CP4892

Deep Learning-Based Named Entity Recognition and Resolution of Referential Ambiguities for Enhanced Information Extraction from Construction Safety Regulations

Xiyu Wang[1] and Nora El-Gohary[2]

[1] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 N. Mathews Ave., Urbana, IL 61801, United States. Email: xiyuw2@illinois.edu
[2] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 N. Mathews Ave., Urbana, IL 61801, United States. Email: gohary@illinois.edu

Abstract

Construction safety regulations and standards contain a massive number of fall protection requirements with respect to different equipment, facilities, and operations. Automated field compliance checking aims to detect field violations to construction safety regulations for improved compliance and safety. Recent research efforts focused on automated tracking of labor and equipment towards improved violation detection and safety compliance. However, extracting and modeling safety requirements for supporting automated violation detection or safety alert systems remains highly manual. Towards addressing this gap, information extraction provides an opportunity to automatically extract requirements from construction safety regulations for comparisons with field information to detect violations (or predict and prevent violations before they occur). However, existing information extraction methods are limited in terms of their scalability and/or accuracy. To address this need, this paper proposes a deep learning-based information extraction method for automatically extracting named entities describing fall protection requirements (e.g., scaffold, horizontal direction, 6 feet) from construction safety regulations and resolving referential ambiguities. The proposed information extraction method consists of three main submethods: (1) a deep learning-based method to recognize entities from the regulations, (2) a deep learning-based method to recognize referential ambiguities in the extracted entities, and (3) a named entity normalization method to resolve these ambiguities. The proposed method was implemented and tested on 20 selected Occupational Safety and Health Administration (OSHA) sections related to

fall protection. An overall information extraction precision, recall, and F-1 measure of 93.2%, 89.6%, and 91.1% were obtained, which indicates good information extraction performance.

1 Introduction

Fall accidents are a major concern for construction safety. A total of 1,102 fatalities in the construction industry were reported in 2019, which represented 20.7% of the total workplace fatalities in the United States (5,333) (OSHA, 2020a; Tang et al. 2020); and 174,100 injuries, with more than 130,000 workers missing days of work, were reported in 2020 (Labor 2021a; Labor 2022). Among all accident types, falling is the leading cause of construction fatalities. It is responsible for more than 30% of all construction deaths (Labor 2021b; Mutual 2020) and is the second leading cause of serious injuries, with higher compensation costs than other types of injuries (OSHA, 2020b).

A large number of fall accidents happen due to field noncompliance with safety regulations, particularly the Occupational Safety and Health Administration (OSHA) regulations. For example, an analysis of fall fatalities in the Construction FACE Database (CFD) revealed that compliance of personal fall arrest systems (PFAS) plays an important role (Dong et al. 2017). Among those fall fatalities, 54% occurred when PFAS were not available, and 23% when the decedents had access to PFAS but were not using them. Some other fall decedents were using PFAS, but the PFAS were either damaged, misused, or did not provide adequate protection. One major reason for such lack of field compliance is that existing onsite safety inspections are not effective. For example, a study showed that 61.5% of the surveyed workers complained that there was no safety supervision during their work (Tadesse et al. 2016). Although OSHA requires “a competent person” to make compliance decisions onsite (OSHA 2020c), those manual observers typically only produce biweekly or monthly safety reports (Tang et al. 2020). At this frequency, potential noncompliance issues can be neglected or stay unresolved, which can possibly lead to serious accidents. It is also difficult to carry out field compliance checking promptly using pure human labor, because (1) skilled supervisory manpower for different activities is not always present onsite (Seo et al. 2015), and (2) manual observation and supervision is labor intensive, time consuming, and error prone (Chen et al. 2019).

Therefore, there is a need for automated field compliance checking to help detect and correct field noncompliance in a timely manner to prevent fall accidents. Automated field compliance checking seeks to automate the process of extracting safety requirements from relevant regulations, capturing relevant site conditions, comparing conditions to requirements to detect violations, and producing prompt feedback to relevant workers. Existing research efforts on field compliance checking have achieved good progress on automated tracking of labor and equipment onsite towards improved safety compliance (e.g., the work by Fang et al. 2018 to detect the existence of PPE). However, extracting construction safety requirements from applicable regulations and representing these requirements in a computable format for subsequent compliance checking is still conducted in a largely manual way. For example, research efforts have proposed manual ontology-based approaches to extract hazard or safety knowledge from fatality reports or industry safety best practice reports (e.g., Zhong et al. 2020b). Such manual process is expensive and unscalable, considering the large number of requirements to be extracted from various safety regulatory documents. A fully automated information extraction method is thus needed.

However, automated information extraction from construction safety regulations is still challenging, despite the current information extraction efforts in the construction domain (Zhong et al. 2022; Zhang and El-Gohary 2021a; Ren and Zhang 2021; Zhou and El-Gohary 2017; Zhang and El-Gohary 2013; Nepal et al. 2013). On one hand, sentences from construction safety regulations are more complex compared to other types of text (e.g., international building codes and energy conservation codes). Such complexity includes (1) the text has referential ambiguities (e.g., multiple expressions are used to refer to the same entity). In other types of text, relative or attributive clauses (which typically contain more coreferents) are less frequently used, and hence the text contains fewer referential ambiguities; (2) different OSHA sections have different text patterns and different ways of organizing requirements for one topic and its subtopics, and (3) a single clause could have nested conditions and exceptions to describe a particular scenario that involves multiple interactions and spatial relations between workers and their environment. Thus, the density of information in a single sentence from safety regulations is relatively high compared to other types of text. It is rather difficult to achieve good performance given such complex text. On the other hand, existing information extraction methods in the construction domain are limited. First, most of the aforementioned efforts used rule-based methods whose performance relies heavily on a set of hand-crafted rules, which require significant amount of human effort to discover the text patterns and develop the corresponding information extraction rules, and are difficult to scale up across other documents with different text patterns. A small number of information extraction methods used traditional machine learning-based methods, which also suffer from level of effort and scalability limitations due to their dependence on traditional feature engineering. In comparison, deep learning-based methods can automatically extract syntactic and semantic features from unstructured text, instead of using hand-crafted rules or highly engineered features, which minimizes the amount of human effort involved in the extraction process and improves the scalability of the approach. Second, none of the aforementioned efforts addressed the problem of referential ambiguity.

To address these limitations, this paper proposes a deep learning-based information extraction method to automatically extract entities that describe fall protection requirements from construction safety regulations, particularly OSHA. The proposed method consists of three main submethods. First, a deep learning-based method is proposed to recognize entities from the regulations. The method uses three types of features to improve performance: GloVe embedding, word-level features, and character-level features. Second, a deep learning-based method is proposed to recognize referential ambiguities in the extracted entities. The method uses transfer learning to deal with the lack of annotated training data, leveraging both out-of-domain, large-scale annotated data together with domain-specific data (construction safety regulations). Third, a named entity normalization method is proposed to resolve these ambiguities. The method measures the similarity between the recognized ambiguous expressions and a list of candidate identifier names to identify their correspondence. The proposed method was tested using fall-related sections from the OSHA 29 CFR 1926 (OSHA 2020c), and the CoNLL-2012 dataset (Pradhan et al. 2012) from the computational linguistic domain.

2 Background

2.1 Named Entity Recognition

Named entity recognition identifies and classifies entities from unstructured text into pre-defined categories (Chiu and Nichols 2016). In the context of construction safety, those entities are semantic information elements describing fall protection requirements such as “scaffold”, “horizontal direction”, “6 feet”. These requirements could be classified as quantitative or existential. Quantitative requirements describe the properties of fall protection measures, e.g., “Each end of a platform, unless cleated or otherwise restrained by hooks or equivalent means, shall extend over the centerline of its support at least 6 inches (15 cm)”. Existential requirements describe the existence of fall protection measures, e.g., “Unstable objects shall not be used to support scaffolds or platform units”.

Various traditional (i.e., not deep learning) machine learning algorithms have been used for named entity recognition, such as support vector machines (SVM) (Isozaki and Kazawa 2002), conditional random fields (CRF) (Lafferty et al., 2001), and hidden Markov model (HMM) (Bikel et at. 1998). Before the advent of deep learning, CRF was the dominant model for named entity recognition. It has, for example, achieved an F-1 measure of 81.15% on the CoNLL-2003 dataset (Yadav and Bethard 2019). Machine learning-based methods have also been utilized within the construction domain. For example, Liu and El-Gohary (2017) used an ontology-based, semi-supervised CRF approach to extract bridge-related entities from bridge inspection reports. Kim and Chi (2019) used a CRF approach to extract safety knowledge from construction accident reports.

