge graphs are constructed from triples (consisting of a subject, predicate, and object), which are a fundamental component of graph data structures, facilitating the description of relationships between data elements using predicates. To construct the triplets from unstructured or semi-structured data, a variety of knowledge extraction techniques have been proposed, including rule-based (Wu, Lin et al. 2022), pattern-based extraction (Liu, Hua and Zhou 2021) and machine learning (Zhang, Xu et al. 2022) have been proposed. While rule-based knowledge extraction offers high precision, it requires a level of expertise in linguistics and sentence structure to create rules that effectively capture concepts (Waltl, Bonczek and Matthes 2018). In relation to that, machine learning, including deep learning, provides automation for knowledge extraction, with a heavy reliance on training sets.

Furthermore, manual development of domain ontologies has enhanced knowledge management in the construction sector. In the computer science sector, ontology has been used as a data representation of data, information, and knowledge that enforce symbolic and explainable artificial intelligence (de Sousa Ribeiro and Leite 2021). As ontology can be stored digitally, its high degree of interoperability and shareability bolsters data distribution, bridges knowledge gaps, and facilitates reasoning across various contexts within the construction sector (Chen, Chen and Cheng 2018, de Aguiar Corrêa, Jehel et al. 2021). Different to knowledge graphs, ontology focuses on explainable conceptual classes (T-box), assertion individuals (A-box) and relationship by defining clear axioms of the concepts for different scope of work and purposes. To enable the data sharing aspect in the construction sector, domain ontologies have been constructed manually by experts and scholars in the construction industry. For instance, El-Diraby (2013) has introduced Domain Ontology for Construction Knowledge (DOCK) as a high level conceptual domain ontologies which include general entities, such as processes, actors, systems and products in the construction industry. In the infrastructure sector, IFCInfra4OM ontology (Ait-Lamallam, Sebari et al. 2021) have extend concepts from the Industry Foundation Classes (IFC). Road, ground, and utility concepts are covered by Assessing The Underworld (ATU) ontology (Du, Wei et al. 2023), in addition, HiOnto (El-Diraby and Kashif 2005) have provided concepts for the highway industry. High-effort works have demonstrated the existence of various ontologies, each with their unique scope. However, expertise is required to manage the complexity before these ontologies can be effectively implemented.

To cumulates shareable knowledge in the highway industry, this study introduces a Chinese knowledge extraction framework that combines manual expert verification with machine learning for automation. Historically, datasets for machine learning-based knowledge extraction have been limited, with varying classes of interest across domains. To enhance automation in highway knowledge extraction, it is crucial to curate datasets following well-established standards in the construction industry. Emphasizing the integration of human intelligence with machine learning, this research aims to create an adaptable and continually improved knowledge extraction framework, enhancing accuracy and reliability. The framework is delineated into three major stages: (1) collecting relevant Chinese standards for the highway sector as the research dataset, (2) identifying entities and properties in the standards aligned with construction industry concepts, and (3) constructing triples by categorizing relationships between identified entities.

METHODOLOGY

Overall, knowledge extraction task has been bifurcated into two minor tasks: Named Entity Recognition (NER) and relationship classification (RC). For NER, the following models have been employed. (1) BiSLTM-CRF (Dai, Xiao et al. 2019) excels in sequence context via bidirectional Long Short-Term Memory networks (BiLSTM) and Conditional Random Fields (CRF). (2) Text-CNN-CRF (Long, Yuan et al. 2018, Zhang and He 2022) uses convolutional neural networks (CNN)s for local feature detection, enhanced by CRFs. (3) BERT (Devlin, Chang et al. 2019) and (4) BERT-CRF (Hu, Zhang et al. 2022), with their pre-trained language understanding, lead in NLP benchmarks and provide a broad contextual view. This selection balances individual model strengths with our dataset's needs, aiming to optimize performance for our specific NER task. For the subsequent task of RC, the BERT sequence classification model has been fine-tuned with two different sets of categories, which involve the type and direction of the relationship in a triplet. With this dual-pronged approach, the subject, predicate, and object of a triplet can be constructed using the outcome of the two minor tasks. Next, manual checking and feedback from highway domain experts have been involved to ensure the quality of the triplets before it being added to the dataset for later round of training (as shown in Figure 1).

