Doctor of Philosophy

Automated Construction Specification Review based on Semantic Textual Analysis

August 2020

Department of Civil and Environmental Engineering
The Graduate School of
Seoul National University

Seonghyeon Moon

Automated Construction Specification Review based on Semantic Textual Analysis

A dissertation submitted to the Graduate School of
Seoul National University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

By
Seonghyeon Moon

July 2020

Approval Signatures of Dissertation Committee

Hyunsoo Lee

Hyun-soo Lee

Seokho Chi

MOONSEO PARK

Moonseo Park

Hyoungkwan Kim

Changwan Kim

Automated Construction Specification Review based on Semantic Textual Analysis

지도교수 지 석 호

이 논문을 박사학위논문으로 제출함

2020 년 6 월

서울대학교 대학원

건설환경공학부

문 성 현

문성현의 박사학위논문을 인준함

2020 년 7 월

위 원 장 이 현 수 (인)

부위원장 지 석 호 (인)

위 원 박 문 서 (인)

위 원 김 형 관 (인)

위 원 金 相 完

DEDICATION

To you, my

ACKNOWLEDGEMENT

I would like to express my gratitude to many people who have supported and encourage me to complete the doctoral dissertation.

First, I sincerely appreciate my supervisor, Prof. Seokho Chi for his invaluable support and trust. He invited me to construction management, and it was a great fortune to learn his passion and insight right next to him. I also deeply appreciate my committee members, Prof. Hyun-soo Lee, Prof. Moonseo Park, Prof. Hyounkwan Kim, and Prof. Changwan Kim for their insightful suggestions and guidance. I was able to improve the quality of my dissertation via in-depth discussions with them.

Next, I want to thank Prof. Duyon Kim and Prof. Bon-Gang Hwang for their sincere advice on my future research career. Besides, many thanks to Dr. Hyun-Chul Oh and Hee-Ung Lee for their insights from the industry, and Dr. Hyunjoong Kim and Dr. Ohchul Kwon for technical assistance related to programming. Without their help, I would not able to develop the model at all. Also, many thanks to Jeong Ah Lee, Clement Nicolas, Chanhyuk Park, Junghoon Kim, and Jeongbin Hwang for their help on data labeling, and Seonsu Mok, Daeyoun Won, Gyueun Lee, and Suho Ahn for participating in the verification experiments.

Thanks to C!LAB members. I loved my six years with you guys. Especially thanks to my neighbors, Dr. Soram Lim, Dr. Jinwoo Kim, Hwikyung Chun, and Taeyeon Chang. Above all, I want to express a large grateful to my best colleague, Gitaek Lee, for the countless meetings and the enormous time on discussion.

Also, I would like to thank my buddies, Donghoon Ji, Inseok Yoon, and Junyeong Heo. Our coffee-break discussions in Mug would be never forgotten. I pray for their efforts to shine in a close future.

Lastly, I would like to give the biggest appreciation to my beloved family, my father, Jonghoon Moon, my mother, Jeonghwa Son, and my only brother, Seongmin Moon for their endless love and encouragement. They were the driving force for me to finish my long journey of Ph.D.

Again, thank you all, my people.

ABSTRACT

Automated Construction Specification Review based on Semantic Textual Analysis

Seonghyeon Moon

Department of Civil and Environmental Engineering

The Graduate School of Seoul National University

The risk management of construction project requires a clear and objective understanding of construction specifications in early phases to ensure that the requirements are appropriate to the site environment. However, the review process is disturbed by the tight schedule of the bidding process, the insufficient number of available experts, and the large volume of contents (generally several thousand pages). Moreover, since the review process is mainly carried out based on human cognitive abilities, it takes considerable time as well as is vulnerable to errors, such as subjective interpretation, misunderstanding, and omitting of requirements. Despite the promising results of previous approaches to automate the process of analyzing construction documents and extracting useful information, they need technical improvements as not considering the semantic textual conflicts of different documents. Since every construction project provides individual specification and even updates the document periodically, the review process requires to analyze different documents that have different semantic features, such as different vocabulary, different sentence structures, and differently organized clauses. Addressing the semantic textual conflicts is challenging to automate the construction specification review process with a sufficient level of applicability and support the project risk management.