2.2 Coreference Resolution

Coreference resolution aims to identify all mentions that refer to the same real-world entity (Lee et al. 2017). In construction safety regulations (e.g., in contrast to building codes), coreferents such as “it” or “them” are frequently used to refer back to an entity mentioned earlier. Depending on the context, sometimes it is even difficult for human to interpret which entity the coreferent is referring to. Such referential ambiguity can cause confusion for information extraction. Similarly, different expressions are used in the same document to refer to the same entity. For example, in OSHA, “two-point adjustable suspension scaffold” and “two-point scaffold” are used at different instances to refer to the same entity. Such use of different expressions can introduce errors in the extracted requirements. Coreference resolution is, thus, vital to prevent referential errors from causing errors in information extraction and propagating into compliance checking errors.

Traditional machine learning-based methods for coreference resolution can be divided into three categories: mention-pair, entity-mention, and mention-ranking models. Mention-pair models, which are the most commonly used for coreference resolution, regard coreference as a pairwise classification task. Entity-mention models classify whether a mention belongs to a preceding coreference cluster. However, these two methods fall short in determining which candidate antecedent is the best for prediction. Mention-ranking models can solve this problem by explicitly ranking all candidate antecedents for each mention. These models have, for example, achieved 54.1%, 54.3%, and 56.6% B^3 F-1 on the ACE2004 dataset, respectively (Rahman and Ng 2009).

2.3 Named Entity Normalization

Named entity normalization seeks to map different mentions of an entity, such as ambiguous surface forms or synonyms, into one canonical form (an identifier name) (Jijkoun et al. 2008). It has been extensively studied in the past few decades and has been adopted for various semantically oriented applications such as question answering, entity retrieval, trend detection, and event tracking (Jijkoun et al. 2008). Especially in recent years, there is a growing body of literature in the biomedical domain proposing different methods for normalizing ambiguous names of chemicals, genes, and diseases (Zhou et al. 2020; Cho et al. 2017; Leaman et al. 2015). In the construction domain, a research effort has proposed to normalize different expressions of the entity names extracted from bridge inspection reports (Liu and El-Gohary 2018).

Methods for named entity normalization can be divided into two categories: dictionary-based and machine learning-based methods. Dictionary-based methods normalize entity names with the help of pre-established knowledge in the identifier names involved, usually stored in the form of gazetteer lists (Yenkar and Sawarkar 2021; Nadeau et al. 2006), lexicons (Névéol et al. 2015), and knowledge bases (Zhou et al. 2007). Machine learning-based methods (Zhou et al. 2020; Cho et al. 2017; Leaman et al. 2015), on the other hand, are more suitable for situations where such pre-established knowledge is unavailable. Instead, it learns to normalize entity names by utilizing a set of features such as frequency, part-of-speech tags, and lemma of each entity name.

2.4 Deep Learning-Based Methods

Deep learning-based methods use stacked neural networks that automatically extract features and patterns from large-scale unstructured text. They can achieve better results than traditional machine learning models by allowing different information paths through the connected neurons. Deep learning has recently been utilized for extracting information from regulatory documents and proved to be promising. For example, Zhang and El-Gohary (2021b) have developed a long short term memory (LSTM) model to extract requirement hierarchies from building codes and standards. Outside the construction domain, many deep learning methods have been proposed to improve the performance of named entity recognition and coreference resolution. Those methods include: (1) proposing new deep learning architectures based on existing ones, (2) creating hybrid models by combining existing ones, such as LSTM combined with convolutional neural network (CNN), and (3) developing state-of-the-art word embeddings as additional features such as embeddings from language models (ELMo) (Peters et al. 2018) or the global vector (GloVe) (Pennington et al. 2014). For example, for named entity recognition, bi-directional LSTM and CNN (BiLSTM-CNN) (Chiu and Nichols 2016) has achieved 91.6% F-1 on the CoNLL-2003 dataset. For coreference resolution, a Neural Coreference Resolution model proposed by Lee et al. (2017) has achieved 67.2% B^3 F-1 on the CoNLL-2012 dataset. Existing research efforts on deep learning-based named entity normalization have mainly focused on experimenting with different word embeddings or similarity metrics for improving the normalization performance (Yuan et al. 2022; Fang et al. 2021; Fakhraei et al. 2019; Roy et al. 2018).

3 State of the Art and Knowledge Gaps

In the area of named entity recognition, a number of research efforts in the construction domain have been undertaken to propose different methods for improving the extraction performance. A large portion of them have focused on developing rule-based extraction methods. For example, Ren and Zhang (2021), Zhou and El-Gohary (2017), and

Zhang and El-Gohary (2013) developed rule-based information extraction methods to extract from construction regulatory documents such as construction procedural documents, building codes, and energy conservation codes, through the use of syntactic and semantic features. Later research efforts have explored utilizing and comparing different traditional machine learning-based methods. For example, Zhang et al. (2019) compared a set of machine learning-based methods such as SVM, linear regression, K-nearest neighbor, decision tree, Naive Bayes, and an ensemble model for text mining from construction accident reports. Liu and El-Gohary (2017) and Kim and Chi (2019) proposed CRF-based methods to extract information from various domain-specific documents such as bridge inspection reports and construction accident reports. Recently, deep learning-based methods have been attracting more research attention. A few LSTM-based methods have been proposed, such as the efforts by Zhang and El-Gohary (2021a) and Zhang and El-Gohary (2021b) to extract semantic and syntactic information elements and requirement hierarchies from building codes.

Despite the importance of these efforts, three main knowledge gaps still exist. First, most of the existing information extraction methods are limited in terms of level of manual effort and scalability. Rule-based methods usually require significant amount of human effort in discovering the text patterns and developing the corresponding information extraction rules. The rules also require adaptation (additions or changes) across different documents, especially if the characteristics of the text change, which further limits scalability. Traditional machine learning-based methods do require less human involvement and offer better scalability, but their performance depends heavily on the quality of the engineered features which are obtained through trial and error. Deep learning-based approaches, on the other hand, are more promising in terms of reducing human effort and improving scalability due to their ability to automatically capture various syntactic and semantic features and patterns from the text, thereby eliminating the manual effort needed to develop the hand-crafted extraction rules or conduct traditional feature engineering. In general, traditional feature engineering can become labor-intensive and time-consuming (Janiesch et al. 2021; Dargan et al. 2020). Second, methods with better performance to deal with the complexity in the text are desired. As discussed in the “Introduction” section, sentences from construction safety regulations are more complex in three aspects. However, most of the existing traditional machine learning-based methods perform shallow extraction by using pattern-based grammars with domain-specific interpretations. Such pattern-based grammars cannot capture complex linguistic features such as long distance dependencies and passive/active voices which are frequently used in construction safety regulations. Those traditional machine learning-based extraction methods, thus, lack the ability to extract the entire semantics which are essential for detecting field noncompliances to prevent fall accidents. Therefore, methods for deeper information extraction need to be developed, to fully understand the semantics in construction safety regulations. Deep learning-based methods are such methods that can perform deeper information extraction. They have outperformed traditional machine learning-based methods in many applications, including information extraction from building codes (Zhang and El-Gohary 2021a), and are flexible in dealing with various patterns in the text. It is then necessary to further explore their use in extracting information from construction safety regulations that contain high complexity, to achieve good extraction performance for supporting automated field compliance checking. Third, previous efforts have only considered information extraction from quantitative requirements (Zhang and El-Gohary 2013). However, as discussed in the "Background" section, there are two types of requirements in construction safety regulations, which are equally considered in this study.

Most importantly, existing information extraction efforts in the construction domain have mainly focused on named entity recognition, without resolving the referential ambiguities in the extracted requirements. The closest effort, for example, is the named entity normalization method proposed by Liu and El-Gohary (2018) to map the referring expressions from bridge reports to their identifier concepts. There are three main knowledge gaps that this research aims to address in this regard. In terms of scope, there are more types of referential ambiguities than referring expressions in construction safety regulations, including coreferents. Both types of referential ambiguities can directly affect the performance and effectiveness of subsequent compliance reasoning. Thus, there is a need to resolve these referential ambiguities in the extracted requirements. In terms of data, since referential ambiguities do not appear in every sentence, there is no sufficient training data from the construction safety domain alone, which is especially a challenge for deep learning based-methods because they typically require more training data than rule-based or traditional machine learning-based methods. There is, thus, a need for leveraging the large amount of annotated data from other domains to resolve referential ambiguities in the construction safety domain. In terms of method, most of the existing rule-based methods for resolving referential ambiguities cannot be applied in the construction safety domain directly. For example, gender agreement and person agreement that could apply for general-domain text or other applications, are not applicable to construction regulations and safety topics. Moreover, these rules and their orderings are often subject to changes from one topic to another. Given the large number of topics involved in constructions safety regulations, there is a need to develop methods that can automatically capture the distinctive patterns in the domain-specific text.