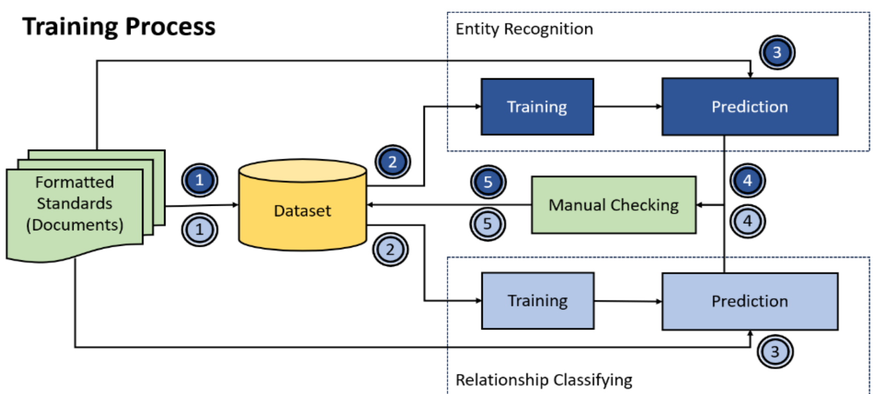

Figure 1: Training process of the knowledge extraction framework: (1) Manual tagging datasets (2) Model training and evaluation (3) Extended dataset for new prediction (4) Manually checking by domain experts (5) Reclaim the finalized triplets for next training task.

In the initial phase of this study, 158 data sources for knowledge extraction have been chosen from various levels of standards. Since consensus is crucial in knowledge management, clear definitions are necessary to prevent misunderstandings and promote effective implementation (Newell 2015). To ensure mutual agreement and clear definitions, standards that have undergone extensive peer review and discussion have been selected as the sources for extracting concepts related to knowledge in this research. Specifically, Chinese standards are categorized into five levels: national standards, sector standards, local standards, association standards and enterprise standards (China Communications Standards Association 2018), with the former considered more significant than the latter. Referring to these levels, the standards with pertinent to the highway sector have been selected for this research, and the number of each standard categories are listed in Table 1.

Table 1: The number of different categories of standards selected for this research.

Chinese National Standards	GB, GB/T	12
Transportation Industry Standards	JT, JT/T	61
Technical Standards of Highway Engineering	JTG, JTG/T	79
Urban Construction Industry Engineering-related Standards	CJJ, CJJ/T	6

In the second stage, datasets have been prepared for the subsequent tasks of NER and RC. To address the absence of domain-specific datasets tailored for NER training and RC within the highway industry, manually crafted datasets are being prepared by domain experts with a background in engineering. This approach ensures a robust foundation for our machine learning and natural language processing (NLP) methodologies. In detail, textualized data has been divided into sections and sentences after collecting standards mentioned above. During separating, data cleaning and preprocessing steps, such as noise removing, character correcting and deduplicating, have been applied to homogenize the data. The preparation of these datasets increases the reliability of the training sets and may improve the performance of the NER and RC models, as well as for ensuring that the extracted knowledge meets the specific requirements and standards of the targeted domain.

Moreover, classes for the NER task have been selected and extended from the ISO 12006-3 standard and the IFC schema. To establish a consensus on domain-specific classes applicable to the NER process, recourse has been taken to the IFC schema and ISO 12006-3. Designed for concepts in BIM and data dictionaries, ISO 12006-3 serves as an upper-level ontology in this research. Nevertheless, changes have been implemented in the 2022 version of ISO 12006-3, resulting in a reduction in the number of constrained concepts when compared to its 2016 counterpart. Despite this reduction, the identification and selection of relevant concepts from both versions for tagging in the NER process have been carried out by this research. This exploration of the standards ensures the incorporation of comprehensive and pertinent concepts in the knowledge extraction approach for the highway domain. Additionally, concepts are being extended to encompass the classes that are of interest in standards and not explicitly covered by IFC and ISO 12006-3. As a result, a comprehensive framework for describing general concepts relevant to the highway industry has been established. As Figure 2 shows, 23 classes under have been selected for the NER tagging task.

Furthermore, to capture the beginning and the end of the character of the entity in the text, the BIO tagging system is being utilized. This system, short for Beginning, Inside, and Outside, is a widely employed technique in natural language processing (NLP) for annotating and labeling entities within a given text. In detail, a Chinese character or an English word in the sentence will be assigned a label indicating whether it is the beginning of an entity, inside an entity, or outside entity.

