Ontology-Based Multi-label Text Classification for Enhanced Information Retrieval for Supporting Automated Environmental Compliance Checking

Peng Zhou[1]and Nora El-Gohary[2]

[1]Graduate Student, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Avenue, Urbana, IL 61801; PH (217) 402-4111; FAX (217) 265-8039; email: pzhou6@illinois.edu
[2]Assistant Professor, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Avenue, Urbana, IL 61801; PH (217) 333-6620; FAX (217) 265-8039; email: gohary@illinois.edu

ABSTRACT

In order to fully automate the environmental regulatory compliance checking process, we need to automatically extract the rules from applicable environmental regulatory textual documents, such as energy conservation codes. In our automated compliance checking (ACC) approach, prior to rule extraction, we first classify the text into pre-defined categories to only retrieve relevant clauses and filter out irrelevant ones, thereby improving the efficiency and accuracy of rule extraction. Machine learning (ML) techniques have been commonly-used for text classification (TC). ML-based TC has, generally, performed well. However, given the need for an exceptionally-high performance (100% recall and >85% precision) for ACC (to avoid consequent compliance reasoning errors), we need further performance improvement. Therefore, in this paper, we present an ontology-based TC algorithm to further improve the classification performance by utilizing the semantic features of the text. We used a domain ontology for conceptualizing the environmental knowledge. In comparison to the ML-based approach, in our ontology-based approach, a document (or clause) is represented in terms of semantic concepts and relations, rather than just terms (words). The semantic concepts and relations in the ontology (e.g. “is-a” relations) help in recognizing the semantic features of the text. Our ontology-based TC algorithm was tested on twelve environmental regulatory documents such as the 2012 International Energy Conservation Code, evaluated in terms of precision and recall, and compared with our previously-utilized ML-based approach. Our results show that our ontology-based approach achieves 96.62% and 96.34% recall and precision, respectively, thereby outperforming the ML-based approach.

INTRODUCTION

Manual compliance checking is error-prone and time-consuming (Eastman *et al.* 2009). Automated compliance checking (ACC) aims at addressing this practical gap by reducing the cost and time of checking the compliance of construction projects to regulatory requirements. Recent efforts of ACC in the construction domain include: 1) modeling regulation constraints using an ontology-based approach for supporting construction quality compliance checking (Zhong *et al*. 2012); and 2) encoding rules manually for building design checking (Eastman *et al.* 2009). Despite the importance of these efforts, existing ACC systems are not fully automated; manual effort is still needed to extract the requirements from textual regulatory documents (e.g. codes) and encode them in a computer-processable format.

To address this gap, in our previous work we proposed a model-based Natural Language Processing (NLP)-enabled approach for ACC in construction. We proposed the use of NLP techniques, including text classification (TC) (Salama and El-Gohary 2013; Zhou and El-Gohary 2014) and information extraction (IE) (Zhang and El-Gohary 2013), to support automated text processing and analysis for automated rule extraction from regulatory documents. In our ACC approach, prior to rule extraction, we first classify the text into pre-defined categories to only retrieve relevant clauses and filter out irrelevant ones, thereby improving the efficiency and accuracy of rule extraction.

Machine learning (ML) techniques have been commonly-used for TC. An example is our previous work in environmental document classification (Zhou and El-Gohary 2014). Supervised ML-based TC has, generally, performed well. However, given the need for an exceptionally-high performance (100% recall and >85% precision) for ACC (to avoid consequent compliance reasoning errors), we need further performance improvement. Therefore, in this paper, we present an ontology-based TC algorithm to further improve the classification performance by utilizing the semantic features of the text. As such, in this paper, we build on our previous work in TC in two main ways. First, compared with the commonly-used supervised ML-based TC approach, we explore the use of ontology to develop a fully semantic TC approach (without supervised ML) for the construction domain. Ontology is a knowledge conceptualization for modeling the concepts and their relationships for capturing the semantics of a domain. Therefore, we propose an ontology-based approach to make use of these semantic information for classifying environmental regulatory documents. Second, we deal with the multi-label classification problem in a direct way, instead of transforming the multi-label classification problem to multiple single-label classification problems (as commonly-used in ML-based TC).

