Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 145 (2016) 504 – 510

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Sustainable Design, Engineering and Construction

Ontology-based sequence labelling for automated information extraction for supporting bridge data analytics

Kaijian Liu[a], Nora El-Gohary[b]*

[a]*Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Ave.,Urbana, IL, 61820, USA*

[b]*Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Ave.,Urbana, IL, 61820, USA*

Abstract

The massive amount of data/information buried in textual bridge inspection reports open opportunities to leverage big data analytics for advanced information-rich bridge deterioration prediction. However, utilizing textual data for bridge deterioration prediction is challenging because of its inherently unstructured nature. To this end, this paper proposes an ontology-based information extraction (IE) framework that automatically recognizes and extracts key data/information from unstructured textual reports, and represents the extracted data/information in a structured way that is ready for data analytics. The proposed IE framework is composed of two primary components: (1) ontology-based sequence labelling for term identification, and (2) ontology-based dependency grammar for relationship association. This paper focuses on presenting the proposed sequence labelling methodology. The methodology utilizes ontology-based begin, inside, and outside (BIO) encoding for phrase-level segmentation and Conditional Random Field (CRF) for ontology-based labelling in both token and phrase levels. The experimental results showed that the proposed methodology has a precision of 97% and a recall of 91%.

Peer-review under responsibility of the organizing committee of ICSDEC 2016

Keywords: Ontology; Sequence labelling; Information extraction; Infrastructure system data analytics; Bridge deterioration prediction.

* Corresponding author. Tel.: +1-217-333-6620; fax: +1-217-265-8039.
E-mail address: gohary@illinois.edu

Peer-review under responsibility of the organizing committee of ICSDEC 2016
doi:10.1016/j.proeng.2016.04.035

1. Introduction

The aged and deteriorated bridges affect the safety, serviceability, and reliability of the U.S. transportation system. For example, the collapse of the I-35W Mississippi River Bridge alone killed 13 people and injured 145 in 2007 [1]. It has long been recognized that the aged and deteriorated bridges are in need of extensive maintenance, repair, and rehabilitation (MR&R) [2]. It is, however, challenging to make MR&R decisions under the existing stringent funding constraints. A $20.5 billion annual investment in the construction and maintenance of bridges is needed to eliminate the nation's bridge deficient backlog by 2028, while only $12.8 billion is being invested currently [3]. To enhance and optimize decision making, bridge management usually relies largely on the predicted performance of bridges. Reliable and information-rich bridge deterioration prediction is essential to enhance bridge operation and maintenance decision making. Nevertheless, the state-of-the-art bridge deterioration prediction models/techniques are limited in fully supporting enhanced decision making for bridge operation and maintenance. Existing research efforts focused on utilizing abstract database data [e.g., National Bridge Inventory (NBI)] along with a small set of context-aware data (e.g., bridge age and location) to develop deterministic, probabilistic, and/or artificial intelligence (AI) models to predict bridge deterioration in a condition-state-based manner. Despite their achievements, their capabilities in supporting enhanced bridge operation and maintenance decision making are limited, with no models or systems making use of the massive amount of data/information buried in textual bridge inspection reports [4]. These inspection reports include critical data/information about bridge deficiencies and bridge maintenance actions, much beyond what can be found in structured data (e.g., NBI data). Such additional data/information about bridge deficiencies and bridge maintenance actions – including corrosion, cracking, decay, delamination, efflorescence, scaling and spalling, scour, and settlement and their associated maintenance methods and material – are expected to enhance bridge prediction results.

To this end, there is a need for information extraction (IE) methods that can automatically recognize and extract data/information from unstructured textual bridge inspection reports, and represent the extracted data/information in a structured way that is ready for data analytics. However, automated IE from unstructured textual bridge inspection reports is challenging, because (1) the capability of machines is limited in understanding natural language, (2) the bridge inspection reports exhibit domain-specific uniqueness that includes complex term identification and relationship association and largely varied text patterns, and (3) the technical criticality of the extracted data/information requires IE to achieve high performance for both precision and recall.

To address this need, this paper proposes an IE framework for supporting automated extraction and representation of key data/information from unstructured textual bridge inspection reports for facilitating information-rich bridge deterioration prediction. The proposed IE framework is composed of two primary components: (1) ontology-based sequence labelling for term identification, and (2) ontology-based dependency grammar for relationship association. Term identification aims to segment key data/information from background data/information. Relationship association aims to analyze how the segmented data/information are related to each other. A bridge domain ontology is used, at the cornerstone of the proposed IE framework, to help recognize domain terminology and meanings for improving the performance of automated IE [5]. In this paper, the authors focus on introducing the proposed ontology-based sequence labelling methodology and the results of testing the methodology in automated term identification from the I-35W Mississippi River Bridge 2006 inspection report.

2. Background

Information extraction (IE) is defined as the automatic extraction of certain information from natural language text [6]. More specifically, IE automatically processes natural language text to extract information of a particular class of entities, relationships, or events [7]. Research efforts in the IE domain can be classified into two common approaches: a rule-based approach and a machine learning (ML)-based approach [8]. The rule-based IE approach involves the development of pattern-matching-based extraction rules for extracting information of interest, where the text patterns consist of syntactic and/or semantic features. For example, Zhang and El-Gohary [9] manually developed pattern-matching-based rules for extracting regulatory requirements from the International Building Code; Fader *et al.* [10] manually developed syntactic and lexical constraint pattern-matching-based rules to improve open IE performance; Wu and Weld [11] automatically developed pattern-matching-based rules based on Wikipedia