By applying this tagging to the text extracted from standards, relevant concepts can be easily identified and extracted with the classes associated to the labels. Subsequently, each class has an associated label. Furthermore, in contrast to the traditional NER approach, the primary factor in deciding whether it should be tagged as a class is the semantics of the entity in the sentence. For example, the entities “soil” and “stone” in the sentence shown in Figure 3 can be tagged as an “ifc:Product” that is cleared. However, the term “cleared” as an indicator of “xtdActivity” should explicitly specify that its targets are the “stone” and "soil" with a status of being “located on the road”. With that in mind, it is better to tag the entire concept with one entity if its components are connected.

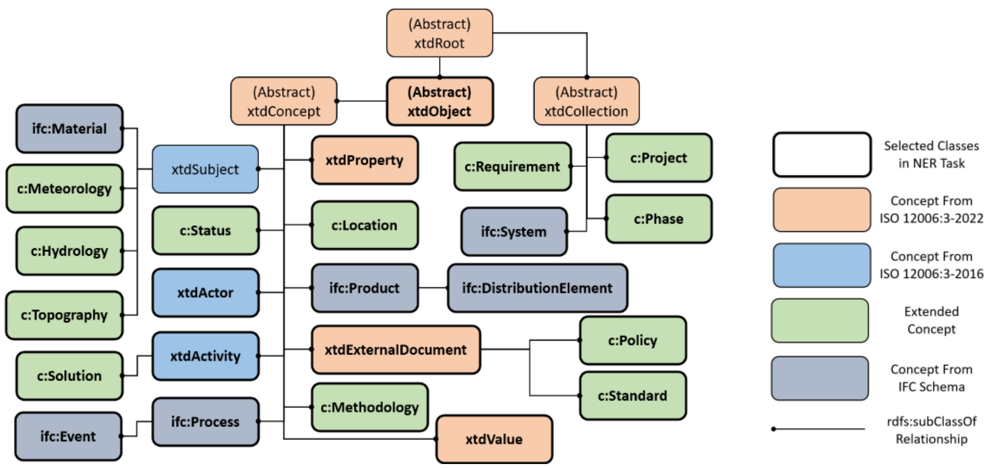

Figure 2: Relationship between classes from ISO 12006-3:2016 (ISO 2016), ISO 12006-3:2022 (ISO 2022), IFC schema (ISO 2018) and extended classes.

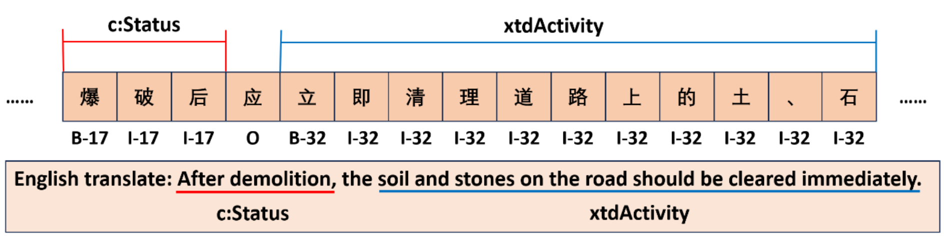

Figure 3: An example of BIO tagging in a sentence for NER task.

On the contrary, datasets for the RC task have been compiled by concatenating paired entities and their corresponding sentences, integrating a classification for relationship with direction. Hence, in preparation for the RC task dataset, 35 relationships, delineating the activities specified in the standards, have been introduced and cataloged in Table 2.

Specifically, the “require” series of relationships elucidates the mandatory elements essential for an activity, while the “has” signifies a containment relationship. Additionally, the “isReferencedTo” series characterizes dependency relationships, and “isRelatedTo” records an ambiguous but perceptible relationship.

Moreover, a singular relationship category has been assigned to entity pairs as exemplified in Figure 3, “requireStatus” has been assigned from the activity “cleared” to the status “demolished”. In parallel, the dataset for fine-tuning BERT for RC is being systematically constructed, as illustrated in Figure 4, with the assignment of a relationship category denoted as “requireStatus-backwards.” Hence, this yields 71 relationship categories, including a “No-relationship” category, associated with 35 original categories, each associated with a direction.