This dissertation aims to develop an automated construction specification review method via semantic textual analysis. First, the author developed a semantic construction thesaurus to understand different vocabulary of the specifications using Word2Vec embedding and PageRank algorithm. Second, the author recognized construction keywords of qualitative requirements from natural language sentences by developing a Named Entity Recognition (NER) model using Word2Vec embedding and the Bi-directional Long Short-Term Memory (Bi-LSTM) architecture with Conditional Random Field (CRF) layer. Third, the author proposed a relevant clause pairing model that identified the most relevant clause from the standard specification for every clause in the construction specification using Doc2Vec embedding and semantic similarity calculation. Eventually, the proposed method would provide a table of clauses, which includes the most relevant clause and the recognized keywords related to construction requirements.

First, to achieve the first research objective, the author analyzed the words that were similarly distributed within the sentence using the Word2Vec model and determined the pivot term for each closed network of converting words. After analyzing 346,950 words (i.e., 19,346 sentences) from 56 construction specifications, the construction thesaurus covered 208 word replacement rules. Second, to achieve the second research objective, the five information types (i.e., persons and organizations in charge, activities required, construction and installation items, quality standards and criteria, and relevant references) that are crucial in the risk management process were determined via in-depth collaboration with experienced contractors. Then, the NER model was developed with 4,659 labeled sentences, where the input was word vectors embedded by Word2Vec and the output was the word categories standing for the determined five information types. The model showed satisfactory results with an F1 score of 0.917 in classifying the word categories within the sentences. The robustness of the model was verified with 30 different sets of randomly split training and validation data. Third, to achieve the third research objective, the manually extracted text data of 2,527 clauses were embedded by Doc2Vec to utilize the semantic features in the pairing process. Then, clause relevance was calculated is based on the cosine similarity between the text vectors to identify the most relevant text. As a result, the relevant clauses were paired with the averaged accuracy of 81.8%.

To validate the proposed approaches, the author conducted experiments. The validation indicators included time efficiency, the accuracy of detecting erroneous provisions, and robustness to subjectivity. The experimental results outperformed the manual review process with reducing working hours, improving performances, and providing more consistent results. Also, the results demonstrated the necessity and practical usefulness of the proposed method for automatic specification review. By utilizing the automated method of semantic text comparison, the users can address the semantic textual conflicts of the specifications (i.e., different vocabulary, different sentence structures, and differently organized clauses), which enables an adequate review of the project requirements.

In conclusion, this dissertation developed the automated construction specification review method by analyzing the semantic textual properties. Particularly, the author identified the semantic textual conflict among construction specifications (i.e., different vocabulary, different sentence structures, and differently organized clauses) that cause difficulty in automating the review process. The author developed the machine learning-based NLP models to facilitate the automated construction specification review. To the best of the author's knowledge, this is the first attempt to handle semantic textual conflict in the field of construction document analysis. The developed method benefits to the contractors who review specifications in the early phases of the construction project, the field engineers who analyze the requirements during the construction phases, and the clients who write a new specification for a project. The proposed approaches enhance the applicability of automated construction specification reviews and can be quickly customized for other types of construction documents, including contract documents, non-conformance reports, accident reports, and inspection reports. Besides, the research would facilitate an in-depth understanding of diverse and complicated construction specifications as well as the review process of the document that could further bring opportunities for improvements in the areas of construction automation and risk management.

Keywords: Construction management; Risk management; Construction specification; Automated document review; Semantic Text Conflict; Natural language processing; Machine Learning

