Automation in Construction 81 (2017) 313–327

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports

Kaijian Liu, Nora El-Gohary *

Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 N. Mathews Ave., Urbana, IL 61801, United States

ARTICLE INFO

Article history:
Received 16 July 2016
Received in revised form 20 January 2017
Accepted 26 February 2017
Available online 20 May 2017

Keywords:
Information extraction
Ontology
Semi-supervised machine learning
Conditional random fields
Bridges
Deterioration prediction
Maintenance decision making

ABSTRACT

A large amount of detailed data about bridge conditions and maintenance actions are buried in bridge inspection reports without being used. Information extraction and data analytics open opportunities to leverage this wealth of data for improved bridge deterioration prediction and enhanced maintenance decision making. This paper proposes a novel ontology-based, semi-supervised conditional random fields (CRF)-based information extraction methodology for extracting information entities describing existing deficiencies and performed maintenance actions from bridge inspection reports. The ontology facilitates the analysis of the text based on content and domain-specific meaning. The proposed semi-supervised CRF simultaneously captures the dependency structures as well as the distributions of labeled and unlabeled data in a concave machine-learning function. It learns from a small set of fixed labeled data and, at the same time, dynamically adapts itself to unseen instances by further learning from a large set of unlabeled data for both reduced human effort and high performance. The proposed algorithm achieved an average precision, recall and, F-1 measure of 94.1%, 87.7%, and 90.7%, respectively.

1. Introduction

Bridges are one of the most important components of the U.S. transportation system. However, the nation's bridges are aging with an average age of 42 years [1], and are deteriorating with a functionally-obsolete rate of 14% and a structurally-deficient rate of 10% [2]. The critical conditions of our national bridges pose great threats to public safety and economic well-being. For instance, the collapse of the I-35W Mississippi River Bridge – one of about 600 bridge failures that occurred in the U.S. between 1989 and 2013 – alone killed 13 people and injured 145 in 2007 [3]. 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 [1]. Bridge maintenance, repair, and rehabilitation (MR&R) is the key to maintain and improve the conditions of our national bridges. Bridge MR&R decisions, to a great extent, depend on bridge deterioration prediction [4–7].

However, existing bridge deterioration prediction models and techniques (e.g., [4,6–9]) are limited in supporting well-informed bridge MR&R decision making, because they have focused on utilizing only a limited set of abstract data [e.g., National Bridge Inventory (NBI) data, which describe bridge conditions mainly by condition ratings]. Such abstract data, although very useful, are not sufficient, because they lack detailed descriptions of bridge conditions and maintenance history, which limits the ability for deficiency-based and maintenance action-based predictions and for learning from these data to inform MR&R decision making. On the other hand, a large amount of data – about bridge conditions and maintenance actions – are buried in textual bridge inspection reports and are not utilized. According to the Federal Highway Administration (FHWA), bridge owners are unable to convert the "mountain" of inspection data in bridge inspection reports into effective maintenance decisions [5]. Bridge inspection reports contain a massive amount of technically-detailed data about the types, onset times, and severities of bridge deficiencies, and the methods, costs, durations, and effectiveness levels of bridge maintenance actions. Such critical data about bridge deficiencies and maintenance actions are much beyond what can be found in NBI data, and are thus expected to improve the ability to predict bridge deterioration.

As such, there is a need for information extraction (IE) methods that can automatically recognize and extract data/information – about bridge conditions and maintenance actions – from unstructured textual bridge inspection reports, and represent the extracted data/information in a structured format that is ready for further data analytics. However, automated IE from bridge inspection reports – compared to other IE efforts such as IE from building codes (e.g., Zhang and El-Gohary [10]) and social media text (e.g., Ritter et al. [11]) – is challenging because of two main reasons. First, compared to building codes, bridge inspection reports are highly variable in terms of text characteristics and patterns, because they are typically written by many different writers/inspectors from various local, state, and federal agencies. Existing rule-based (e.g., [12–16]) and supervised machine learning (ML)-based (e.g., [17,18]) IE

* Corresponding author.
E-mail addresses: kliu15@illinois.edu (K. Liu), gohary@illinois.edu (N. El-Gohary).

http://dx.doi.org/10.1016/j.autcon.2017.02.003
0926-5805/

methods would, thus, require an unaffordable amount of human effort for developing a comprehensive set of representative pattern-matching-based rules or annotated training data, in order to capture the variability in text patterns. Second, compared to social media text, on one hand, bridge inspection reports exhibit domain-specific uniqueness that involves high levels of technical detail and complex concept identification and relationship association (i.e., identifying complex technical concepts about bridge elements, deficiencies, and maintenance actions, etc., and their associated relations). On the other hand, because of the technical criticality of the extracted data/information, high performance in both precision and recall is required for automated IE from bridge inspection reports. Existing semi-supervised ML-based (e.g., [19–22]) and unsupervised ML-based (e.g., [23–26]) IE methods cannot deal with such complexities and variabilities with high precision and recall performance.

