Automation in Construction 74 (2017) 103–117

Contents lists available at ScienceDirect

Automation in Construction

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

Ontology-based automated information extraction from building energy conservation codes

Peng Zhou, 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 2 January 2016
Received in revised form 22 August 2016
Accepted 19 September 2016
Available online xxxx

Keywords:
Information extraction
Ontology
Natural language processing
Automated compliance checking
Energy conservation codes

ABSTRACT

An ontology-based information extraction algorithm for automatically extracting energy requirements from energy conservation codes is proposed. The proposed algorithm aims to support fully-automated energy compliance checking in the construction domain by allowing automated extraction of the requirements from the codes instead of the status quo which relies on manual extraction of requirements from codes and manual formalization of those requirements in a computer-processable format. Automated information extraction from energy conservation codes, compared to other building codes, is a far complex task because many code provisions are long, hierarchically-complex, and with exceptions. A combination of text classification methods, domain-specific preprocessing techniques, ontology-based pattern-matching extraction techniques, sequential dependency-based extraction methods, and cascaded extraction methods is proposed to deal with such complexity in extraction. The proposed algorithm was tested in extracting energy requirements from Chapter 4 of the 2012 International Energy Conservation Code, and the results showed 97.4% recall and 98.5% precision.

1. Introduction

Environmental compliance checking aims to help construction projects comply with environmental codes and regulations such as the International Energy Conservation Code (IECC). Because manual compliance checking is time-consuming and costly [10,39], a number of research efforts aimed to automate the compliance checking process. Examples of automated compliance checking (ACC) efforts in the past five years include checking of building envelope performance [39], building safety design [33], building structural design [30], construction quality [45], building safety design and planning [27], building water network design [25], building fire safety [9,23], building evacuation [6], building sustainability [5], and formwork constructability [17]. Despite the importance of these efforts, existing ACC systems and methods are not fully automated; they require (1) intensive manual effort in extracting requirements from regulatory documents and encoding these requirements in a computer-processable format (e.g., [17]), or (2) substantial manual effort in annotating regulatory documents (e.g., [5,14]).

To address this gap, Zhang and El-Gohary [43,44] proposed an information extraction (IE) methodology for automatically extracting information from building codes [43] and an information transformation methodology for automatically transforming the extracted information into a computer-processable rule format [44]. Compared to building codes, automatically extracting requirements from energy codes is more challenging because of (1) longer provisions: provisions in energy codes are longer, which indicates that requirements are more likely to be noisy and/or semantically complex; (2) hierarchically-complex sentence structure s: text in energy codes has more complex sentence structures, in which one provision may contain multiple levels of subprovisions, and one subprovision may contain multiple requirements; and (3) more exceptions: a requirement in energy codes may contain one or multiple exceptions for waiving the compliance with the requirement if one or all of a set of exception conditions are met.

In this paper, an ontology-based information extraction (OBIE) algorithm for automatically extracting regulatory requirements from energy conservation codes is proposed. The proposed algorithm advances existing IE methods in the construction domain in four main ways. First, it extracts regulatory requirements from pre-classified text rather than unclassified text, which aims to improve the efficiency (by avoiding unnecessary computational processing of irrelevant text) and performance (by avoiding potential noise and errors resulting from processing irrelevant text) of IE. Second, it uses a deeper (more detailed) ontology, which aims to better capture domain-specific meaning. Third, it applies conceptual dependency theory to build a conceptual dependency structure and proposes a sequential dependency-based extraction method, which aim to reduce text ambiguities. Fourth, it proposes domain-specific preprocessing techniques and cascaded extraction methods, which aim to deal with the complexity of the text (i.e., longer provisions, hierarchically-complex sentence structures, and more exceptions). The proposed algorithm was tested in extracting

* Corresponding author.
E-mail address: gohary@illinois.edu (N. El-Gohary).

http://dx.doi.org/10.1016/j.autcon.2016.09.004
0926-5805/

commercial building energy efficiency regulatory requirements from the 2012 IECC [15].

2. Background

2.1. Information extraction

Natural language processing (NLP) is a subdiscipline of artificial intelligence that aims to enable computers to understand human language [24]. IE applies NLP techniques [e.g., part-of-speech (POS) tagging, morphological analysis, etc.] to recognize information from unstructured data and formalize it into structured data [18]. According to the level of complexity, IE can be categorized into four types: (1) named entity recognition, which aims to identify particular entities [18]; (2) relation detection, which aims to discern the relationships among the identified entities [18]; (3) event extraction, which aims to identify events from text (each event has a trigger and a number of associated arguments, and each event may be composed of a number of entities and their relationships) [13,32]; and (4) full IE, which aims to extract all information expressed by a sentence based on a full analysis of the sentence [43]. Named entity recognition, relation detection, and event extraction can be classified as shallow IE because they aim to extract partial information from a sentence, whereas full IE could be classified as deep IE because it aims to extract all information from a sentence [43].