In the remainder of this paper, we present our proposed ontology-based TC approach for classifying environmental clauses according to a set of predefined, semantic labels. The labels are defined based on a hierarchy of topics in our ACC model. For each topic, an ontology is built to model the concepts and relationships that are related to this topic. Then we apply a deep learning algorithm to learn the similarities between each clause (based on the terms in a clause) and each topic (based on the ontology concepts related to this topic) for classifying each clause into zero or more topics. We finally compare the performance of our proposed ontology-based TC approach to the performance of our previous ML-based approach (Zhou and El-Gohary 2014).

BACKGROUND AND KNOWLEDGE GAPS

NLP aims at enabling computers to analyze and process natural language in a meaningful way to facilitate a range of tasks (e.g. automated machine translation) (Manning and Schutze 1999). TC, a subfield of NLP, aims at classifying documents (like paragraphs or clauses) to one or more categories (Manning and Schutze 1999). A category is represented by a label, and may refer to a class or concept.

A TC problem could be categorized as a multi-label or single-label TC (Katakis and Tsoumakas 2007). Multi-label TC can assign more than one label to a document, while single-label TC can only assign one label for each document. In our application, we address a multi-label TC problem, since multiple labels could be assigned to one clause.

ML techniques have commonly been used for TC (e.g. Salama and El-Gohary 2013; Zhou and El-Gohary 2014). While generally successful, ML-based TC usually discards semantic text information (e.g. meaning of words), which is potentially very useful in identifying the correct label(s) of a document. Semantic-based TC has, thus, been introduced to capture and use the semantics of the text for improving the TC performance. The use of ontologies (a type of semantic model) in TC has, therefore, recently attracted much research efforts. In this regard, we identified two main research gaps: 1) To the best of our knowledge, there have been no research efforts for using ontology-based TC in the construction domain. This is a lost opportunity for improving the performance of TC-based applications in construction; and 2) Outside of the construction domain, ontology-based TC efforts: a) still rely on supervised ML for training the classifier – using labelled training data – to learn the rules for labelling any given text. This involves much manual effort in labelling the training data (Vogrincic and Bosnic 2011); and/or b) are unable to deal with a multi-label TC problem directly. This requires transformation to multiple single-label problems (Vogrincic and Bosnic 2011). To address these knowledge gaps, in this paper, we propose an ontology-based multi-label TC approach for semantic TC in the construction domain. Our approach does not rely on supervised ML and is able to deal with the multi-label classification problem directly.

PROPOSED METHOD FOR ONTOLOGY-BASED TEXT CLASSIFICATION OF ENVIRONEMTAL REGULATORY DOCUMENTS

We are proposing a four-phase method for ontology-based domain-specific TC, including: 1) TC topic hierarchy and ontology development; 2) data preparation; 3) ontology-based classification; and 4) classification result evaluation.

Step1: TC Topic Hierarchy and Ontology Development

A topic hierarchy is first developed to identify the labels that will be used for TC. In this paper, we focus on analyzing the ‘energy efficiency topic’, which is a subtopic of ‘environmental topic’ (as per Figure 1). All the leaf nodes (11 subtopics) in the taxonomical topic hierarchy, in addition to the “Total Building System Energy Efficiency Topic”, were used as labels of classification. For more details on the topic hierarchy and its development methodology, the reader is referred to Zhou and El-

Gohary (2014). For modeling the semantic information associated with each topic, we then extended the hierarchy into an ontology using the ontology development methodology by El-Gohary and El-Diraby (2010). For each topic, a sub-ontology is built to model the concepts and relationships that are related to this topic. Figure 2 shows a partial sub-ontology developed for the "Air Leakage Topic".