Table 2: Relationship categories (without direction)

has	hasTestRequirement	isReferencedToStandard	isRelatedToTask	requireMethodology
hasPart	hasTopography	isReferencedToTopic	prevent	requireProperty
hasStatus	isReferencedTo	isRelatedTo	require	requireStatus
hasSubTask	isReferencedToDocument	isRelatedToEntity	requireActor	requireTargetEntity
hasSystem	isReferencedToEntity	isRelatedToLocation	requireEntity	requireTargetProperty
hasTarget	isReferencedToPhase	isRelatedToProperty	requireEquipment	subClassOf
hasTask	isReferencedToRequirement	isRelatedToRequirement	requireLocation	sameAs

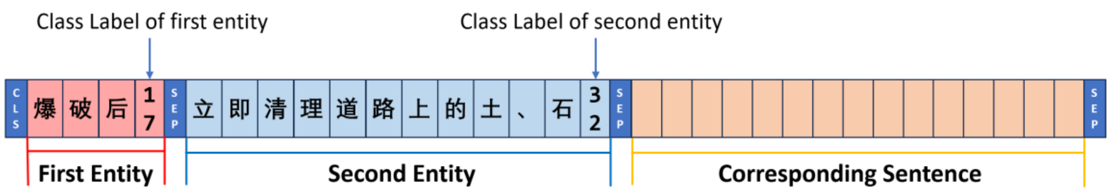

Figure 4: Concatenation of Entities (with the class label), and the corresponding sentence. (CLS and SEP are token required in BERT for sentence separation)

After setting up the datasets for both NER and RC, various models have undergone training, and their performance is being assessed using the F1 score. In the realm of NER, models such as BiSLTM-CRF, TextCNN-BiLSTM-CRF, BERT, and BERT-CRF have been fine-tuned to effectively discern and extract entities from textual data. In this research, bert-base-chinese (Devlin, Chang et al. 2019) has been utilized as the pretrained model of the BERT fine-tune exercise in NER. The NER task utilizes 30 epochs for training, in contrast, RC models, particularly those based on BERT, are trained for 20 epochs to understand associations between entities within the given context. The evaluation method involving the F1 score serves as a robust metric, considering both precision and recall. It provides a comprehensive assessment of the models' performance in correctly identifying entities and their relationships by considering the precision and the recall from the test dataset, which is 0.1 size of the training set. This iterative process of training and evaluation ensures the refinement and optimization of the models, ultimately enhancing their accuracy and efficacy in the knowledge extraction framework. The training specifics involve utilizing the best outcomes from five random shuffles of the dataset, with the optimal configuration chosen for processing the 158 standards. For RC, three random shuffles are conducted to achieve a robust and diverse training process.

RESULT AND DISCUSSION

In the results of NER, BERT-CRF emerged as the top-performing model among the four variants assessed and BERT model perform well in the RC task. The evaluation, conducted using the F1 score, showcased BERT-CRF's notable proficiency in accurately identifying and extracting entities within the dataset. The F1 score for BERT-CRF reached 87.54, outperforming

BiSLTM-CRF, TextCNN-BiLSTM-CRF, and BERT (as shown in Figure 5). This outcome underscores the efficacy of BERT-CRF in capturing patterns within the textual data, emphasizing its superiority in the NER task within the knowledge extraction framework. Moreover, the accuracy of each model demonstrated a consistent and gradual improvement as the size of the training set increased from 500 to 1800 (as shown in Figure 5). Moreover, the results of RC in three rounds steadily increase throughout the epochs, and the increase is also observed as the size of the training set grows. (shown in Figure 5). It is noteworthy to mention that the steady increase in accuracy may be attributed to the uniform distribution of the dataset, contributing to the models' ability to generalize effectively across varying instances and resulting in a systematically improved performance (as shown in Figure 6). All four models exhibited incremental accuracy gains as the training set size expanded, showcasing their adaptability and capacity to capture intricate patterns within larger datasets.