4 Proposed Method for Information Extraction from Construction Safety Regulations

This study proposes a deep learning-based information extraction method to automatically extract entities that describe fall protection requirements from construction safety regulations and to resolve referential ambiguities in the extracted results. The proposed information extraction method consists of three main submethods for named entity recognition, coreference resolution, and named entity normalization. Named entity recognition aims to recognize entities (e.g., “scaffolds”) and classify them into pre-defined entity classes (e.g., equipment). The entity classes were pre-defined based on a review of 20 OSHA sections related to fall protection, and a review of ontology-based modeling of construction safety knowledge (Zhang et al. 2014; Lu et al. 2015; Zhang et al. 2015; Zhong et al. 2020a; Fang et al. 2020). The pre-defined entity classes include person, equipment, reference, hazard, facility, location, operation, material, property, date, other attribute, quantity value, quantity unit, and other entity. Table 1 shows examples of the most frequent entity names within each class. A BiLSTM-CNN-based model was trained to automatically recognize and classify the entities based on their syntactic and semantic features. Coreference resolution aims to identify all mentions that refer to the same entity, including coreferents and referring expressions. Transfer learning was used to deal with the lack of training data. For the first two submethods, three types of features were used to improve the performance of the models: GloVe embedding, word-level features, and character-level features. Named entity normalization aims to map the recognized coreferents and referring expressions of an entity to one identifier name to remove the referential ambiguities in the extracted requirements. Similarity assessment was conducted to map the recognized different mentions to their closest identifier names. Fig. 1 summarizes the research methodology, which includes six primary tasks: data preprocessing, feature preparation, named entity recognition, coreference resolution, named entity normalization, and evaluation. Fig. 2 further illustrates the application of the proposed method, with an example.

4.1 Data Preprocessing

Data preprocessing aims to process the raw text to be ready for the subsequent steps of information extraction. Two preprocessing methods were used: tokenization and sentence splitting. Tokenization divides a character sequence in the text into units (words). Sentence splitting detects the boundary of each sentence by recognizing the sentence-ending characters such as periods and questions marks.

4.2 Feature Preparation

Three types of features were utilized to improve the performance of both the named entity recognition model and the coreference resolution model: GloVe embedding, word-level features, and character-level features. GloVe embedding is the state-of-the-art word embedding, which is pre-trained on Wikipedia and Web text of 6 billion words. It represents the semantics of words in the form of rich and dense feature vectors. Additional word-level features and character-level features, which are not included in the GloVe embedding, were added to the embedding because they help differentiate the different entities (e.g., equipment vs. quantity value). The word-level features include four types of information: (1) if the word is all lower-cased, (2) if the word is all upper-cased, (3) if the word contains numbers, and (4) if the word contains capital letters. The character-level features include: (1) if the character is punctuation, (2) if the character is a digit, (3) if the character is uppercase, and (4) if the character is lowercase.

4.3 Named Entity Recognition

The deep neural network, BiLSTM-CNN, by Chiu and Nichols (2016) was adopted for named entity recognition. This hybrid model was chosen for its potential to achieve better extraction performance, because it can combine the benefits of both BiLSTM model and CNN model which are designed with different strengths. BiLSTM is better at capturing context and long dependency, while CNN is better at capturing character-level information, both of which can contribute to more accurate predictions. The proposed BiLSTM-CNN-based model contains three main types of layers: embedding layers, BiLSTM layer, and multi-layer perceptron (MLP) layers. The embedding layers consist of a GloVe embedding layer, a word-level feature embedding layer, and a CNN-extracted character embedding layer. The GloVe embedding layer uses the pre-trained embedding as a starting point, then adjusts itself to the semantics of construction safety text during training. The word-level embedding layer represents the word-level features described in Section 4.2. The CNN-extracted character embedding layer is used to represent the character-level features prepared in Section 4.2, as well as other character-level features extracted using CNN, such as prefix and suffix. The outputs from the three embedding layers are concatenated before being fed into the BiLSTM layer. The BiLSTM layer is used to compute the feature values using the output from the embedding layers of the current word and its context words. The MLP layers consist of a linear layer and a softmax layer, which transform the feature values from the BiLSTM layer into log probabilities for the tag categories, where tags with the highest probabilities are returned as predictions. Cross entropy was used as the loss function. The BiLSTM-CNN architecture is illustrated in Fig. 3.

4.4 Coreference Resolution

Coreference resolution aims to identify referential ambiguities in the document, i.e., all mentions that refer to the same entity, including coreferents and referring expressions. Different types of coreferents and referring expressions were first identified based on an analysis of the selected OSHA sections. Examples of different types of coreferents and referring expressions are shown in Tables 2 and 3, respectively.

A deep learning model was adapted and trained using transfer learning strategy for coreference resolution. Transfer learning is used to leverage rich syntactic and semantic information from existing large-scale annotated source-domain data for solving problems in a target-domain (construction safety domain in the problem at hand). The CoNLL-2012 dataset was used as source-domain data. It was developed in the computational linguistic domain to predict coreferences in English, Chinese, and Arabic. The English portion contains around one million words from various sources such as newswire, magazines, broadcasts, weblogs, and speeches. The deep neural network by Lee et al. (2017) was adopted. The task is formulated as finding antecedent y_i for every possible span i in the document, where each span is represented by considering two main factors: the headword and the context. For a given span, possible antecedents could be all the spans before itself. If no antecedent is found for a span, it is because either the span is not an entity mention, or the span is an entity mention but is not a coreferent with any previous span. A pairwise score is used to measure the similarity between two spans by considering three factors: (1) if span i is a mention, (2) if span j is a mention, and (3) if j is an antecedent of i. To reduce computational complexity, low scoring spans are pruned.

The deep learning model contains three main types of layers: embedding layers, BiLSTM layer, and the MLP layers. The embedding layers and BiLSTM layer are similar to the model used for named entity recognition. However, an attention mechanism (Vaswani et al. 2017) was added to the BiLSTM layer to model the headwords of each span in the form of a weighted vector. The output of the BiLSTM layer and the headwords vectors are then concatenated to produce the span representation. In the MLP layers, pairwise scores are calculated using the span representation. Headwords of the spans with the highest pairwise score are considered as coreferents. The marginal log-likelihood of correct coreferents implied by the gold standard was used as the loss function. The architecture of the coreference resolution model is illustrated in Fig. 4.

4.5 Named Entity Normalization

Named entity normalization aims to map different mentions of an entity to one canonical identifier name to resolve the referential ambiguities for supporting subsequent compliance reasoning. Those different mentions include the coreferents and referring expressions recognized during coreference resolution. Therefore, the proposed named entity normalization method seeks to map the recognized different mentions to their identifier names automatically with the help of a domain-specific dictionary. The proposed named entity normalization method includes three main steps: candidate identifier name extraction, similarity assessment, and identifier name mapping.

The OSHA sections that cover the scope, application, and definitions of the OSHA subparts were used as a domain-specific dictionary. This is because these sections include definitions of the terms covered in each subpart, such as specific tools and equipment. Each of these sections typically contains named entities related to various accident types, not only fall protection-related topics, because each OSHA subpart consists of a few sections addressing different safety topics. For example, section 1926.1400 describes the scope for 1926 Subpart CC, which includes a variety of topics such as power line safety, signal person qualification, and fall protection, and thus contains more entities than fall protection-related entities. Therefore, they are sufficient in covering most of the fall-related entity names. Entity names were automatically extracted from this domain-specific dictionary (i.e., the scope, application, and definitions sections) using the trained named entity recognition model (in Section 4.3). The extracted entities were then used as candidate identifier names.

Similarity assessment aims to assess the similarities between the extracted identifier names, as well as the output from the coreference resolution task, i.e., multiple clusters each containing different mentions of an entity in one clause. The proposed similarity assessment is an embedding-based method, similar to the method proposed by Farouk (2020). To measure the similarities, these two sets of entities (candidate identifier names and identified different mentions) were converted to a vectorized representation, using the embedding layer from the trained named entity recognition model. This is because the trained embeddings can encode the semantic information of domain-specific entities such that entities that are similar in meanings are closer in the embedding space. A similarity matrix was calculated between each cluster of different mentions and each candidate identifier name, using an average embedding for one entity name (excluding the embeddings of the stopwords which are too frequently used to provide distinctive information). The similarity scores in the similarity matrix were calculated using cosine similarity, which has shown better performance in capturing the similarity between texts (Sitikhu et al. 2019), compared with the Euclidean similarity (which is frequently used in other applications). This is because cosine similarity measures the directions of vectors to ensure that entities containing different meanings (i.e., vectors pointing in different directions in the embedding space) receive lower similarity scores. Then, for different mentions in one cluster, the identifier name with the highest similarity score among all other candidates is selected for normalization.

After similarity assessment, identifier name mapping is conducted to convert the different mentions in each cluster to their corresponding selected identifier name. This mapping process is conducted automatically through a lookup table storing the correspondence between these two sets of entities. In this way, different mentions of an entity, including coreferents and referring expressions, can all be replaced with a well-established entity name in the construction safety domain.