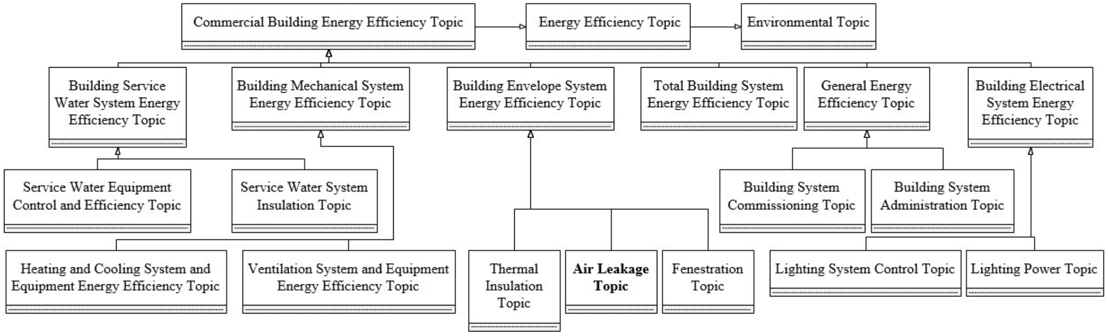

Commercial Building Energy Efficiency Topic
Energy Efficiency Topic
Environmental Topic
Building Service Water System Energy Efficiency Topic
Building Mechanical System Energy Efficiency Topic
Building Envelope System Energy Efficiency Topic
Total Building System Energy Efficiency Topic
General Energy Efficiency Topic
Building Electrical System Energy Efficiency Topic
Service Water Equipment Control and Efficiency Topic
Service Water System Insulation Topic
Building System Commissioning Topic
Building System Administration Topic
Heating and Cooling System and Equipment Energy Efficiency Topic
Ventilation System and Equipment Energy Efficiency Topic
Thermal Insulation Topic
Air Leakage Topic
Fenestration Topic
Lighting System Control Topic
Lighting Power Topic

Figure 1. Text classification topic hierarchy.

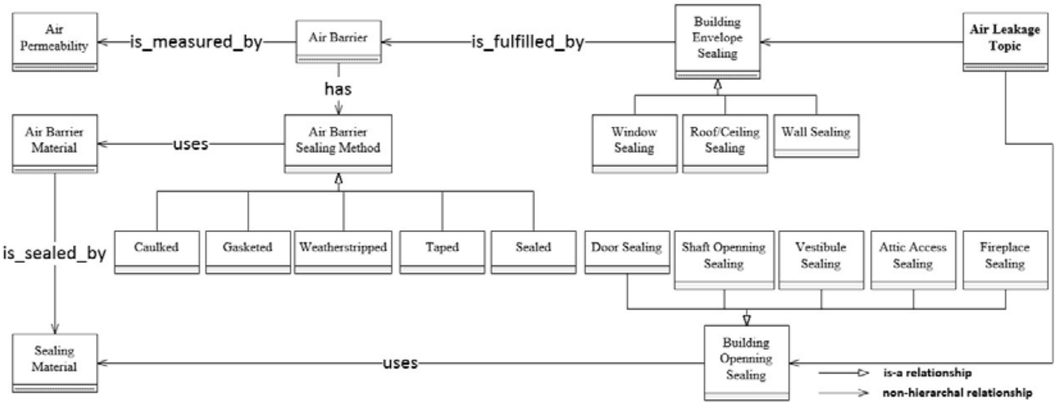


Figure 2. Partial sub-ontology for "Air Leakage Topic".

Step2: Data Preparation

Data Collection and Labelling. We collected around 1,600 clauses from 12 documents. In collecting the data (clauses), we considered data sufficiency in terms of quantity and quality. The performance of our proposed methodology highly depends on the quantity and quality of data (Mikolov 2013) due to the use of deep learning (in Step 3). The use of large amounts of data can facilitate accurate learning of semantics. Good quality data refers to the words in sentences being logical and coherent. Based on our preliminary experimental results, the good performance results indicate that our data is sufficient in terms of quantity and quality. We then split the 1,600 clauses into two sets, training set and testing set, at a ratio 2:1. Because our approach does not rely on supervised ML, there was no need to label the training set. Therefore, we only manually labelled the testing set. The testing set is further used in performance evaluation (in Step 4).