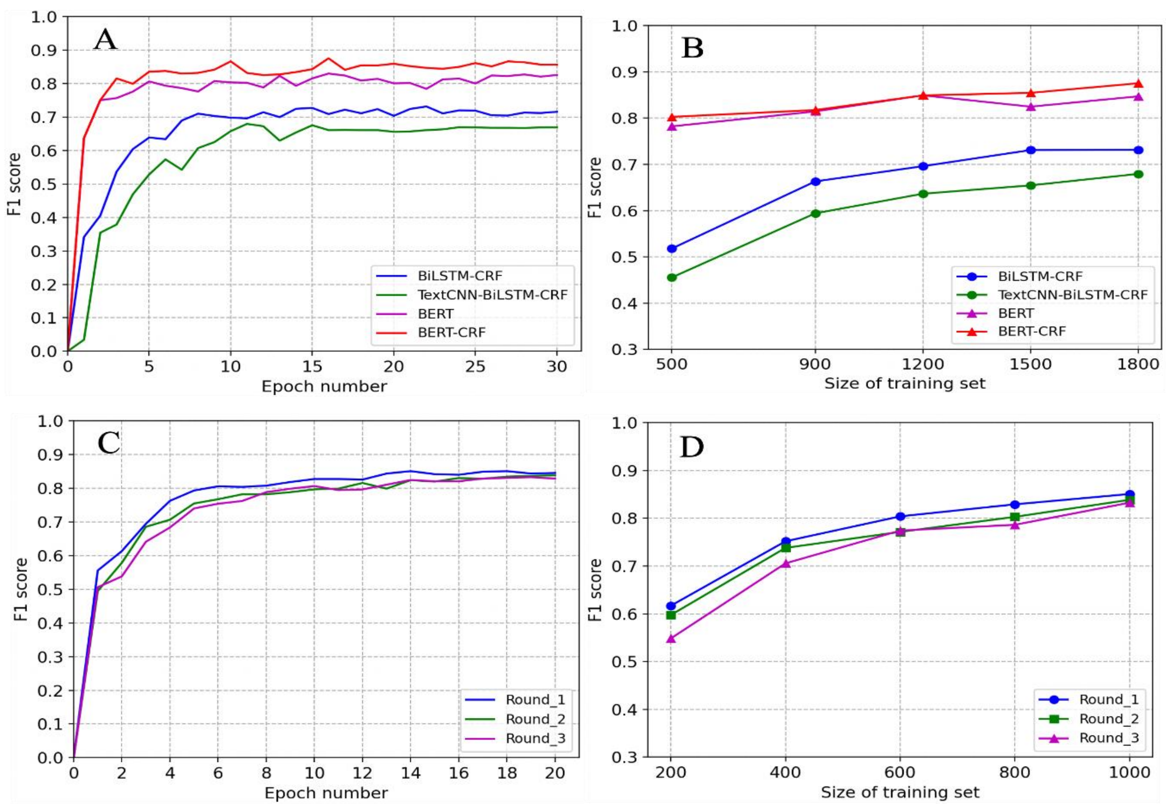

Figure 5: Evaluation of each model in NER and RC task according to epoch number and size of the training set. (A) Epoch-wise F1 Score in NER Task; (B) F1 Score with Varied Training Set Sizes in NER Task; (C) Epoch-wise F1 Score in RC Task; (D) F1 Score with Varied Training Set Sizes in RC Task

In the results of both NER and RC, a total of 125,692 entities and 183,582 relationships have been identified within the 158 standards. The class with the highest count is xtdActivity, totaling 12,107 entities. This is understandable as standards predominantly encompass activities relevant at various stages in the highway life cycle. Most of these activities are associated with relationships such as “hasSubTask”, denoting hierarchical connections between activities, and the “require” series, indicating constraints on each activity. The “isReferencedTo” series of relationships is also prevalent, linking the current standard to policies and external documentation. This summary encapsulates the graph generated from the highway sector standards. However, it's important to note that relationship categories may vary for different sets of textualized text beyond standards. This framework accommodates such variations with slight modifications to the relationship categories.

