Automation in Construction 62 (2016) 45–56

Contents lists available at ScienceDirect

Automation in Construction

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

Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports

Antoine J.-P. Tixier [a], Matthew R. Hallowell [a,*], Balaji Rajagopalan [b], Dean Bowman [c]

[a] Department of Civil, Environmental, and Architectural Engineering, University of Colorado at Boulder, Boulder, CO 80309, USA
[b] Department of Civil, Environmental, and Architectural Engineering, Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado at Boulder, Boulder, CO 80309, USA
[c] Bentley Systems, Exton, PA, USA

ARTICLE INFO

Article history:
Received 12 March 2015
Received in revised form 20 October 2015
Accepted 7 November 2015
Available online 21 November 2015

Keywords:
Automated content analysis
Natural language processing
Text mining
Knowledge extraction
Accident
Injury
Safety
Attribute
Risk
R

ABSTRACT

In the United States like in many other countries throughout the globe, construction workers are more likely to be injured on the job than workers in any other industry. This poor safety performance is responsible for huge human and financial losses and has motivated extensive research. Unfortunately, safety improvement in construction has decelerated in the last decade and traditional safety programs have reached saturation. Yet major construction companies and federal agencies possess a wealth of empirical knowledge in the form of huge databases of digital construction injury reports. This knowledge could be used to better understand, predict, and prevent the occurrence of construction accidents. Unfortunately, due to the lack of a clear methodology and the high costs of manual large-scale content analysis, these valuable data have yet to be extracted and leveraged. Recently, researchers have proposed a framework allowing meaningful empirical data to be extracted from accident reports. However, the resource limitations inherent to manual content analysis still remain. The present study tested the proposition that manual content analysis of injury reports can be eliminated using natural language processing (NLP). This paper describes (1) the overall strategy and methodology used in developing the system, and specifically how key challenges with decoding unstructured reports were overcome; (2) how the system was built through an iterative process of coding and testing against manual content analysis results from a team of seven independent analysts; and (3) the implications and potential uses of the data extracted. The results indicate that the NLP system is capable of quickly and automatically scanning unstructured injury reports for 101 attributes and outcomes with over 95% accuracy. The main contribution of this research is to empower any organization to quickly obtain a large and highly reliable structured attribute and outcome data set from their databases of unstructured accident reports. Such structured data are a necessary prerequisite to the application of statistical modeling techniques, allowing the extraction of new safety knowledge and finally the amelioration of safety management.

1. Motivation

Construction is constantly ranked as one of the most dangerous industries worldwide [51]. In the United States, despite the improvements that followed the Occupational Health and Safety Act of 1970, construction still accounts for 17% of all work-related deaths while only employing 7% of the national workforce [17]. In fact, according to the Bureau of Labor Statistics [61], approximately 700 workers die each year. Construction fatalities and injuries result in immense societal costs, totaling approximately $15 billion in lost revenue every year [61]. What is even more alarming is that these colossal human and financial costs are expected to escalate with the 33% construction employment growth projections in the 2010–2020 decade, which is more than twice the overall anticipated economic growth [17].

Despite the abundant research that has been motivated by the aforementioned alarming injury and fatality rates, safety performance in construction has been plateauing in recent years, and the implementation of effective injury prevention practices has reached saturation [19]. Fortunately, risk-based approaches are emerging and show promise for safety improvement through proactive decision making. For example, Baradan and Usmen [6] compared the risk of building trades, Hallowell and Gambatese [62] quantified the safety risk for various activities required to construct concrete formwork, and Shapira and Lyachin [63] studied the impact of tower cranes on jobsite safety. However, these approaches are currently limited because (1) they

* Corresponding author.
E-mail addresses: antoine.tixier-1@colorado.edu (A.J.-P. Tixier), matthew.hallowell@colorado.edu (M.R. Hallowell), rajagopalan.balaji@colorado.edu (B. Rajagopalan), dean.bowman@bentley.com (D. Bowman).

http://dx.doi.org/10.1016/j.autcon.2015.11.001
0926-5805/

focus on specific activities and trades without considering the temporal and spatial interactions among risk factors, (2) they are not based on empirical data, and (3) they are limited in scope of application [48,51]. Consequently, existing models do not translate well to other work scenarios and do not capture the dynamics of construction work, where trades and activities constantly interact [31,51]. To overcome these limitations, Esmaeili and Hallowell [19] proposed a unified attribute-based framework that allows standard risk factor and outcome variables to be extracted from naturally occurring accident reports. Although this method shows promise, it requires the analysis of large numbers of reports to see patterns and trends emerge from the data. Such manual content analysis is laborious and resource-intensive [18,48].

To remove the needs for manual analysis of injury reports and allow the large-scale use of the attribute-based framework, the present study tests the proposition that attributes and safety outcomes can be automatically and accurately extracted from unstructured injury reports using natural language processing (NLP).

1.1. Background: attribute-based approach to construction safety

The attribute-based approach to construction safety theorizes that any construction situation can be uniquely and comprehensively characterized by a finite number of observable fundamental construction site attributes [18,48]. These basic elements are context-free, universal, and pertain to construction means and methods, environmental conditions, and human factors. For instance, in the following excerpt of an injury report, "employee was welding overhead and wind shifted, resulting in discomfort to left eye," three fundamental attributes can be identified: (1) welding, (2) working overhead, and (3) wind.