Step3: Ontology-Based Classification

Data Preprocessing. Data preprocessing is the process of transforming the raw text into the format required by our ontology-based TC approach. We implemented three steps for data processing: 1) Tokenization: Tokenization aims at segmenting the text into words or tokens; 2) Stemming: Stemming aims at stripping off word suffixes (in some cases prefixes) to its root or stem. We implemented stemming using Porter2 stemming algorithm in Python programming language; and 3) Stopword Removal: Stopwords refer to those high-frequency and low-content words like "am", "is", "a", "the", "of". We coded a Python preprocessing program for implementing the above-mentioned three sub-tasks. The input to the program are two sets of raw text (.txt) files (training and testing clause sets, as obtained from Step 2), and the output are two pre-processed datasets (training dataset and testing dataset). The training and testing datasets are the input the ontology-based TC algorithm (discussed in the following section).

Ontology-Based TC Implementation. After the datasets have been preprocessed, we first applied a deep learning algorithm "word2vec" (Mikolov 2013) on the training dataset to capture similarities between each concept in the ontology and each term in a clause. Deep learning is a branch of artificial intelligence (AI)-oriented ML, which aims at learning the semantic meanings behind words, in an unsupervised way. Deep learning can measure the similarity between two concepts in order to capture ontological relationships like is-a relationship. In our approach, the similarities between each term in a testing clause (i.e. a clause in the testing dataset) and each concept related to a topic are summed up for each topic to compute the total similarity between a clause and a topic. All topics with a positive total similarity with a clause are selected as potential labels for that clause. If there is exactly one topic with a positive total similarity with a clause, then that topic is assigned as the only label for that clause. If there are no topics with a positive total similarity with a clause, then no topics are assigned to that clause. If there are more than one topic with a positive total similarity with a clause, then at least the topic with the highest total similarity should be assigned as the primary label for that clause (the corresponding topic is then referred to as a primary topic). In this case, the total similarity difference between this primary topic and the other topic(s) are measured. Then, a total similarity difference threshold is used to further identify secondary labels; topics with a total similarity difference less than the threshold value are selected as secondary labels for that clause. The total similarity difference threshold is set experimentally for maximizing the recall. We have implemented a python programming language version of word2vec, Gensim, to compute the similarity values (Radim and Petr 2010).

In comparison to other existing ontology-based TC methods (e.g. Vogrincic and Bosnic 2011; WijewickremaI and Gamage 2013), our proposed method is different in the following ways: 1) we do not use supervised ML for training the classifier to learn the rules of labelling. Instead, we use unsupervised ML (deep learning), and we use it for a different TC task; we use deep learning for learning the semantic similarity between a term of a clause and a concept related to a topic. In unsupervised ML, the learning is an unsupervised task (i.e., the training data does not specify what we are trying to learn), and thus labelling of the training data is not needed. Therefore, in our approach, we only label the testing data, which saves much manual effort; and 2) we use a direct (without problem transformation) multi-label ontology-based TC method. We can deal with multi-label classification problems directly without transforming the multi-label problem to multiple single-label problems. This reduces the data preparation effort.

Step4: Classification Result Evaluation

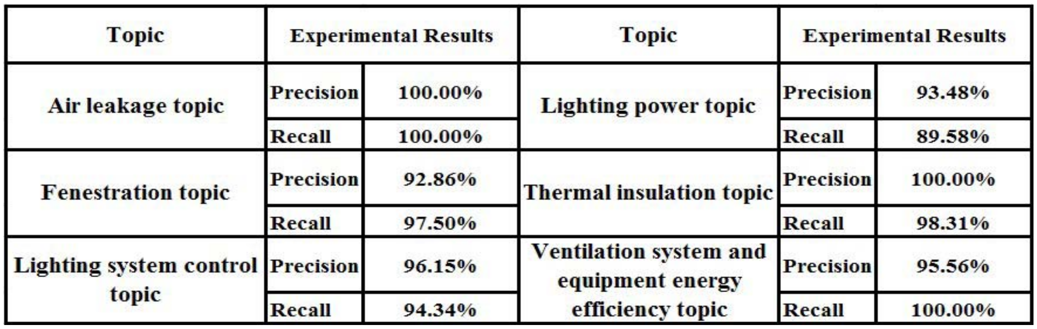
We first constructed a Confusion Matrix (CM) to collect the classification results for analyzing the performance of the classifier (Manning 2009). A CM is a number-of-classes rows and number-of-classes columns matrix, in which the main diagonal shows how many clauses are labelled correctly and other entries of the matrix show how many errors are made by classifying a clause from one class to another class. A CM can, thus, help visualize the overall classification results and compute recall and precision for evaluating each topic. For each topic, recall measures the number of correctly labelled clauses as a percentage of the total number of clauses that should be labelled for that topic. Precision measures the number of correctly labelled clauses as a percentage of the total number of labelled clauses for that topic. Typically, there is a tradeoff between recall and precision. In our application, recall is given higher priority than precision since missing to recall one clause means overlooking a relevant clause, which may undermine the performance of the ACC system.