4.6 Evaluation

The performance of named entity recognition, named entity normalization, and overall information extraction (after coreference resolution and named entity normalization) was evaluated by comparing the recognized/extracted entities with the gold standard using three metrics: precision (*P*), recall (*R*), and F-1 measure, as per Eqs. 1-3. Precision is defined as the number of correctly recognized/extracted entities divided by the total number of recognized/extracted entities. Recall is defined as the number of correctly recognized/extracted entities divided by the total number of entities that should be recognized/extracted. F-1 measure is the weighted harmonic mean of precision and recall. Due to data imbalance, macro average of precision, recall, and F-1 measure were used to evaluate performance – to avoid majority classes (those with larger instances) skewing the results. Macro average provides an average over classes, thereby weighing all classes equally, as opposed to micro average that provides an average over instances.

$$P = \frac{\text{number of correctly recognized/extracted entities}}{\text{total number of all recognized/extracted entities}} \tag{1}$$

$$R = \frac{\text{number of correctly recognized/extracted entities}}{\text{total number of entities that should be recognized/extracted}} \tag{2}$$

$$F\text{-}1 = \frac{2\times \mathrm{P}\times \mathrm{R}}{P+R} \tag{3}$$

For coreference resolution, the performance was evaluated using the B^3 precision, recall, and F-1 measure (Bagga and Baldwin 1998), as per Eqs. 4-6, where is *n* is the number of entities that have coreferents.

$$B^3\ P = \frac{1}{n}\sum_i \frac{\text{number of correctly recognized coreferents for entity } i}{\text{total number of all recognized coreferents for entity } i} \tag{4}$$

$$B^3\ R = \frac{1}{n}\sum_i \frac{\text{number of correctly recognized coreferents for entity } i}{\text{total number of coreferents that should be recognized for entity } i} \tag{5}$$

$$B^3\ F\text{-}1 = \frac{2\times \mathrm{B}^3\ \mathrm{P}\times \mathrm{B}^3\ \mathrm{R}}{\mathrm{B}^3\ \mathrm{P}+\mathrm{B}^3\ \mathrm{R}} \tag{6}$$

5 Experimental Results and Discussion

The proposed information extraction method was tested using OSHA sections related to fall protection. A set of experiments were conducted to evaluate: (1) the performance of the proposed deep learning-based named entity recognition method by comparing it with a baseline method (Section 5.2); (2) the impact of different transfer learning strategies on the proposed coreference resolution method (Section 5.3); (3) the impact of the three types of features used in the proposed models on named entity recognition and coreference resolution (Sections 5.2 and 5.3); (4) the performance of the named entity normalization method (Section 5.4); and (5) the overall information extraction performance, including the three constituent submethods and their collective effectiveness in addressing referential ambiguities (Section 5.5). The hyperparameters of the named entity recognition and coreference resolution models were fine-tuned during the experiments for improved performance. The experiments were implemented using Keras and tensorflow on NVIDIA GeForce RTX 2070 SUPER.

5.1 Data Preparation and Gold Standard Development

All 20 OSHA sections that are related to fall protection were selected for developing the dataset for training and testing, which cover a number of topics such as personal fall arrest systems, fall protection systems, guardrail systems, positioning device systems, scaffolds, ladders, and aerial lifts, as per Table 4. The resulting dataset included 2,091 sentences, which were split into a training and validation dataset and a testing dataset at a ratio of 8:2. The testing dataset included 418 sentences, 7312 words (prior to resolving referential ambiguities), 169 cluster of different mentions, and 7334 (after resolving referential ambiguities). The first dataset was further split into a training set and a validation set at the same ratio. The dataset was annotated to create the gold standard for training and testing. The gold standard was developed by three annotators who have background in both civil engineering and natural language processing. An inter-annotator agreement of 94.5% in F-1 measure was achieved, which indicates the reliability of the gold standard (Artstein 2017). For named entity normalization, the following sections (those acting as the dictionary, which cover all necessary sections for describing the scope and applications) were used for extracting the identifier names: 1926.20, 1926.107, 1926.450, 1926.500, 1926.750, 1926.751, 1926.1050, and 1926.1400.

For named entity recognition, the entities were annotated using the following labels: person (PER), equipment (EQU), reference (REF), hazard (HAZ), facility (FAC), location (LOC), operation (OPE), material (MAT), property (PRO), date (DAT), other attribute (ATT), quantity value (QUA), quantity unit (UNI), and other entity (ENT). To distinguish adjacent entities with the same tag, a BIO tagging scheme was used, where “B” denotes the beginning of an entity, “I” stands for “inside”, and “O” means the absence of an entity. An example sentence annotated using the BIO tagging scheme is shown in Table 5. Fig. 5 depicts the entity class distribution, which shows that the developed dataset is unbalanced, with equipment (EQU) being the most frequent class and date (DAT) the least frequent class.

For coreference resolution, the sentences (i.e., our domain-specific data) were annotated following the same tagging scheme as the CoNLL-2012 dataset (i.e., the general-domain data), where each unique entity name and its different mentions in one clause were assigned the same index. An example clause annotated using this tagging scheme is shown in Fig. 6(a). Two special situations were considered during the annotation. First, an equipment and its components were not regarded as coreferents, because OSHA regulations can contain specific requirements about components as well. This is different from other types of text (e.g., social media text) where one feature or one component of an entity can be used to refer to the whole. Second, due to the hierarchical structure of the OSHA regulations, one clause and its subclause(s) can refer to the same entity (e.g., bricklayers’ square scaffolds) even though a more general entity name (e.g., scaffolds) is used in the subclause(s). An example of this case is shown in Fig. 6(b). The annotated OSHA clauses were then combined with the CoNLL-2012 dataset for transfer learning. For named entity normalization, each cluster of different mentions was annotated with the correct identifier name.

The final gold standard for evaluating the overall information extraction performance was developed based on the gold standard for named entity recognition, but with coreferents and referring expressions normalized.

5.2 Named Entity Recognition

5.2.1 Optimization and Performance Results

To optimize the performance of the named entity recognition model, the hyperparameters of the model were fine-tuned. During optimization, the model used a kernel size of 3 and a dropout rate of 0.5, and was trained for 42 epochs with a learning rate of 0.001 using a Nadam optimizer. The three embedding layers of the BiLSTM-CNN model had an output dimension of 100, 8, and 30, respectively, which were concatenated into a dimension of 138, to be fed into the BiLSTM layer. The BiLSTM layer had an output dimension of 400, and the MLP layers output tag probabilities with a dimension of 29 (“B-” and “I-” tags for each of the 14 entity classes and “O” tag).

To evaluate the effectiveness of using deep neural networks on named entity recognition, the performance of the proposed method was compared with a CRF-based method as a baseline. CRF was selected for comparison because it was the dominant traditional machine learning-based model for analyzing sequential data such as text, prior to the advent of deep learning-based methods (see Section 2.1). The baseline CRF method achieved an average precision, recall, and F-1 measure of 76.7%, 66.2%, and 69.8%, respectively, while the proposed BiLSTM-CNN model achieved 91.6%, 88.7%, and 89.9%, respectively, which indicates better named entity recognition performance. The improved performance is likely due to the ability of BiLSTM to better adapt to domain-specific semantics and better capture word dependencies. For example, “flight of stairs” (as a whole) was correctly classified as FAC (facility) using the BiLSTM-CNN model, but was incorrectly broken down into “flight” (ENT, i.e., other entity) and “stairs” (FAC) using the CRF-based method. A comparison of the two methods with respect to each entity class is summarized in Table 6. Improved performance using the proposed method was observed for every entity class, except for DAT (date) class because it appears the least frequently in the training dataset.

To evaluate the impact of the three types of features (i.e., GloVe embedding, word-level features, and character-level features, as per Section 4.2) used in the proposed method, an experiment was conducted to compare the named entity recognition performance with and without these features. As shown in Fig. 7, incorporating these features resulted in a 3.1% increase in precision, 4.5% in recall, and 3.3% in F-1 measure, which indicates that the three types of features are effective in improving the named entity recognition performance.

5.2.2 Error Analysis

Fig. 8 shows the confusion matrix for named entity recognition (excluding results of “O” tags which are irrelevant). The most frequent misclassification is seen in predicting an entity as EQU (equipment). For example, 13 ATT (other attribute) entities were recognized as EQU (total extracted ATT = 146) and 12 ENT (other entities) were recognized as EQU (total extracted ENT = 237). This is mainly due to the imbalance of the dataset. As shown in Fig. 5, EQU constitutes a large portion of the dataset.

