Deep Learning-Based Sequence Labeling for Information Extraction from Multiple Types of Textual Bridge Reports

Qiyang Chen, S.M. ASCE,[1] and Nora EI-Gohary, A.M. ASCE[2]

[1]PhD. Student, Dept. of Civil and Environmental Engineering, Univ. of Illinois at Urbana-Champaign, 205 North Mathews Ave., Urbana, IL 61820. E-mail: qiyangc2@illinois.edu
[2]Associate Professor, Dept. of Civil and Environmental Engineering, Univ. of Illinois at Urbana-Champaign, 205 North Mathews Ave., Urbana, IL 61820. E-mail: gohary@illinois.edu

ABSTRACT

A massive amount of data/information that is buried in unstructured data sources such as bridge rehabilitation reports offer opportunities to complement other data sources such as structural health monitoring data for improved prediction of future bridge conditions. Extracting these data/information from their sources is, however, challenging for two reasons: (1) utilizing textual data for analytics remains to be a challenge due to the inherently unstructured nature of these data; and (2) existing information extraction methods, which aim to extract structured data/information from text, are limited in their capabilities and generalizability (e.g., a method developed to extract certain information from one type of document may not perform well on another type). To address these gaps, this paper proposes a deep learning-based information extraction (IE) model that automatically recognizes and extracts bridge-related data/information (e.g., bridge deficiencies) from multiple types of unstructured textual sources and represents the extracted data/information in a structured format to support further data analytics. The proposed IE model utilizes deep learning-based (BiLSTM); begin, inside, and outside (BIO) encoding for the phrase-level segmentation; and Conditional Random Fields (CRF) for both word-level and phrase-level labeling for automatic sequence labeling. The paper discusses the proposed model and its performance results.