PRELIMINARY EXPERIMENTAL RESULTS AND ANALYSIS

Evaluation of the Ontology-Based TC Algorithm. We tested the performance of our ontology-based TC algorithm on six topics. The preliminary experimental results are summarized in Table 1. We achieved an overall average recall and precision of 96.62% and 96.34%, respectively. The “Ventilation System and Equipment Energy Efficiency Topic” achieved 100% recall, while the “Air Leakage Topic” even reached 100% recall and precision. The “Lighting Power Topic” and “Lighting System Control Topic” achieved the least recalls, with 93.48% and 93.34%, respectively. This is expected since, intuitively, there are semantic relationships that are making the two topics overlapping. For example, installing an automatic lighting shut-off can save energy to help meet the requirements of lighting power. In our future work, we will explore if/how refining our ontology (e.g. adding more detailed concepts) could address cases of overlapping topics.

Performance Comparison: Ontology-based Approach vs. ML-based Approach. We further compared the performance of our proposed ontology-based approach to that of the ML-based approach (that we proposed in our previous work (Zhou and El-Gohary 2014)). The performance of both approaches in terms of recall and precision are illustrated in Figure 3. In terms of overall average recall and precision, the ontology-based approach achieves 96.62% and 96.34%, respectively. It, thus, highly outperforms the ML-based approach which achieved 90.17% and 88.50% overall average recall and precision, respectively. In terms of the standard deviations of recall and precision for all topics, the ontology-based approach reaches 4.03% and 3.09%, while the ML-based approach reaches 4.26% and 5.58%, respectively. Though the standard deviation of the ML-based approach is small, the even smaller standard

Table 1. Preliminary Experimental Results.



Topic	Experimental Results		Topic	Experimental Results	
Air leakage topic	Precision	100.00%	Lighting power topic	Precision	93.48%
	Recall	100.00%		Recall	89.58%
Fenestration topic	Precision	92.86%	Thermal insulation topic	Precision	100.00%
	Recall	97.50%		Recall	98.31%
Lighting system control topic	Precision	96.15%	Ventilation system and equipment energy efficiency topic	Precision	95.56%
	Recall	94.34%		Recall	100.00%

deviation values for the ontology-based approach show its potentially more robust performance. In terms of analyzing the performance of each topic, only the recall of the "Lighting Power Topic" yielded by the ML-based approach exceeds that of the ontology-based approach by 4.42%. The ontology-based approach outperforms for all other 5 topics.

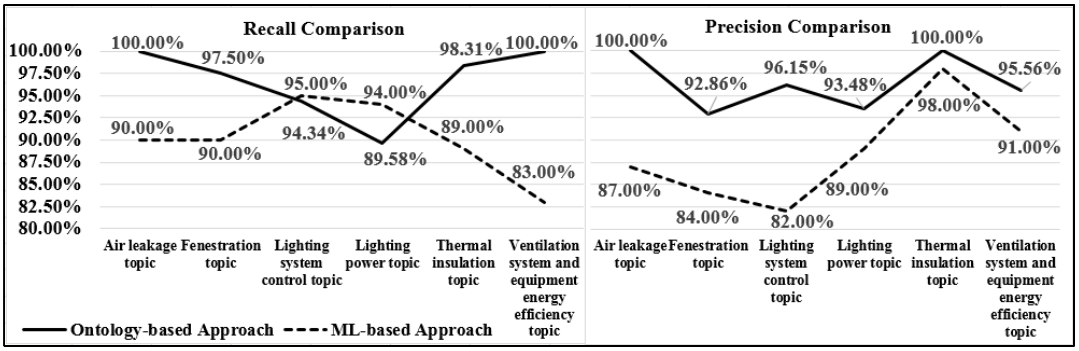


Figure 3. Performance comparison of ontology-based approach and ML-based approach in terms of recall and precision.