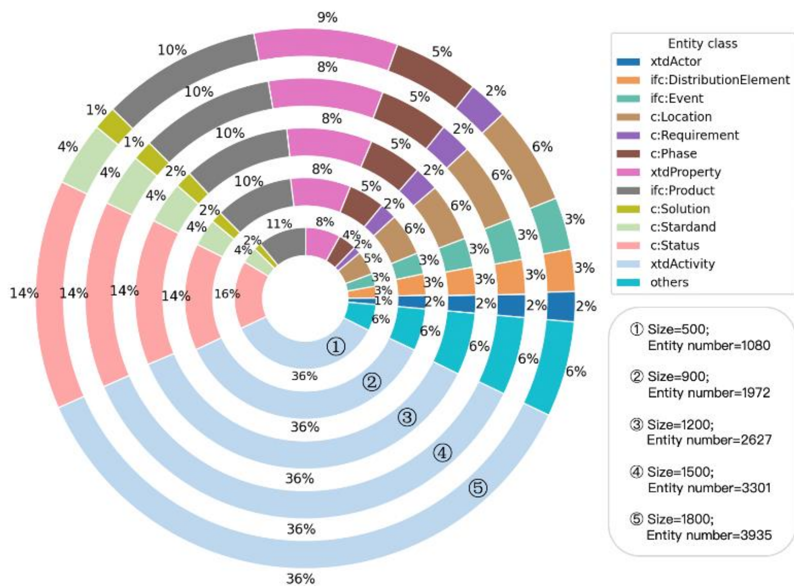

Figure 6: Distribution of each class in NER training set

As a limitation of this research, deep semantics from the statements haven’t been extracted. Since entities are extracted based on the sentence's meaning, some components are not well broken down due to the limitations of POS Tagging in NER. This is because BIO POS tagging only allows a single interpretation of the sentence. With more fragmented tagging, the relationships become more complex, and fragmented entities may not be the most effective way to record the relationship. This may be a direction of the future research regards to deep semantic extractions from the result produced by this research.

CONCLUSION

In conclusion, our study successfully employed machine learning, specifically BERT-CRF, for knowledge extraction in the highway sector. The integration of domain expertise and various models showcased superior performance in extracting entities and relationships from 158 standards. The NER task, evaluated with the F1 score, highlighted BERT-CRF's effectiveness. The analysis revealed 125,692 entities and 183,582 relationships, with xtdActivity dominating, reflecting the emphasis on highway life cycle stages. Key relationships such as “hasSubTask”, “require”, and “isReferencedTo” played pivotal roles in defining structures and references. The scalability of models, evidenced by accuracy growth with an expanded training set, underscores their adaptability. Despite variations in relationship categories in different contexts, our flexible framework accommodates such changes through slight modifications. This study contributes to advancing knowledge extraction in the highway sector, emphasizing the synergy of machine learning and domain expertise. The findings pave the way for further applications in complex domains, establishing a foundation for future research.

ACKNOWLEDGEMENT

This work was supported by the National Key Research and Development Program of China under Grant 2022YFC3801100.

REFERENCES

Ait-Lamallam, S., I. Sebari, R. Yaagoubi, and O. Doukari. (2021). “IFCInfra4OM: An ontology to integrate operation and maintenance information in highway information modelling.” *ISPRS International Journal of Geo-Information* 10(5): 305.

American Road & Transportation Builders Association. (2023). *Accelerating Digital Construction for Transportation Projects: Recommendations and Root Issues*, American Road & Transportation Builders Association.

Bolton, A., L. Butler, I. Dabson, M. Enzer, M. Evans, T. Fenemore, F. Harradence, E. Keaney, A. Kemp, and A. Luck. (2018). *The Gemini principles*.

Chen, W., K. Chen, and J. C. P. Cheng. (2018). “Towards an Ontology-based Approach for Information Interoperability Between BIM and Facility Management.” *Advanced Computing Strategies for Engineering*, Cham, Springer International Publishing.

China Communications Standards Association. (2018). “Standardization Law of the People’s Republic of China.” <http://oldmh.ccsa.org.cn/article_new/show_article.php?categories_id=428b9f25-50a5-cf06-3677-44b1c0410df5&article_id=fagui_2049b387-fbb9-701d-81c9-5ab70c510ffa>(Mar. 25, 2018).

Dai, D., X. Xiao, Y. Lyu, S. Dou, Q. She, and H. Wang. (2019). “Joint extraction of entities and overlapping relations using position-attentive sequence labeling.” *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. Honolulu, Hawaii, USA, AAAI Press: Article 773.

de Aguiar Corrêa, L., P. Jehel, F. F. Marchiori, and R. F. T. da Silva. (2021). “Ontology and interoperability in construction planning automation with 4D BIM: a systematic review.” *Simpósio Brasileiro de Gestão e Economia da Construção* 12: 1-8.

de Sousa Ribeiro, M., and J. Leite. (2021). “Aligning Artificial Neural Networks and Ontologies towards Explainable AI.” *Proceedings of the AAAI Conference on Artificial Intelligence* 35(6): 4932-4940.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (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* 1. Minneapolis, Minnesota, Association for Computational Linguistics.

Du, H., L. Wei, V. Dimitrova, D. Magee, B. Clarke, R. Collins, D. Entwisle, M. E. Torbaghan, G. Curioni, and R. Stirling. (2023). “City infrastructure ontologies.” *Computers, Environment and Urban Systems* 104: 101991.