Misclassification can be seen for other entity classes such as LOC (location), PRO (property), ATT (other attribute), and ENT (other entity), as shown in Fig. 8. A major cause for those misclassifications can be word-sense ambiguities, especially if a single document uses two (or more) different meanings for the same word at different instances, which makes it difficult for the model to distinguish which meaning is intended at which case. For example, the word “standard” can be regarded as REF in the phrase “anchorage standard”, since it refers to a subset of regulations whose topic is anchorage, but as ATT in the phrase “at least a standard 7 inch steel I-beam”, since it means that such measurement of the I-beam is typical. Similarly, the word “level” can be regarded as LOC in the phase “lower level”, since it means some surface or position, but as ATT in the sentence “footings should be level, sound and rigid”, since it means that the footings should have no slopes or bumps. The word “direction” can be regarded as PRO in the phrase “in an upward direction”, since it means a path for movement, but as ENT in the phrase “under the supervision and direction of a competent person”, since it means guidance or management. It is, thus, rather difficult for the model to distinguish which meaning is applied for which case in the absence of sufficient training samples and/or context information.

ENT is the entity class where most classification errors happen. This is mainly due to the coreference words such as "they", "it", “that”, "those", or "the" plus an adjective. For example, the word “it” is not referring to any specific entity in the sentence “It is infeasible or creates a greater hazard to use these systems”, but can be regarded as ENT in the sentence “Each platform greater than 10 feet shall… unless it is designed so…”, which can cause confusion for the model to make correct predictions. Moreover, a few entities in the ENT class do not appear as frequently as necessary for the model to make correct predictions, which caused misclassification errors.

5.3 Coreference Resolution

5.3.1 Optimization and Performance Results

The hyperparameters of the model were fine-tuned. During optimization, the model used a kernel size of 3, a dropout rate of 0.4, and a learning rate of 0.001 using Adam optimizer. The embedding layers had an output dimension of 350. The BiLSTM layer had an output dimension of 400. The weighted vectors produced by the attention mechanism were then converted into headword vectors of dimension 450. Each span representation was obtained by concatenating two vectors at the boundary of the span from the BiLSTM layer (of size 400) with a headword vector (of size 450), whose final dimension was 1250. Such span representation was then used to calculate the pairwise scores whose dimension equals to the number of maximum possible antecedents.

To evaluate the impact of different transfer learning strategies on the coreference resolution model, the model was trained and tested using two strategies: model-based two-stage training and model-based alternating training (following the method by Zhang and El-Gohary 2021a). During the two-stage training, the model was first trained on the source-domain data, then trained further on the target-domain data with the last layer (source output layer) replaced by a target output layer, whereas during alternating training, the model was alternating between training on the source-domain data using the source output layer and training on the target-domain data using the target output layer (for more details on the two training strategies, the readers are referred to Zhang and El-Gohary 2021a). As shown in Fig.

9(a), the proposed coreference resolution method that utilizes the model-based alternating training strategy achieved a B^3 precision, B^3 recall, and B^3 F-1 measure of 77.6%, 70.3%, and 73.8%, respectively, while the same method that utilizes the model-based two-stage training strategy instead achieved a B^3 precision, B^3 recall, and B^3 F-1 measure of 60.7%, 75.5%, and 67.3%, respectively. The achieved performance (with alternating strategy) is comparable with the state-of-the-art LSTM-based coreference resolution performance using general-domain text (69.9%, 64.7%, and 67.2%, see Section 2.2) (Lee et al. (2017), which indicates that the proposed coreference solution method is effective in addressing domain-specific referential ambiguities. Similar to the findings in Section 5.2, incorporating the three types of features also improved the performance of coreference resolution. As shown in Fig. 9(b), an increase of 6.6% in precision, 7.7% in recall, and 7.3% in F-1 measure was observed for the model that uses an alternating strategy.

5.3.2 Error Analysis

One major source of error is that the proposed deep learning model can only consider a fixed span width *L*. Spans that are longer than the fixed width are pruned to reduce computational complexity. However, different from general-domain text where coreferents occur relatively close to each other, due to the hierarchical structure of OSHA regulations, coreferents can occur across a longer span than the fixed span width *L*. For example, if a clause has several subclauses, then the coreferent in the last subclause can be far away from the first coreferent. Therefore, the model could fail to capture all coreferents across spans that are longer than the fixed span width *L*.

The most difficult type of coreferent for the proposed model to recognize is the discontinuous sets where the pronoun refers to more than one antecedent. For example, in the sentence "lifelines, lanyards, and deceleration devices should be … as they would be …", the pronoun "they" refers to three entities: lifelines, lanyards, and deceleration devices. However, the proposed model predicted only one of the three entities as the coreferent, instead of all of them. There are multiple causes for this error. First, there is no enough training data for this case because it is not considered in the CoNLL-2012 dataset. Second, since the three coreferents are all in plural forms, none of the number agreement constraint or verb agreement constraint can work in this case. Third, the proposed deep learning model is designed to output coreferents with the highest probability, and therefore cannot output more than one coreferent.

The lack of domain-specific context is also causing difficulty for the model to decide whether two mentions are coreferents, referring expressions, or not. For example, phrases of "safety monitor", "competent person", and "the person making the determination and certification" are the same expressions by the meaning they convey. However, it is difficult for the model to understand their interconnections and therefore predict them as referring expressions. This is because no background knowledge is provided in the relevant OSHA clauses in terms of role definitions or how construction teams are organized. To solve problems such as this, more domain-specific prior knowledge needs to be incorporated in future work.

5.4 Named Entity Normalization

For the list of candidate identifier names, a total of 1246 candidate identifier names (e.g., “body_belt”, “personal_fall_arrest_system”, and “positioning_device_system”, as per Fig. 2) were collected from the aforementioned OSHA sections (see Section 5.1), which covers most of the fall protection-related entities that could have referential ambiguities. For similarity assessment, example results are shown in Fig. 10 (with an embedding size of 2 for visualization purpose). The figure shows that the vectorized representation has the ability to encode the semantics of the domain-specific entities – entity names that are similar in meaning are closer in their embedding space. For example, “rope” and “pendant_rope” are closer than “rope” and “walkway”. After identifier name mapping, this normalization process achieved an average precision, recall, and F-1 measure of 93.0%, 93.0%, and 93.0%, indicating good normalization performance.

The proposed method for named entity normalization was also helpful in correcting some errors in named entity recognition (i.e., in resolving referential ambiguities). For example, the phrases “edge of the walking/working surface”, “the working edge”, and “the edge” (all in one cluster) were initially recognized as LOC (correct), LOC (correct), and ENT (incorrect), respectively. In this normalization step, all three phrases were however mapped to the identifier name “walking_working_edge”. As a result, the tag of “the edge” was corrected to LOC (tag of “walking_working_edge”). More example results of the proposed named entity normalization method are shown in Table 7.

5.5 Overall Information Extraction

The overall information extraction performance was evaluated by comparing the final extraction results (after named entity recognition, coreference resolution, and named entity normalization) with the final gold standard (see Section 5.1). As shown in Table 8, a macro precision, recall, and F-1 measure of 93.2%, 89.6%, and 91.1% was achieved, which indicates that the proposed named entity recognition method is effective in extracting named entities from construction safety regulations, and that the coreference resolution method and the named entity normalization method are effective in addressing referential ambiguities in the text. To provide a comparative benchmark, recent efforts (outside the construction domain), which covered both named entity recognition and resolution of referential ambiguities, showed comparable performance results [e.g., 83% precision, 87% recall, and 85% F-1 measure in Agrawal et al. (2022) from the medical domain]. An example of the overall information extraction results is illustrated in Fig. 11.

An additional experiment was conducted to further evaluate the effectiveness of the proposed information extraction method (including the three submethods) in resolving referential ambiguities – i.e., the results with and without resolving referential ambiguities were compared. The named entity recognition results from Section 5.2 (i.e., without resolving referential ambiguities) were compared with the final gold standard (which has the coreferents and referring expressions normalized), as per Table 8. The results showed 5.1% (93.2% vs. 88.1%), 10.9% (89.6% vs. 78.7%), and 8.9% (91.1% vs. 82.2%) improvement in precision, recall, and F-1 measure, respectively, which indicates the effectiveness of the proposed approach in addressing referential ambiguities.

6 Limitations

Five limitations of this research are acknowledged. First, the proposed coreference resolution method is limited in recognizing cataphoras. This is because the proposed deep learning model only considers spans that happen before the current one for identifying potential coreferents. To address cataphoras, the authors plan to further modify the proposed deep learning model, in their future work, to allow for considerations of the spans that happen ahead. Second, this study proposed BiLSTM-based models for named entity recognition and coreference resolution, without testing on other deep learning model structures. Although the BiLSTM-based models are suitable for addressing sequential data, validating the proposed information extraction method, and evaluating the impacts of the features and strategies, they fall short in dealing with long sentences. In future work, the authors plan to explore different types of models, such as transformer-based models, to deal with variable sentence lengths for achieving better extraction performance. Third, like any other dictionary- or ontology-based method, the performance of the proposed named entity normalization method depends on the quality of the selected domain-specific dictionary. As the proposed method is applied to a different regulation or subdomain of safety knowledge, a different dictionary/sections would be used, and hence the performance results may vary, although a similar performance level is expected if the quality/coverage of the dictionary/sections are similar. In their future work, the authors will test the proposed method in extracting information from other documents to further verify if a similar level of performance can be achieved in such case. Fourth, the scope of this study was limited to referential ambiguities; word-sense ambiguities, which are a major source of error during named entity recognition, were not addressed. To further improve the extraction performance, the authors plan to conduct word-sense disambiguation, in future work, to determine which sense a word is being used in a particular situation. Fifth, the proposed method was only tested on one source of construction safety regulations – OSHA. In future work, the authors plan to test the proposed method using more construction safety regulatory documents such as the fall-related standards from the American National Standards Institute (ANSI).