CONCLUSION, CONTRIBUTION, AND FUTURE WORK

This paper proposed an ontology-based multi-label TC approach for classifying environmental regulatory clauses for supporting ACC in construction. We achieved 96.6% and 96.3% overall average recall and precision, respectively. Compared with the supervised ML-based approach (proposed in our previous work), our ontology-based approach outperforms and is promising. The advantages of using our ontology-based approach include: 1) the time-consuming effort in labelling the training dataset is eliminated; and 2) the semantic information of the text is captured and used for improving the TC performance.

This work contributes to the body of knowledge, in the domain of TC in construction, in two main ways. First, compared with the commonly-used supervised ML-based TC, we proposed an ontology-based TC approach for classifying environmental regulatory documents in the construction domain. Based on our preliminary experimental results, we showed that the ontology-based approach outperforms the ML-based approach. Second, we proposed a method for ontology-based TC that deals with the multi-label classification in a direct way, without the need for transforming the multi-label classification problem into multiple single-label ones. This reduces the data preparation effort.

In future work, we will continue to 1) conduct experiments on other topics (showed in our topic hierarchy); 2) explore the extension of our ontology-based TC approach for classifying regulatory documents of other types of topics (e.g. safety topics); and 3) work on automating the ontology construction process.

ACKNOWLEDGEMENT

The authors would like to thank the National Science Foundation (NSF). This material is based upon work supported by NSF under Grant No. 1201170. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

REFERENCES

Eastman, C., Lee, J., Jeong, Y., and Lee, J. (2009). “Automatic rule-based checking of building designs.” *Automat. Constr.*, 18(8), 1011-1033.

El-Gohary, N., and El-Diraby, T. (2010). “Domain ontology for processes in infrastructure and construction.” *J. Constr. Eng. and Manage.*, 136(7), 730-744.

Manning, C. D., and Schutze, H. (1999). *Foundations of statistical NLP*, MIT Press.

Manning, C. D., Raghawan, P., and Schutze, H. (2009) *Introduction to information retrieval*, Cambridge University Press.

Mikolov, T. (2013). “word2vec: Tool for computing continuous distributed representations of words.” https://code.google.com/p/word2vec/(Dec.12, 2013).

Rehurek, R., and Sojka, P. (2010). “Software framework for topic modelling with large corpora.” *Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45-50,Valletta, Malta.

Salama, D., and El-Gohary, N. (2013). “Semantic text classification for supporting automated compliance checking in construction.” *J. Comput. Civ. Eng., accepted.*

Tsoumakas, G., and Katakis, I. (2007). “Multi-label classification: an overview.” *Int. J. of Data Warehousing and Mining*, 3(3), 1-13.

Vogrincic, S., and Bosnic, Z. (2011). “Ontology-based multi-label classification of economic articles.” *Comput. Sci. Inf. Syst.*, 8(1), 101-119.

Wijewickrema, C. M., and Gamage, R. (2013). “An ontology based fully automatic document classification system using an existing semi-automatic system.” *Proc. IFLA 2013 World Library and Information Congress*,Singapore.

Zhong, B. T., Ding, L. Y., Luo, H. B., Zhou, Y., Hu, Y. Z., and Hu, H. M. (2012). “Ontology-based semantic modeling of regulation constraint for automated construction quality compliance checking.” *Automat. Constr.*, 28(0), 58-70.

Zhou, P., and El-Gohary, N. (2014). “Semantic-based text classification of environmental regulatory documents for supporting environmental compliance checking in construction.” *Proc., 2014 Construction Research Congress,* Georgia Institute of Technology, Atlanta, GA.