El-Diraby, T. E. (2013). “Domain Ontology for Construction Knowledge.” *Journal of Construction Engineering and Management* 139(7): 768-784.
El-Diraby, T. E., and K. Kashif. (2005). “Distributed ontology architecture for knowledge management in highway construction.” *Journal of Construction Engineering and Management* 131(5): 591-603.
Hodapp, D., and A. Hanelt. (2022). “Interoperability in the era of digital innovation: An information systems research agenda.” *Journal of Information Technology* 37(4): 407-427.
Hu, S., H. Zhang, X. Hu, and J. Du. (2022). “Chinese Named Entity Recognition based on BERT-CRF Model.” *2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS)*, Zhuhai, China, 2022, pp. 105-108.
Hu, Z.-Z., S. Leng, J.-R. Lin, S.-W. Li, and Y.-Q. Xiao. (2022). “Knowledge extraction and discovery based on BIM: a critical review and future directions.” *Archives of Computational Methods in Engineering* 29(1): 335-356.
ISO. (2016). ISO 12006-3:2016 *Building construction-Organization of information about construction works - Part 3: Framework for object-oriented information*. Geneva, ISO.
ISO. (2018). ISO 16739-1:2018 *Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries Part 1: Data schema*. Switzerland, ISO.
ISO. (2022). ISO 12006-3:2022 *Building construction-Organization of information about construction works - Part 3: Framework for object-oriented information*. Geneva, ISO.
Liepert, C., C. Stary, A. Lamprecht, and D. Zügn. (2023). *Improving Interoperability in the Exchange of Digital Twin Data Within Engineering Processes*, Cham, Springer Nature Switzerland.
Liu, Y., W. Hua, and X. Zhou. (2021). “Temporal knowledge extraction from large-scale text corpus.” *World Wide Web* 24(1): 135-156.
Long, S., R. Yuan, L. Yi, and L. Xue. (2018). “A Method of Chinese Named Entity Recognition Based on CNN-BILSTM-CRF Model. ” *Data Science*, Singapore, Springer Singapore.
Ma, R., and B. Lin. (2023). “Digital infrastructure construction drives green economic transformation: evidence from Chinese cities.” *Humanities and Social Sciences Communications* 10(1): 460.
National Railway UK. (2021). “Digital Roads 2025 Roadmap: Digitally Enabled - Connected - Data Driven.”
Newell, S. (2015). “Managing Knowledge and Managing Knowledge Work: What we know and what the Future holds.” *Journal of Information Technology* 30(1): 1-17.
Pauwels, P., A. Costin, and M. H. Rasmussen. (2022). “Knowledge Graphs and Linked Data for the Built Environment.” *Industry 4.0 for the Built Environment: Methodologies, Technologies and Skills*. M. Bolpagni, R. Gavina and D. Ribeiro. Cham, Springer International Publishing: 157-183.
Petrasch, R. J., and R. R. Petrasch. (2022). “Data Integration and Interoperability: Towards a Model-Driven and Pattern-Oriented Approach.” *Modelling* 3(1): 105-126.
Ramonell, C., R. Chacón, and H. Posada. (2023). “Knowledge graph-based data integration system for digital twins of built assets.” *Automation in Construction* 156: 105109.
Waltl, B., G. Bonczek, and F. Matthes. (2018). *Rule-based Information Extraction: Advantages, Limitations, and Perspectives*.
Wu, L.-T., J.-R. Lin, S. Leng, J.-L. Li, and Z.-Z. Hu. (2022). “Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web.” *Automation in Construction* 135: 104108.

Zhang, J.-R., and B. He. (2022). “Named Entity Recognition Model Based on TextCNN-BiLSTM-CRF with Chinese Text Classification.” *Journal of Computers* 33(2): 115-126.
Zhang, N., X. Xu, L. Tao, H. Yu, H. Ye, S. Qiao, X. Xie, X. Chen, Z. Li, and L. Li. (2022). “Deepke: A deep learning based knowledge extraction toolkit for knowledge base population.” arXiv preprint arXiv:2201.03335.
Zhang, N., J. Ye, Y. Zhong, and Z. Chen. (2023). “Digital Transformation in the Chinese Construction Industry: Status, Barriers, and Impact.” *Buildings* 13(4): 1092.