There are different approaches to IE [18,28,29], including rule-based and supervised machine learning (ML)-based approaches. A rule-based approach requires human effort to analyze the text features in a relatively small set of text corpus (sometimes called developing data, which is analogous to training data in the case of ML), define the text patterns in terms of the text features, and then develop extraction rules based on the defined patterns. Text features may include [28]: (1) syntactic features, which refer to syntax-related features that are determined based on grammatical analysis, such as POS tags (e.g., tag "IN" represents a preposition like "for"); and/or (2) semantic features, which refer to concepts that capture the meaning of the information (e.g., "mass wall" is a concept that represents a type of wall). The patterns may be defined in terms of combinations of different syntactic and/or semantic features via regular expressions. Regular expressions is a language that is implemented by computers for pattern matching to characterize possible sequences of text [18].

A supervised ML-based approach requires human effort to collect a relatively large set of training data and annotate them with the relevant text features and with the information that should be extracted. Then, an ML algorithm (e.g., using Support Vector Machines, Hidden Markov Model, or Conditional Random Fields) is used to automatically learn the extraction rules from the annotated training data. Compared with the rule-based approach, the ML-based approach (1) requires a much larger size of annotated training data: because the performance of an ML-based IE algorithm depends on the training data for learning, a sufficiently large size of training data is required to accurately learn the text patterns and the extraction rules; and (2) does not require manual effort in pattern definition and extraction rule development: an ML algorithm automatically learns the patterns of the text and the extraction rules.

Although an ML-based approach can save the manual effort in pattern definition and extraction rule development, a rule-based approach is adopted in this research for two main reasons. First, a rule-based approach tends to yield higher performance, because human expertise usually results in more accurate patterns and extraction rules [28]. The performance of ML in a complex task such as IE is usually inconsistent and insufficient [16]. In this specific application, the level of complexity in IE is even much higher, compared to the state-of-the-art IE, which makes a rule-based approach especially suitable in this case; deep IE is needed to extract all information that describes a regulatory requirement and high performance is needed to support high performance ACC – both making the IE problem quite challenging. Second, in this application, the manual effort in pattern definition and extraction rule development in the rule-based approach is expected to be less than that required for manually annotating a sufficiently large size of training data if taking an ML-based approach.

2.2. Ontology-based information extraction

OBIE is a subfield of IE. Comparing to non-ontology-based IE, which only depends on the lexical and/or syntactic information of the text, OBIE further relies on semantic information to extract information based on meaning. In many cases, OBIE is domain and application-oriented, when a domain and/or an application ontology is used to assist in extracting semantic information that is specific to a particular domain and/or application [19,42,43]. In this case, OBIE captures domain-specific semantic information as semantic features, which are then used in the patterns in the extraction rules. Compared with non-ontology-based IE, the domain-specific semantic information that is used in OBIE is promising in improving the IE performance for a specific domain [42,43].

OBIE has been explored in different domains such as biology (e.g., [29]), business (e.g., [3,40]), law (e.g., [28]), medicine (e.g., [38]), mechanical engineering (e.g., [22]), and civil engineering (e.g., [43]). OBIE has also been explored in different complexity levels of IE: named entity recognition (e.g., [29]), relation detection (e.g., [22,38,40]), event extraction (e.g., [3]), and full IE (e.g., [43]). The most complex level (i.e., full IE) is the most challenging and the least explored. In terms of approach, all these efforts used a rule-based approach to deal with the OBIE problem.

3. State of the art and knowledge gaps in automated information extraction in construction

Despite the large number of IE efforts outside the construction domain, the number of IE efforts, especially OBIE efforts, are limited in the construction domain. For non-ontology-based IE efforts, Al Qady and Kandil [2] used limited syntactic features [i.e., specific phrases like VP (i.e., verb phrase) segment and its role ACTIVE_VERB] to extract concepts and relations from contract documents, with the aim to improve construction document management. Abuzir and Abuzir [1] used document structure features (i.e., HTML tags) and simple lexico-syntactic features (e.g., "such as" is a lexico-syntactic feature that was used to extract the terms following it because it usually indicates a synonym relationship among these terms) to extract terms and their relations from web pages, with the aim to construct a thesaurus of civil engineering. For OBIE efforts, Zhang and El-Gohary [43] used a combination of syntactic and semantic features to extract regulatory requirements from building codes for supporting automated code compliance checking, where the semantic features were extracted using a building ontology. Despite the importance of these efforts, they are still limited in one or more of the following four main ways. First, existing efforts extract information from unclassified text, which may result in unnecessary processing effort and may increase extraction errors due to processing irrelevant text. None of these efforts explored the use of text classification techniques to filter out irrelevant text prior to IE to improve the efficiency and performance of IE. Second, existing efforts were not tested in deep IE from long provisions with multiple exceptions. For example, Abuzir and Abuzir [1] and Al Qady and Kandil [2] conducted shallow IE (extracting partial information from a sentence, whereas deep IE aims to extract all information expressed by a sentence based on a full analysis of the sentence). Zhang and El-Gohary [43], on the other hand, conducted deep IE, but tested their algorithms in extracting requirements from international building codes, which include relatively shorter provisions with fewer exceptions in comparison to energy conservation codes; energy conservation codes include relatively long provisions with several exceptions. Third, existing efforts are limited in automatically dealing with text that includes hierarchically-complex sentence structures. For example, Al Qady and Kandil [2] used a manual approach