Although this approach is simple, it is very powerful. First, from this perspective, any incident can be viewed as the resulting outcome of the presence of a worker and the joint occurrence of some fundamental attributes. Consequently, attributes are also called *injury precursors* or simply *precursors*. In what follows, the terms *attribute* and *precursor* are used interchangeably. It is important to note that precursors should be observable before an injury occurs. *Falling object*, for example, is not a precursor, it is an outcome. On the other hand, o*bject at height* is a precursor. As illustrated in the previous example, descriptors of the work environment and outcomes can be extracted even from brief reports. Finally, this information is authentic since it is simply based on objective narratives of discrete events.

A connection with genetics can naturally be made with this style of analysis: every person is unique, but their genetic information is entirely encoded by combinations of a finite number of basic universal building blocks that constitute their DNA. The attribute-based approach to construction safety is built upon a similar theory that by identifying fundamental and universal construction injury precursors, understanding how they interact, and modeling how they shape risk and create unsafe work conditions, it may be possible to better understand the true nature of, predict, and prevent the occurrence of construction injuries. Historically, scientific understanding of complex phenomena has improved when breaking down convoluted systems into fundamental constituents that individually are easier to comprehend. A fascinating recent example is the Human Genome Project [16], which allowed sequencing and mapping of about 30,000 genes, unlocking the structure of human DNA. Similarly, the finite element method, a numerical technique used in many quantitative disciplines of engineering, is built on the theory that complicated continuous structures and objects can be represented by a finite number of geometrically simpler pieces [60].

Esmaeili and Hallowell [19] conducted the first attribute-level risk analysis in construction by analyzing 300 struck-by injury cases from national databases. Through this analysis, they identified 34 fundamental attributes. More recently, a team of eight researchers performed a manual content analysis of 2201 industrial injury reports gathered from 476 contractors, allowing the initial list of Esmaeili and Hallowell [19] to be refined and broadened to 80 precursors [18,48]. These precursors are summarized in Table 1. The validity of the content analysis and the relevance of the attributes presented by Prades [48] and Desvignes [18] were ensured by adhering to a strict coding scheme, implementing an iterative process with team-based calibration meetings, and using peer reviews and random checks by external reviewers. In these studies, attributes were classified in three categories: upstream, transitional, and downstream. Upstream precursors can be anticipated as soon as during the design phase, transitional precursors are generally not identifiable by designers but can be detected before construction begins based on knowledge of construction means and methods, and downstream precursors are mostly related to human behavior and can only be observed during the construction phase.

In addition to the 80 precursors presented in Table 1, Prades [48] and Desvignes [18] also extracted various safety outcomes from accident reports, namely, injury type, injury severity, body part affected, and energy source involved. The variables belonging to these categories are listed in Table 2. The injury codes, severity levels, and body divisions included in Table 2 are consistent with OSHA definitions and past research [30]. Energy sources were extracted based on the theory that any injury can be associated with the release of an energy source [23,28]. For instance, a falling load can be seen as a *gravity* release, a welding flash burn involves *radiation*, and waterproofing substances, solvents, or concrete in its liquid form can cause *chemical* burns. Additional definitions and examples can be found in Albert et al. [3]. Attributes and outcomes are occasionally referred to as variables in the rest of this paper.

Although the work of Esmaeili and Hallowell [19], Prades [48], and Desvignes [18] made important contributions to attribute-level safety analysis, the manual content analysis procedures used were time consuming, limiting the number of reports that could be analyzed in a reasonable research effort, and thereby the emergence of trends and patterns in the data extracted. For example, Desvignes [18] only used a random set of 1280 reports from a larger set of 4458 available reports because of time and resource limitations. In addition, even when a rigorous protocol is followed, it is never possible to entirely eliminate inconsistencies among human coders. For all these reasons, resorting

Table 1
Context-free validated injury precursors from Desvignes [18].

Upstream	Rebar	Screw
Cable tray	Scaffold	Slag
Cable	Soffit	Spark
Chipping	Spool	Slippery walking surface
Concrete liquid	Stairs	Small particle
Concrete	Steel sections	Adverse low temperatures
Conduit	Stripping	Unpowered tool
Confined workspace	Tank	Unstable support/surface
Congested workspace	Unpowered transporter	Wind
Crane	Valve	Wrench
Door	Welding	Lifting/pulling/manual handling
Dunnage	Wire	Light vehicle
Electricity	Working at height	Exiting/transitioning
Formwork	Working below elevated workspace/material	Sharp edge
Grinding	Drill	Splinter/sliver
Grout	**Transitional**	Repetitive motion
Guardrail/handrail	Bolt	Working overhead
Heat source	Cleaning	**Downstream**
Heavy material/tool	Forklift	Improper body position
Heavy vehicle	Hammer	Improper procedure/inattention
Job trailer	Hand size pieces	Improper security of materials
Lumber	Hazardous substance	Improper security of tools
Machinery	Hose	No/improper PPE
Manlift	Insect	Object on the floor
Stud	Ladder	Poor housekeeping
Object at height	Mud	Poor visibility
Piping	Nail	Uneven walking surface
Pontoon	Powered tool	