7 Contributions to the Body of Knowledge

This research offers a deep learning-based information extraction method for automatically extracting named entities that describe fall protection requirements from construction safety regulations for supporting automated field compliance checking and resolving referential ambiguities. The proposed method improves the information extraction methodology and application in the construction domain in five primary ways. First, to the best of the authors' knowledge, it is the first effort in the construction domain to address referential ambiguities through coreference resolution and named entity normalization. Resolving referential ambiguities is important to maintain consistent expressions in the extracted entities and prevent/reduce referential errors in the information extraction, which could cause serious errors in the compliance checking results. Second, the proposed method uses a combination of three types of features – GloVe embedding, word-level features, and character-level features – to improve the feature representation of the text. Complementing the GloVe embedding features with the proposed word-level and character-level features helps better capture the syntactic differences across different entities, which helps better differentiate these entities. Third, the proposed method leverages both source- and target-domain data, using transfer learning, for enhanced coreference resolution. The use of transfer learning allows the model to leverage the rich syntactic and semantic text patterns from the large-scale general-domain data while adapting these patterns to be closer to those found in the domain-specific text, thereby improving both performance and scalability. Fourth, this study identified a set of entity classes, different types of coreferents, and different types of referring expressions, which were effective in extracting fall protection-related information from OSHA and can be applied in other applications (e.g., extracting information related to other OSHA topics or from other regulations). This new knowledge can bring additional insights to better understand the types of referential ambiguities in this domain-specific text (construction domain or more specifically construction safety domain), and their different syntactic and semantic forms, and how to best resolve these ambiguities for improved document analytics and related artificial intelligence (AI) applications. Fifth, the application of the proposed method could provide a better understanding of construction safety requirements and potential fall protection accidents, and how to improve field compliance and prevent such accidents, considering different contexts and scenarios. The proposed deep learning models for named entity recognition, coreference resolution, and named entity normalization – including their features and transfer learning strategies – could also be applied to more accident types such as struck-by and caught-in-between, or to different regulatory documents such as company-level construction safety rules and reports.

8 Conclusions and Future Work

This paper proposed a deep learning-based information extraction method to extract named entities from construction safety regulations for supporting automated field compliance checking and to resolve referential ambiguities in the extracted results. The proposed information extraction method consists of three submethods: named entity recognition, coreference resolution, and named entity normalization. For named entity recognition, a BiLSTM-CNN model was proposed and trained to recognize and classify named entities from unstructured text. Three types of features (GloVe embedding, word-level features, and character-level features) were used in the proposed method for improving the extraction performance. For coreference resolution, a deep learning-based model was proposed and trained, using transfer learning strategy to leverage the rich semantics from the source-domain data. Two types of training data, including the CoNLL-2012 dataset from the computational linguistic domain and documents from construction safety regulations, were prepared for transfer learning. For named entity normalization, the proposed method mapped different mentions of an entity to its identifier name by considering the similarity between them.

The proposed method was tested on 20 OSHA sections related to fall protection. The proposed named entity recognition method achieved an average precision, recall, and F-1 measure of 91.6%, 88.7%, and 89.9%, respectively, showing better performance than the baseline CRF-based method. The proposed coreference resolution method using model-based alternating training strategy achieved an average B^3 precision, B^3 recall, and B^3 F-1 measure of 77.6%, 70.3%, and 73.8%, respectively, which indicates good coreference resolution performance given such complex text. The proposed named entity normalization method achieved an average precision, recall, and F-1 measure of 93.0%, 93.0%, and 93.0%, indicating its effectiveness in mapping different mentions to their corresponding identifier names. The performance of named entity recognition and coreference resolution with the three sets of features also outperformed those without these features, which indicates that the proposed features have a positive impact on the extraction performance. The proposed information extraction method achieved an overall precision, recall, and F-1 measure of 93.2%, 89.6%, and 91.1%, respectively, which indicates that the proposed coreference resolution method and named entity normalization method were effective in addressing the referential ambiguities in the domain-specific text.

In their future work, the authors plan to focus their research efforts on four main research directions. First, different deep learning model structures or algorithms could be explored to further enhance the extraction performance. For example, multiple transformer-based models that are able to deal with variable sentence lengths, especially Bidirectional Encoder Representations from Transformers (BERT), can be adapted and trained for coreference resolution. Second, the proposed information extraction method (including all three submethods) could be further tested in extracting requirements from other construction safety regulations. Additional adaptation effort may be needed depending on the experimental results and the variability in the text characteristics. Third, the authors will further adapt the proposed method to extract requirements about other types of accidents. Such application could provide additional insights into the generalizability of the proposed method and could help identify ways to improve the field compliance checking process. Fourth, beyond extracting named entities, the authors will develop a relation extraction method, to further add the needed interlinks to the named entities extracted in this work. These interlinks are important in determining compliance, because they would help describe the requirements in terms of the spatial relations between site objects, interactions of workers and their environment, and comparisons of the site objects' attributes with certain values. With the help of such interlinks, the extracted safety requirements can be represented in a structured way for supporting subsequent analytics. All the aforementioned efforts would lead to a better understanding of how to automatically analyze construction safety regulations and how to advance the underlying models for supporting automated and AI-based field compliance checking processes.

9 Data Availability Statement

Some data, models, or code generated or used during the study (the labeled gold standard for evaluation) are available from the corresponding author by request.

10 Acknowledgements

The authors would like to thank the National Science Foundation (NSF). This paper is based on work supported by NSF under Grant No. 1827733. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

11 References

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. (2022). “Large Language Models are Zero-Shot Clinical Information Extractors.” *arXiv preprint arXiv*:2205.12689.

Artstein, R. (2017). “Inter-annotator agreement.” *Handbook of linguistic annotation* (pp. 297-313). Springer, Dordrecht.

Bagga, A., & Baldwin, B. (1998). “Algorithms for scoring coreference chains.” *Proc., 1st Lang. Resour. Eval.*, 563-566.

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1998). “Nymble: a high-performance learning name-finder.” arXiv preprint cmp-lg/9803003.

Chen, H., Luo, X., Zheng, Z., & Ke, J. (2019). “A proactive workers' safety risk evaluation framework based on position and posture data fusion.” *Automat. Constr.*, *98*, 275-288.

Chi, C. F., & Lin, S. Z. (2018). “Classification scheme and prevention measures for caught-in-between occupational fatalities.” *Appl. Ergon.*, 68, 338-348.

Chiu, J. P., & Nichols, E. (2016). “Named entity recognition with bidirectional LSTM-CNNs.” *Trans. Assoc. Comput. Linguist.*, *4*, 357-370.

Cho, H., Choi, W., & Lee, H. (2017). “A method for named entity normalization in biomedical articles: application to diseases and plants.” *BMC bioinformatics*, 18(1), 1-12.

Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). “A survey of deep learning and its applications: a new paradigm to machine learning.” *Arch. Comput. Methods Eng*., 27(4), 1071-1092.

Dong, X. S., Largay, J. A., Choi, S. D., Wang, X., Cain, C. T., & Romano, N. (2017). “Fatal falls and PFAS use in the construction industry: Findings from the NIOSH FACE reports.” *Accident Analysis & Prevention*, *102*, 136-143.

Fakhraei, S., Mathew, J., & Ambite, J. L. (2019). “Nseen: Neural semantic embedding for entity normalization.” *In ECML PKDD* (pp. 665-680). Springer, Cham.

Fang, L., Cao, Y., & Zheng, Z. (2021). “Biomedical Entity Normalization based on Pre-trained Model with Enhanced Information.”

Fang, W., Ding, L., Luo, H., & Love, P. E. (2018). “Falls from heights: A computer vision-based approach for safety harness detection.” *Automat. Constr*., 91, 53-61.

Fang, W., Ma, L., Love, P. E., Luo, H., Ding, L., & Zhou, A. (2020). “Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology." *Automat. Constr.*, *119*, 103310.

Farouk, M. (2020). "Measuring text similarity based on structure and word embedding." *Cogn. Syst. Res*., 63, 1-10.

Isozaki, H., & Kazawa, H. (2002). "Efficient support vector classifiers for named entity recognition." *Proc., 19th Int. Conf. Comput. Linguist*., Taipei, Taiwan 2002, 1–7.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). "Machine learning and deep learning." *Electronic Markets*, 31(3), 685-695.