Student Number: 2015-21288

CONTENTS

Chapter 1. Introduction 1

1.1. Research Background 1

1.2. Problem Statement 6

1.3. Research Objectives 7

1.4. Research Process and Scope 11

1.5. Dissertation Outline 14

Chapter 2. Theoretical Background and Related Works 17

2.1. Construction Specification 18

2.2. Automated Text Analysis in Construction Industry 21

2.2.1. Document Interpretation 22

2.2.2. Provision Classification 23

2.2.3. Compliance Checking 25

2.3. Limitations of Previous Research 27

2.4. Summary 30

Chapter 3. Analysis of Construction Text Ambiguity 31

3.1. Research Method: Semantic Construction Thesaurus 33

3.1.1. Thesaurus 33

3.1.2. Text Embedding: Word2Vec 34

3.1.3. Word Weighting: PageRank...38

3.2. Data Preparation 40

3.2.1. Data Collection 40

3.2.2. Text Preprocessing...44

3.3. Development of Semantic Construction Thesaurus 49

3.3.1. Word Embedding 51

3.3.2. Pivot Term Determination based on Semantic Similarity ..53

3.4. Results of Semantic Construction Thesaurus 57

3.4.1. Results of Word Embedding 57

3.4.2. Semantic Word Similarity..59

3.4.3. Semantic Construction Thesaurus 61

3.5. Summary..64

Chapter 4. Qualitative Requirement Recognition on Construction Clauses 65

4.1. Research Method: Construction Keyword Recognition................67

4.1.1. Named Entity Recognition 67

4.1.2. Recurrent Neural Network...69

4.1.3. Bi-directional Long Short-Term Memory 72

4.1.4. Conditional Random Field...75

4.2. Development of NER Model for Construction Keyword Recognition...77

4.2.1. Data Labeling 77

4.2.2. NER Model Development 82

4.3. Results of Construction Keyword Recognition............................87

4.3.1. Results of Thesaurized Word Embedding 87

4.3.2. NER Model Validation 88

4.3.3. Evaluation of Impact of Thesaurus.....................................93

4.4. Summary...95

Chapter 5. Identification of Relevant Clauses from Different Construction Specifications 96

5.1. Research Method: Relevant Clause Pairing 98

5.1.1. Analyzed Unit of Text Relevance: Clause 98

5.1.2. Text Embedding: Doc2Vec 99

5.1.3. Cosine Similarity 102

5.2. Relevant Clause Pairing Framework 103

5.2.1. Development of Clause Corpus and Clause Embedding..104

5.2.2. Estimation of Semantic Relevance of Clauses 106

5.3. Results of Relevant Clause Pairing 107

5.3.1. Results of Clause Embedding..107

5.3.2. Identification of Relevant Clauses...................................108

5.4. Summary...124

Chapter 6. Experimental Results and Discussions 125

6.1. Experimental Design 126

6.2. Experimental Results 133

6.2.1. Review of QCS 2014 against QCS 2010..........................133

6.2.2. Review of QCS 2014 against Connecticut of United States 135

6.3. Evaluation of Automated Specification Review..........................138

6.4. Industrial Applications..142

6.5. Summary...144

Chapter 7. Conclusions 146

7.1. Achievements to Research Objectives..147

7.2. Contributions 151

7.3. Opportunities for Improvement and Future Research.................154

Bibliography 157

국문 초록 168

Appendix A. Research Prototype: DICCI............................172

A.1. UI Functions 173

A.2. Data Selection for Analysis 174

A.3. Selection of Analysis Type 178

A.4. Clause Pairing...179

A.5. Paragraph Pairing 181

A.6. Informative Keywords Extraction 183

LIST OF TABLES

Table 3.1 Data collection 40

Table 3.2 Data exploration..44

Table 3.3 Example of dictionary of similarly used words 49

Table 3.4 Hyperparameters of Word2Vec model..51

Table 3.5 Samples of Word2Vec embedding result 58

Table 3.6 Sample of similar word dictionary 59

Table 3.7 Word replacement rules of semantic construction thesaurus........62

Table 4.1 Personal information of consultants 78

Table 4.2 Identification of Information Types..79

Table 4.3 Examples of NER categories 81

Table 4.4 Hyperparameters of NER model 86

Table 4.5 Samples of thesaurized Word2Vec embedding result...................88

Table 4.6 Confusion matrix of NER results 90

Table 4.7 Classification performance for each category 91

Table 4.8 F1 scores of 30 randomly split data sets.......................................91

Table 5.1 Hyperparameters of Doc2Vec model..104

Table 5.2 Result of relevant clause pairing (‘Qatar_2014’ and ‘Qatar_2010’) 110

Table 5.3 Confusion matrix of clause pairing (‘Qatar_2014’ and ‘Qatar_2010’) 117

Table 5.4 Result of clause pairing (‘Qatar_2014’ and ‘United States_ Connecticut_2018’) 118

Table 5.5 Confusion matrix of clause pairing (‘Qatar_2014’ and ‘United States_Connecticut_2018’)...123

Table 6.1 Personal information of experiment participants........................127

Table 6.2 Experimental results of control group (‘Qatar_2014’ and ‘Qatar_ 2010’) 134

Table 6.3 Experimental results of experimental group (‘Qatar_2014’ and ‘Qatar_2010’) 135

Table 6.4 Experimental results of control group (‘Qatar_2014’ and ‘United States_Connecticut_2018’)...136

Table 6.5 Experimental results of experimental group (‘Qatar_2014’ and ‘United States_Connecticut_2018’) 137

