Interpreting accident reports by integrating a heterogeneous graph neural network and factor analysis

Junyu Chen, Hung-Lin Chi, Bo Xiao, Rongyan Li
The Hong Kong Polytechnic University, Hong Kong SAR
hung-lin.chi@polyu.edu.hk

Abstract. Occupational safety in the construction industry is one highly prioritized concern around the globe. Accident reports are considered valuable recourses preserving information about corresponding risk factors. Many efforts in the literature have demonstrated that deep learning models are readily applicable to processing and analyzing narrative reports. However, the heterogeneous semantic information was rarely considered. This research utilizes knowledge graph-based accident analysis to provide a machine-assisted approach for construction accident report interpretation. To validate the proposed approach, this research labels 320 crane-related accident reports from the US OSHA database and develops a Crane Safety Knowledge Graph (CSKG) as a case study. Then, a Heterogeneous Graph Attention Network (HAN) is trained to explore the accident features and the importance of various risk factors. Through mapping and clustering the accident data points, the results reveal the capability of the proposed approach to learn the accident patterns and generate safety rules for construction cranes.

1. Introduction

Improving occupational safety is a challenging issue across the globe. In Europe, a total of 3581 occupational fatal accidents were recorded in the year 2018, where 716 were from the construction industry (Eurostat, 2022). In the United States, 154 construction work-related fatalities and catastrophes were reported in the year 2022 (Occupational Safety and Health Administration, 2023). The situations are even worse in developing countries and one common feature revealed by statistical data from different countries is that the construction industry is liable for recommendation systems (Wang et al., 2019). In knowledge representation and reasoning, KG has become an effective tool to retrieve, analyze and visualize heterogeneous semantic information. A generic KG is composed of a data layer and a schema layer. The data layer contains domain knowledge, utilizing entities, attributes, and relations to represent semanti case basis, seeking attention to corresponding precautionary measures. However, for different accident cases, there is a variety of risk scenarios that present a wide spectrum of on-site conditions such as misoperations, malfunctions, and constraints. Hence, the first challenge associated with the utilization of textual accident data can be identified as how to properly extract and manage the accident information for inference, such as eliciting general safety rules. As defined by Gruber, ontology denotes the specification of conceptualization in a formal and explicit manner; ontology modeling is the process to model a set of concepts and their relationships into ontologies within a knowledge domain (Gruber, 1993). In this research, ontology modeling is potentially a promising approach to accommodate the multifariousness of different accidents while preserving the generalization ability of the developed KG, transforming the unstructured textual accident data into a structured accident database.

The latent features of accidents (e.g., the importance of various risk factors and the accident patterns) preserved in the accident-enabled KG can be learned by an HGNN, which is generally comprised of layers for input features, knowledge inference, and output predictions. Referring to previous research: 1) one-sentence accident summaries can be used as inputs through word embedding. This process utilizes a corpus of text and an embedding method to reconstruct the word sequence of accident summaries into a vector space. 2) the prediction of accident consequences, represented as accident types, are the expected outputs from

where *q* is the vector containing semantic attention values of node pairs from various meta-paths between the entity set ; evaluates the information contribution of the meta-path .

Step 3: Fuse all the semantic-specific embeddings with the corresponding meta-path weights to generate the final embedding of node *i*:

Figure 2: The hierarchical attention structure in the HAN model

Analysis of The Overall HAN Model.

4. Case Study

Data Collection. Fatality and Catastrophe Investigation Summaries reported by the US OSHA were considered in this research. The case study focused on construction crane safety in the past two decades. The scope of data collection was hereby specified by using the keyword "construction crane" for retrieval on the US OSHA website. Through a careful interpretation process, the accident reports not involving crane usage or construction activities were ruled out and a total of 320 cases were compiled in the final accident database in this research.

Ontology-Based Knowledge Extraction. As shown in Table 1, the labeling for the accident consequences was following the Top Four construction hazards identified by the US OSHA. The basic accident information considered the features of involved construction sites and activities. For accident causations, the primary causes were categorized as human errors, mechanical problems, and environmental hazards. The original accident narratives were labeled from each information aspect and category, determining the corresponding attributes of the accident entities.

Implementation of The HAN Model. Referring to (Xiao Wang et al., 2019), the configuration training loss was close to 0; the validation loss was smaller than 0.3. The training results demonstrated the reliability of the trained HAN model for the sampling accident data.

Figure 4: The iterative process of model training

Table 1: The aspects of accident information and the corresponding ontology relations.

Aspects	Ontology relations	Aspects	Ontology relations
Accident Type	Fall from the personnel basket (AT0)	Human Error	Operator misoperation (A_HEa_A)
	Fall from the extension ladder (AT1)		Rigger misoperation (A_HEb_A)

Representation of Accident Entities and Weights of Meta-Paths. Through the training process, each accident node was represented as a 64-dimensional vector. To visualize the 320 accident entities as data points as well as retain the information from the final embeddings, the t-SNE method was used to map the 320 data points as shown in Figure 5(a). The normalized weights of meta-paths were also obtained from the trained HAN model, indicating the different importance of various causal factors leading to crane accidents. As indicated in Table 2, considering the information revealed by meta-paths with a weight higher than 0.01, research findings were summarized from three aspects: 1) from the human error aspect, the misoperation of maintenance workers or inspectors was identified as an important risk factor for crane operation; 2) from the mechanical problem aspect, the fall of the crane boom or jib was revealed as an essential risk factor for crane operation, followed by malfunction or failure of the crane, and crane collapse; 3) from the environmental hazard aspect, poor weather or operation conditions were essential for crane operation, followed by a lack of clear division of work area, and a lack of PPE or communication devices.

Figure 5: Visualization of the 320 accident data points (a) using t-SNE; (b) using fuzzy c-means

Table 2: Obtained weights of meta-paths for three time periods.

Meta-path code	**Normalized Weight**	**Meta-path code**	**Normalized Weight**	**Meta-path code**	**Normalized Weight**	**Meta-path code**	**Normalized Weight**

3) From Cluster 6, it was found that some cases of the body caught in or between (AT7) and the finger/hand/foot caught in or between (AT8) were highly associated, indicating the most significant contributing factor as not properly separating the operating crane from surrounding workers.

4) 95.32 841.0 g0 G6BT/F1 12 Tf1 0 0 1 112.58 718.66 Tm0 g0 G[( )]TJETQEMC

References

Davies, D. L. and Bouldin, D. W. (1979) 'A cluster separation measure', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2), pp. 224 227.

Dhalmahapatra, K., Shingade, R. and Maiti, J. (2020) 'An innovative integrated modeling of safety data using multiple correspondence analysis and fuzzy discretization techniques', *Safety Science*, 130(January), p. 104828. doi: 10.1016/j.ssci.2020.104828.

Dunn, J. C. (1974) '