Jijkoun, V., Khalid, M. A., Marx, M., & De Rijke, M. (2008). "Named entity normalization in user generated content." *Proc., 2nd workshop on Analytics for noisy unstructured text data* (pp. 23-30).

Kim, T., & Chi, S. (2019). "Accident case retrieval and analyses: using natural language processing in the construction industry." *J. Constr. Eng. Manag*., 145(3), 04019004.

Labor (2021a). "Employer-reported workplace injuries and illnesses-2020." *Bureau of Labor Statistics*. Retrieved from <https://www.bls.gov/news.release/pdf/osh.pdf> (Jan. 20, 2022).

Labor (2021b). "Census of Fatal Occupational Injuries (CFOI) – Current." *Bureau of Labor Statistics*. Retrieved from <https://www.bls.gov/iif/oshcfoi1.htm> (Jan. 20, 2022).

Labor (2022). "Labor Force Statistics from the Current Population Survey." *Bureau of Labor Statistics*. Retrieved from <https://www.bls.gov/cps/cpsaat47.htm> (Jan. 20, 2022).

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." *Proc., 18th Int. Conf. Mach. Learn*., Williamstown, Massachusetts 2001, 282–289.

Leaman, R., Wei, C. H., & Lu, Z. (2015). "tmChem: a high performance approach for chemical named entity recognition and normalization." *J. Cheminformatics*, 7(1), 1-10.

Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). "End-to-end neural coreference resolution." *arXiv preprint arXiv:1707.07045*.

Liu, K., & El-Gohary, N. (2017). "Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports." *Automat. Constr.*, *81*, 313-327.

Liu, K., & El-Gohary, N. (2018). "Unsupervised named entity normalization for supporting information fusion for big bridge data analytics." *Proc., the European Group for Intelligent Computing in Engineering* (pp. 130-149). Springer, Cham.

Lu, Y., Li, Q., Zhou, Z., & Deng, Y. (2015). "Ontology-based knowledge modeling for automated construction safety checking." *Saf. Sci.*, *79*, 11-18.

Mutual, L. (2020). "Workplace safety index 2020: Construction." *Liberty Mutual*. Retrieved from <https://business.libertymutual.com/wp-content/uploads/2021/04/WSI_1002.pdf> (Jan. 20, 2022).

Nadeau, D., Turney, P. D., & Matwin, S. (2006). "Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity." *Proc., bienn. conf. Can. Soc. Comput. Stud. Intell.* (pp. 266-277). Springer, Berlin, Heidelberg.

National Safety Council (2020). "Work Injury Costs." *National Safety Council*. Retrieved from < https://injuryfacts.nsc.org/work/costs/work-injury-costs/> (Jan. 20, 2022).

Nepal, M. P., Staub-French, S., Pottinger, R., & Zhang, J. (2013). "Ontology-based feature modeling for construction information extraction from a building information model." *J. Comput. Civ. Eng*., 27(5), 555-569.

Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., & Zweigenbaum, P. (2015). "CLEF eHealth Evaluation Lab 2015 Task 1b: Clinical Named Entity Recognition." *In CLEF* (Working Notes).

OSHA (2020a). "Commonly Used Statistics." *Occupational Safety and Health Administration*. Retrieved from <https://www.osha.gov/data/commonstats> (Jan. 20, 2022).

OSHA (2020b). "Fall Prevention - General Statistics Related to Slips, Trips, & Falls." *Occupational Safety and Health Administration*. Retrieved from <https://www.oshatraining.com/fall-protection-and-prevention-training.php> (Jan. 20, 2022).

OSHA, U. (2020c). "Construction Industry: OSHA Safety and Health Standards (29 CFR 1926/1910)." *US Department of Labor, Occupational Safety and Health Administration*, Washington, DC.

Pennington, J., Socher, R., & Manning, C. (2014, October). "Glove: Global vectors for word representation." *Proc., Empirical Methods in Natural Language Processing (EMNLP) Conf.,* 1532-1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). "Deep contextualized word representations." *arXiv preprint arXiv:1802.05365*.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012, July). "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes." *In Joint Conference on EMNLP and CoNLL-Shared Task*, 1-40.

Ren, R., & Zhang, J. (2021). "Semantic Rule-Based Construction Procedural Information Extraction to Guide Jobsite Sensing and Monitoring." *J. Comput. Civ. Eng*., 35(6), 04021026.

Rahman, A., & Ng, V. (2009). “Supervised models for coreference resolution.” *Proc., Empirical Methods in Natural Language Processing (EMNLP) Conf.,* 968-977.

Roy, D., Ganguly, D., Bhatia, S., Bedathur, S., & Mitra, M. (2018). “Using word embeddings for information retrieval: How collection and term normalization choices affect performance.” *Proc., 27th Int. Conf. Inf. Knowl. Manag. Proc.* (pp. 1835-1838).

Seo, J., Han, S., Lee, S., & Kim, H. (2015). “Computer vision techniques for construction safety and health monitoring.” *Adv. Eng. Inform., 29*(2), 239-251.

Sitikhu, P., Pahi, K., Thapa, P., & Shakya, S. (2019). “A comparison of semantic similarity methods for maximum human interpretability.” *Proc., artificial intelligence for transforming business and society (AITB)* (Vol. 1, pp. 1-4). IEEE.

Tadesse, S., Kelaye, T., & Assefa, Y. (2016). “Utilization of personal protective equipment and associated factors among textile factory workers at Hawassa Town, Southern Ethiopia.” *J. Occup. Med. Toxicol.*, *11*(1), 6.

Tang, S., Roberts, D., & Golparvar-Fard, M. (2020). “Human-object interaction recognition for automatic construction site safety inspection.” *Automat. Constr.*, *120*, 103356.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). “Attention is all you need.” *Proc. Adv. Neural Inf. Process Syst.,* 5998-6008.

Yadav, V., & Bethard, S. (2019). “A survey on recent advances in named entity recognition from deep learning models.” *arXiv preprint arXiv:1910.11470*.

Yenkar, P., & Sawarkar, S. D. (2021). “Gazetteer based unsupervised learning approach for location extraction from complaint tweets.” *IOP Conf. Ser.: Mater. Sci. Eng*. (Vol. 1049, No. 1, p. 012009). IOP Publishing.

Yuan, Z., Zhao, Z., Sun, H., Li, J., Wang, F., & Yu, S. (2022). “CODER: Knowledge-infused cross-lingual medical term embedding for term normalization.” *J. Biomed. Inform*., 103983.

Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2019). “Construction site accident analysis using text mining and natural language processing techniques.” *Automat. Constr*, 99, 238-248.

Zhang, J., & El-Gohary, N. M. (2013). “Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking.” *J. Comput. Civ. Eng.*, 30(2), 04015014.

Zhang, R., and El-Gohary, N. (2021a). “A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking.” *Automat. Constr.*, in press.

Zhang, R., and El-Gohary, N. (2021b). "Hierarchical representation and deep learning-based method for automatically transforming textual building codes into semantic computable requirements." *J. Comput. Civ. Eng*., accepted.

Zhang, S., Boukamp, F., & Teizer, J. (2014). "Ontology-based semantic modeling of safety management knowledge." In *Computing in Civil and Building Engineering,* 2254-2262.

Zhang, S., Boukamp, F., & Teizer, J. (2015). "Ontology-based semantic modeling of construction safety knowledge: Towards automated safety planning for job hazard analysis (JHA)." *Automat. Constr.*, *52*, 29-41.

Zhong, B., Li, H., Luo, H., Zhou, J., Fang, W., & Xing, X. (2020a). "Ontology-based semantic modeling of knowledge in construction: classification and identification of hazards implied in images." *J. Constr. Eng. Manag.*, *146*(4), 04020013.

Zhong, B., Pan, X., Love, P. E., Ding, L., & Fang, W. (2020b). "Deep learning and network analysis: Classifying and visualizing accident narratives in construction." *Automat. Constr.*, 113, 103089.

Zhong, B., Wu, H., Xiang, R., & Guo, J. (2022). "Automatic Information Extraction from Construction Quality Inspection Regulations: A Knowledge Pattern–Based Ontological Method." *J. Constr. Eng. Manag*., 148(3), 04021207.

Zhou, H., Ning, S., Liu, Z., Lang, C., Liu, Z., & Lei, B. (2020). "Knowledge-enhanced biomedical named entity recognition and normalization: application to proteins and genes." *BMC bioinformatics*, 21(1), 1-15.

Zhou, P., & El-Gohary, N. (2017). "Ontology-based automated information extraction from building energy conservation codes." *Automat. Constr.*, *74*, 103-117.

Zhou, W., Yu, C., Smalheiser, N., Torvik, V., & Hong, J. (2007). "Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature." *In SIGIR* (pp. 655-662).

Table 1. Examples of most frequent entities in each pre-defined class.