To address these challenges, this paper proposes an ontology-based, semi-supervised conditional random fields (CRF)-based IE methodology for extracting information about bridge conditions and maintenance actions from bridge inspection reports. The proposed IE methodology utilizes the semantics of a bridge domain ontology to assist in analyzing the text in bridge inspection reports based on content and domain-specific meaning. A novel semi-supervised CRF-based IE methodology is proposed. It allows for (1) simultaneously capturing the dependency structures and distributions of labeled and unlabeled data in a concave machine-learning function, and (2) dynamically adapting itself to unseen instances by further learning from a large set of unlabeled data – in addition to learning from a small set of fixed labeled data. This helps to deal with the complexities and variabilities of the text, with both reduced human effort and high IE performance. The proposed methodology was evaluated in extracting information entities about existing deficiencies, performed maintenance actions, and their related attributes from 11 bridge inspection reports collected from different state Departments of Transportation (DOTs).

2. Background

IE is an automated process that aims to recognize and extract information of a particular class of entities, relations, or events from natural language text [27,28]. Existing IE methods can be classified into two primary categories: rule-based methods and ML-based methods [27–29].

2.1. Rule-based information extraction

Rule-based IE methods rely on hand-crafted pattern-matching-based rules for guiding the recognition and extraction of target information from unstructured textual data [29,30]. The pattern-matching-based rules are constructed with syntactic and/or semantic text features. Outside of the construction domain, many rule-based IE techniques have been proposed (e.g., [12–14,16,31,32]). In the construction domain, a limited number of research efforts have focused on developing rule-based IE methods to support various domain-specific tasks (e.g., [10,33–36]). For example, Zhang and El-Gohary [10] and Zhou and El-Gohary [34] developed pattern-matching-based rules with both syntactic and semantic features to extract building regulatory information for automated compliance checking.

2.2. Machine learning-based information extraction

ML-based IE methods utilize ML algorithms to automate the extraction of information from text [29,30]. ML-based IE methods differ from each other primarily based on the types of ML algorithms used. ML-based IE methods are supervised, semi-supervised, or unsupervised.

2.2.1. Supervised machine learning-based information extraction

Supervised ML-based IE methods learn from labeled training data how to extract information. A number of supervised ML algorithms have been utilized in supporting IE, including decision trees [37], support vector machines [38], structural support vector machines [39], hidden Markov models [40], maximum-entropy Markov models [41], and CRF [42]. Among these IE methods, CRF has been widely recognized for its suitability in supporting IE. This is because it: (1) is a graphical model that offers a natural formalism for representing the dependency structures of natural language [43]; (2) is a discriminative model that captures conditional probabilities to allow for the exploration of a rich set of interdependent features [43]; and (3) models conditional probabilities globally to prevent label bias issues [42].

2.2.2. Semi-supervised machine learning-based information extraction

Semi-supervised ML-based IE methods learn from both labeled and unlabeled data how to extract information. Existing semi-supervised ML-based IE methods have been proposed using bootstrapping strategy (e.g., [19–22]), information-theoretic regularization (e.g., [44,45]), or robust representations of unlabeled data as inputs (e.g., [46,47]). The bootstrapping strategy relies on an iterative process of adding confidently-extracted unlabeled data and re-training a ML model based on the new dataset [21]. It might be, thus, prone to noises and requires heuristic determination of stopping criteria [48]. Information-theoretic regularization aims to regularize learning functions of labeled data through minimizing the entropy of unlabeled data. The regularization process results in a non-concave objective function [44]. Concavity is especially important for ML-based IE; otherwise, the IE performance could be negatively affected by suboptimal initializations and only reaching to local maxima. Robust representations of unlabeled data are achieved under the cluster assumption, which assumes that if two data points lay in the same cluster, they are likely to have a similar class label [45]. Utilizing the cluster assumption has been proven to be an effective way for developing semi-supervised ML-based IE methods [49,50].

2.2.3. Unsupervised machine learning-based information extraction

Unsupervised ML-based IE methods learn how to extract information, without learning from labeled training data. In the absence of labeled data, some unsupervised ML-based IE methods attempted to group similar entities into a cluster merely based on similarities measured from unlabeled text (e.g., [23–26]). Others also proposed to utilize topic modeling methods, such as probabilistic latent semantic indexing [51] and latent Dirichlet allocation [52], in order to dynamically cluster similar entities (e.g., [53]). Because of the existence of statistical dependencies between entities in natural language [43], without formally representing and utilizing such dependencies revealed by labeled data, unsupervised ML-based IE methods might be inclined to generate incoherent clusters [54].

3. State of the art and knowledge gaps

There is a body of research efforts – inside and outside of the construction domain – that have been undertaken towards extracting information from unstructured text. Despite their achievements, existing IE methods are still limited in supporting automated IE from highly heterogeneous and complex text – such as that in bridge inspection reports. Two primary knowledge gaps are identified.

First, there is a lack of IE methods that can simultaneously reduce human effort and achieve high performance when extracting information from highly heterogeneous and complex text. On one hand, most of the existing IE methods have taken a rule-based approach or a supervised ML-based approach. For example, almost all IE efforts in the construction domain have used rule-based IE methods (e.g., [10,33,34,55]). Rule-based and supervised ML-based IE methods might be able to address the heterogeneity and complexity of text and thus achieve high IE performance by learning from a large set of representative examples, but they require a high amount of human effort. This is because such IE methods involve a highly human-intensive process for developing IE rules (in the case of rule-based IE) or annotating training