Table 6.6 Evaluation indicators 138

LIST OF FIGURES

Figure 1.1 Manual review process of construction specification 5

Figure 1.2 Research framework..8

Figure 1.3 Automated review process of construction specification...........10

Figure 1.4 Research process 12

Figure 2.1 Origins of references in QCS 2014 19

Figure 3.1 Research process of semantic construction thesaurus................32

Figure 3.2 Application of thesaurus..33

Figure 3.3 Example of one-hot encoding 35

Figure 3.4 Word2Vec architecture: (a)CBOW, (b)skip-gram 37

Figure 3.5 PageRank algorithm..39

Figure 3.6 Example of data format conversion: (a)original PDF, (b)manually extracted TXT 43

Figure 3.7 Most frequent n-grams (n: 2 to 5) 45

Figure 3.8 Text preprocessing steps..48

Figure 3.9 Example of word link graph..50

Figure 3.10 Word2Vec embedding architecture 52

Figure 3.11 Sample of recursive word replacement....................................54

Figure 3.12 Word network of similar word dictionary................................60

Figure 3.13 Semantic construction thesaurus 61

Figure 4.1 Research process of construction keyword recognition 66

Figure 4.2 RNN model architecture 70

Figure 4.3 Example of word classification result: (a)conventional classification model, (b)RNN model 71

Figure 4.4 LSTM model architecture 72

Figure 4.5 Bi-LSTM model architecture 73

Figure 4.6 LSTM unit structure 74

Figure 4.7 CRF framework 76

Figure 4.8 NER model architecture 83

Figure 4.9 NER model framework 83

Figure 4.10 Sentence padding: (a)sentence less than 50 words, (b)sentence more than 50 words 85

Figure 4.11 Samples of NER results: (a)original text, (b)NER tagged text . 89

Figure 5.1 Research process of relevant clause pairing 97

Figure 5.2 Doc2Vec architecture: (a)PV-DM, (b)PV-DBOW 101

Figure 5.3 Relevant clause pairing framework 103

Figure 5.4 Doc2Vec embedding architecture 105

Figure 5.5 Sample of relevant clause pairing (the same clauses): (a)'Qatar_ 2014,' (b)'Qatar_2010' 115

Figure 5.6 Sample of relevant clause pairing (the similar clauses): (a)'Qatar_2014,' (b)'Qatar_2010' 116

Figure 5.7 Sample of relevant clause pairing: (a)'Qatar_2014,' (b)'United States_Connecticut_2018' 123