Class	Examples
Person	Employer, employee, competent person, qualified person, engineer
Equipment	Scaffold, ladder, safety net system, lanyard
Reference	This section, paragraph (k) of 1926.502, paragraphs (a), (b), or (c)
Hazard	Falling, damage, snag of clothing, tripping, shake
Facility	Beam, roof, foundation, metal decking, other structure
Location	Walking/working surface, work area, level
Operation	Hoisting, dropping, rigging, lifting
Material	Reinforcing steel, wood, metal, debris
Property	Weight, length, diameter, direction, speed
Date	January 1, 1998, Apr. 6
Other attribute	In use, center-to-center, under construction
Quantity value	6, 3.3, ½, half, one
Quantity unit	Feet, times, meters, inches, mps
Other entity	Fall protection plan, requirement, test

Table 2. Examples of different types of coreferents.

Type of coreferent	Description	Example
Demonstrative pronoun	Comparison exists with something that occurred earlier.	In the sentence “*scaffold strength*(1) is less than *that*(2) required by paragraph (a) of this section”, the anaphor (2) refers to (1).
Discontinuous set	Pronoun refers to more than one antecedent.	In the sentence “*lifelines*(1), *lanyards*(2), and *deceleration devices*(3) should be … as *they*(4) would be …”, pronoun (4) refers to (1), (2), and (3) together as a single entity.
One anaphora	Pronoun “one” refers to antecedent.	In the sentence “that *component*(1) should be replaced by a stronger *one*(2)”, anaphora (2) refers back to (1).
Definite pronominal	Pronoun “it” refers to antecedent.	In the sentence “*sliding hitch knot*(1) should never be used because *it*(2) is unreliable in stopping a fall”, pronoun (2) refers back to (1).
Adjectival pronominal	Coreferent refers to adjective form of entity that occurred earlier.	In the sentence “under *the walking/working surface*(1), but in no case more than 30 feet below *such level*(2)”, (1) is an adjectival form that has been referred to by (2).
Cataphora	Opposite of anaphora.	In the sentence “only *those items*(1) specifically designed as *counterweights*(2)”, (1) refers to (2) that precedes it.
Inferable or bridging anaphora	Coreferent refers to the antecedent ambiguously through context.	In the sentence “*the person making the determination and certification*(1)… until inspected and determined by a *competent person*(2) shall be removed from service”, (2) refers to (1) though not stated explicitly, but can be inferred by their context.

Table 3. Examples of different types of referring expressions.

Type of referring expression	Example of two expressions referring to same entity
Possessive	“*edge of roof*” and “*roof edge*”
Article	“*edge of the higher level*” and “*edge of a higher level*”
Hyphen	“*double-pole scaffold*” and “*double pole scaffold*”
Synonym	“*metatarsal protection*” and “*metatarsal guards*”
Space	“*eye splice*” and “*eyesplice*”
Active-passive voice	“*complete system*” and “*completed system*”
Detailed description omitted	“*two-point adjustable suspension scaffold*” and “*two-point scaffold*”
Abbreviation	“*controlled decking zone*” and “*CDZ*”

Table 4. Selected OSHA sections.

Topic	Section(s)
General requirement	1926.451, 1926.501, 1926.1051
Fall protection systems	1926.502, 1926.760, 1926.1423, 1926 Subpart R App G
Guardrail systems	1926 Subpart M App B
Personal fall arrest systems	1926 Subpart M App C
Positioning device systems	1926.104, 1926.105, 1926 Subpart M App D
Personal protective equipment	1926.95, 1926.96, 1926.100
Scaffolds	1926.452, 1926 Subpart L App A
Ladders	1926.1053
Aerial lifts	1926.453
Housekeeping	1926.25

Table 5. Example clause annotated using BIO tagging scheme.

Original sentence	Annotated sentence
Each employee on a walking/working surface shall be protected from objects falling through holes by covers .	<O>Each</O> <B-PER>employee</B-PER> <O>on</O> <O>a</O> <B-LOC>walking/working</B-LOC> <I-LOC>surface</I-LOC> <O>shall</O> <O>be</O> <O>protected</O> <O>from</O> <B-ENT>objects</B-ENT> <O>falling</O> <O>through</O> <B-ENT>holes</B-ENT> <O>by</O> <B-EQU>covers</B-EQU> <O>.</O>

Note: "B"=beginning of an entity; "I"=inside of an entity; "O"=absence of an entity; PER=person; EQU=equipment; LOC= location; and ENT=other entity.

Table 6. Performance of proposed named entity recognition method compared to baseline.

Class	CRF (baseline)			BiLSTM-CNN		
	Precision	Recall	F-1 measure	Precision	Recall	F-1 measure
PER	93.3%	75.7%	83.6%	95.5%	95.5%	95.5%
EQU	65.9%	88.9%	75.7%	92.9%	96.8%	94.8%
REF	74.2%	85.2%	79.3%	84.5%	97.3%	90.4%
HAZ	76.9%	43.5%	55.6%	91.4%	78.0%	84.2%
FAC	75.9%	47.3%	58.3%	88.4%	92.4%	90.4%
LOC	68.7%	61.2%	64.7%	94.6%	91.8%	93.2%
OPE	46.5%	29.9%	36.4%	84.7%	81.3%	83.0%
MAT	62.5%	30.6%	41.1%	97.3%	92.3%	94.7%
PRO	77.0%	73.8%	75.4%	86.5%	92.8%	89.5%
DAT	100.0%	100.0%	100.0%	100.0%	80.0%	88.9%
ATT	80.1%	47.6%	59.7%	85.6%	64.4%	73.5%
QUA	99.5%	92.2%	95.7%	99.0%	97.6%	98.3%
UNI	94.3%	94.3%	94.3%	96.2%	98.1%	97.1%
ENT	59.4%	56.6%	58.0%	85.4%	84.1%	84.7%
Macro average	76.7%	66.2%	69.8%	91.6%	88.7%	89.9%

Note: PER=person; EQU=equipment; REF=reference; HAZ=hazard; FAC=facility; LOC=location; OPE=operation; MAT=material; PRO=property; DAT=date; ATT=other attribute; QUA=quantity value; UNI=quantity unit; and ENT=other entity.

Table 7. Examples of named entity normalization results.

Different mentions	Entity name after normalization	Entity class tag after normalization
walking/working level, walking/working surface, work surface, such level, work platform	walking_working_surface	LOC
supporting formwork, supporting surface, their member, supporting structure	supporting_structure	FAC
the member to which they are connected, connected object, itself, object	connected_member	ENT
registered professional engineer, safety monitor, the person making the determination and certification, competent person	competent_person	PER
qualified person with appropriate education and experience, qualified person	qualified_person	PER
personal protective equipment, such equipment, PPE	personal_protective_equipment	EQU
two-point scaffolds, scaffold, two-point adjustable suspension scaffold	two_point_adjustable_suspension_scaffold	EQU

Table 8. Comparisons of the proposed method before and after resolving referential ambiguities.

Gold standard for comparison	Proposed Method											
	NER (%)			CR (%)			NEN (%)			NER + CR + NEN (%)		
	P	R	F-1	B^3 P	B^3 R	B^3 F-1	P	R	F-1	P	R	F-1
Named entity recognition (NER) gold standard	91.6	88.7	89.9	NA	NA	NA	NA	NA	NA	NA	NA	NA
Coreference resolution (CR) gold standard	NA	NA	NA	77.6	70.3	73.8	NA	NA	NA	NA	NA	NA
Named entity normalization (NEN) gold standard	NA	NA	NA	NA	NA	NA	93.0	93.0	93.0	NA	NA	NA
Final gold standard (after resolving referential ambiguities)	88.1	78.7	82.2	NA	NA	NA	NA	NA	NA	93.2	89.6	91.1

Note: P=precision; R=recall; F-1=F-1 measure; B^3 P=B^3 precision; B^3 R=B^3 recall; B^3 F-1=B^3 F-1 measure; NA=not applicable.

Fig. 1. Research methodology.

Fig. 2. Application of proposed information extraction method, with example.

Fig. 3. Architecture of proposed BiLSTM-CNN model for named entity recognition.

Fig. 4. Architecture of proposed deep learning model for coreference resolution.

Fig. 5. Distribution of entity classes.

Fig. 6. Examples of annotated clauses for coreference resolution: (a) Example clause annotated by following the tagging scheme of the CoNLL-2012 dataset; and (b) Coreferents in one example clause and its subclause.

Fig. 7. Performance of proposed named entity recognition method with and without the three types of features.

Fig. 8. Confusion matrix for proposed named entity recognition method.

Fig. 9. Performance of proposed coreference resolution method: (a) Comparison of different transfer learning strategies; and (b) Comparison with and without the three types of features.

Fig. 10. Similarity of entity names in the embedding space.

Fig. 11. Example of extracted semantic information elements using the proposed information extraction method.