Figure 6.1 Sample of analyzed specification..126

Figure 6.2 Report cover of experimental result..129

Figure 6.3 Report format of experimental result..130

Figure 6.4 Example of experimental result 132

Chapter 1. Introduction

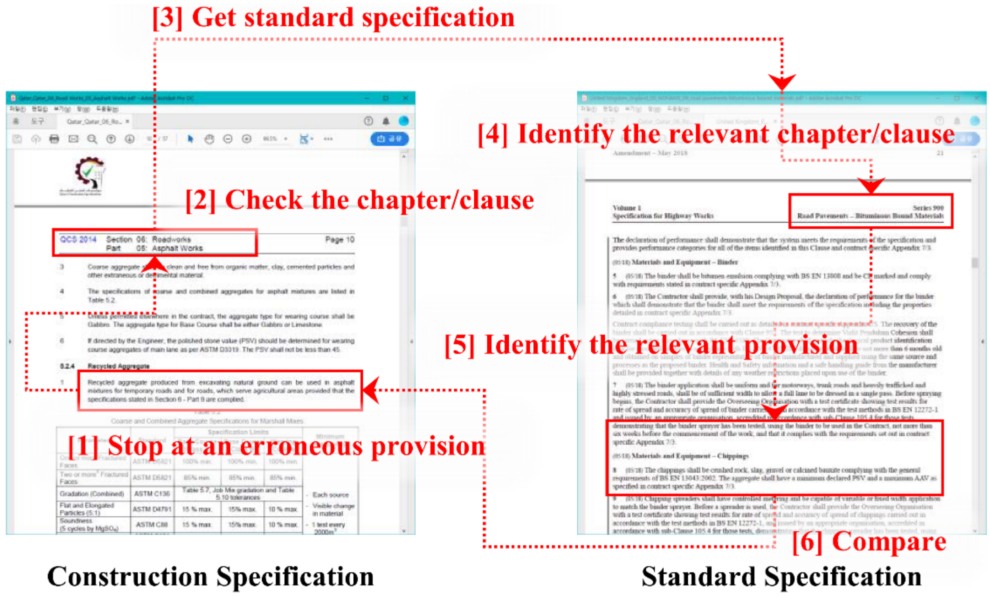
1.1. Research Background

A construction specification is a document that specifies the qualitative requirements for performing work in a construction project, which covers technical construction issues. Since the document is a legally binding contract, the contractors should follow the requirements thoroughly during the construction project (Ryoo et al. 2010). However, the contractors might face inappropriate provisions that require unrealistic standards and criteria because the requirements are commonly generated by the project client who lacks practical expertise. These erroneous provisions cause problems to the project, such as wasted resources due to design changes, increased risks of accidents due to construction errors or unsafe installation, and conflicts or lawsuits between stakeholders due to non-compliance (Zhang and El-Gohary 2017; Zhong et al. 2012). For example, the contractors of a road construction project in Qatar suffered from erroneous provisions. The construction specification provided from the client required inappropriate criterion for the asphalt aggregate mixture, which might suit to mild weather conditions rather than the hot and dry weather of Qatar. Besides, some of the requirements were even impractical or unrealistic for the site, where no materials and local engineers existed to meet the criteria. More than 700 cases of non-compliance issues occurred during the four-year project due to the inappropriate requirements, and the project finally encountered conflicts because of the non-compliances and project delay. If the contractor discovered the erroneous provisions in the earlier phases (e.g., planning or bidding), the provisions could be corrected, and the project risk might decrease.

A clear and objective understanding of construction specifications in the early phases of the project is particularly important for project risk management. The contractors should review the provisions thoroughly and analyze each requirement to ensure it is appropriate to the site environment. The field engineers should carefully examine the requirements of provisions before working in order to reduce the risk of low performance, safety accidents, and reworks. Even the client should refer to other specifications to discover erroneous provisions and deliver the requirements free from misconceptions (Lam et al. 2007). However, the review process is disturbed by the tight schedule of the bidding process, the insufficient number of available experts, and the large volume of contents (generally several thousand pages) (Lee and Yi 2017). Moreover, since the review process is mainly carried out based on human cognitive abilities, it takes considerable time as well as is vulnerable to errors, such as subjective interpretation, misunderstanding, and omitting of requirements.

Many researchers in the construction domain have attempted to automate document analysis processes based on natural language processing (NLP) and text mining techniques. The previous research parsed the construction documents to identify the informative instances, analyzed the relationship between the instances, and defined several rules and patterns to extract useful information from the text data (Xiao et al. 2018; Zhang and El-Gohary 2016; Zhong et al. 2020a). Despite the promising results under experimental conditions, the previous approaches included several critical limitations in terms of practical applicability. Since the approaches were based on the pre-defined rules and patterns, they required to develop a new model for every new data to analyze. The ability to process new data is one of the most significant requirements for the automated document analysis to be applied in practice since every construction project provides individual specification and even updates the document periodically.

The author analyzed the review process of construction specifications to understand the practical requirements for automation. In many cases, the construction specification has been compared to the standard specification that describes the national standards and criteria, which can be generally applied to the construction sites considering the environmental properties. If a contractor stops at an erroneous provision, he will check the chapter and clause to which the provision is relevant. Then, the contractor would identify the relevant chapter and clause from the standard specification to find the relevant provision. Finally, the paired relevant provisions should be compared to each other (Figure 1.1). Practically, every construction project provides individual specifications and even updates the document periodically; thus the practitioner should review different documents each time. As different specifications show different textual properties (e.g., different vocabulary, different sentence structures, and differently organized clauses), pre-defined rules and patterns would not be able to analyze the different documents (Moon et al. 2019). Therefore, in order to automate the review process, the methods should be able to (1) understand the different vocabulary, (2) recognize qualitative requirements from natural language sentences, and (3) identify the relevant provision of which topic is the same.


[3] Get standard specification
[4] Identify the relevant chapter/clause
[2] Check the chapter/clause
[5] Identify the relevant provision
[1] Stop at an erroneous provision
[6] Compare
Construction Specification
Standard Specification

QCS 2014 Section 06: Roadworks
Part 05: Asphalt Works
Page 10

3 Coarse aggregate shall be clean and free from organic matter, clay, cemented particles and other extraneous or detrimental material.

4 The specifications of coarse and combined aggregates for asphalt mixtures are listed in Table 5.2.

5 Unless permitted elsewhere in the contract, the aggregate type for wearing course shall be Gabbro. The aggregate type for Base Course shall be either Gabbro or Limestone.

6 If directed by the Engineer, the polished stone value (PSV) should be determined for wearing course aggregates of main lane as per ASTM D3319. The PSV shall not be less than 45.

5.2.4 Recycled Aggregate

1 Recycled aggregate produced from excavating natural ground can be used in asphalt mixtures for temporary roads and for roads, which serve agricultural areas provided that the specifications stated in Section 6 - Part 9 are complied.

Table 5.2
Coarse and Combined Aggregate Specifications for Marshall Mixes

Parameter	Standard	Specification Limits			Minimum
		[illegible]	[illegible]	[illegible]	[illegible]
One or more Fractured Faces	ASTM D5821	100% min.	100% min.	100% min.	- Each source - Visible change in material - 1 test every 2000m³
Two or more¹ Fractured Faces	ASTM D5821	85% min.	85% min.	85% min.	
Gradation (Combined)	ASTM C136	Table 5.7, Job Mix gradation and Table 5.10 tolerances			
Flat and Elongated Particles (5:1)	ASTM D4791	15 % max.	15% max.	10 % max.	
Soundness (5 cycles by $MgSO_4$)	ASTM C88	15 % max.	15% max.	10 % max.	

Amendment – May 2018 21

Volume 1
Specification for Highway Works

Series 900
Road Pavements – Bituminous Bound Materials

The declaration of performance shall demonstrate that the system meets the requirements of the specification and provides performance categories for all of the items identified in this Clause and contract specific Appendix 7/3.

(05/18) Materials and Equipment – Binder

5 (05/18) The binder shall be bitumen emulsion complying with BS EN 13808 and be CE marked and comply with requirements stated in contract specific Appendix 7/3.

6 (05/18) The Contractor shall provide, with his Design Proposal, the declaration of performance for the binder which shall demonstrate that the binder shall meet the requirements of the specification including the properties detailed in contract specific Appendix 7/3.

Contract compliance testing shall be carried out as detailed in contract specific Appendix 7/5. The recovery of the binder shall be carried out in accordance with Clause 95[illegible]. The test to determine Vialit Pendulum Cohesion shall [illegible] product identification [illegible] not more than 6 months old and obtained on samples of binder representative of binder manufactured and supplied using the same source and processes as the proposed binder. Health and Safety information and a safe handling guide from the manufacturer shall be provided together with details of any weather restrictions placed upon use of the binder.

7 (05/18) The binder application shall be uniform and for motorways, trunk roads and heavily trafficked and highly stressed roads, shall be of sufficient width to allow a full lane to be dressed in a single pass. Before spraying begins, the Contractor shall provide the Overseeing Organisation with a test certificate showing test results for rate of spread and accuracy of spread of binder carried out in accordance with the test methods in BS EN 12272-1 and issued by an appropriate organisation, accredited in accordance with sub-Clause 105.4 for those tests, demonstrating that the binder sprayer has been tested, using the binder to be used in the Contract, not more than six weeks before the commencement of the work, and that it complies with the requirements set out in contract specific Appendix 7/3.

(05/18) Materials and Equipment – Chippings

8 (05/18) The chippings shall be crushed rock, slag, gravel or calcined bauxite complying with the general requirements of BS EN 13043:2002. The aggregate shall have a minimum declared PSV and a maximum AAV as specified in contract specific Appendix 7/3.

9 (05/18) Chipping spreaders shall have controlled metering and be capable of variable or fixed width application to match the binder sprayer. Before a spreader is used, the Contractor shall provide the Overseeing Organisation with a test certificate showing test results for rate of spread and accuracy of spread of chippings carried out in accordance with the test methods in BS EN 12272-1, and issued by an appropriate organisation, accredited in accordance with sub-Clause 105.4 for those tests, demonstrating that the chippings spreader has been tested, using

Figure 1.1 Manual review process of construction specification

1.2. Problem Statement

The necessity of automation to review construction documents (e.g., construction specifications) has resonated with many researchers. Many researchers have been working on how to analyze construction documents and extract information automatically by applying NLP and text mining techniques. Despite the promising results under experimental conditions, the previous approaches need technical improvements to satisfy the practical requirements in automating the review process. As every construction project provides individual specification and even updates the document periodically, the review process requires to analyze different documents each time. Since the existing approaches are based on pre-defined rules and patterns, the user should develop a new model for every new data. The semantic textual properties of the different documents (i.e., different vocabulary, different sentence structures, and differently organized clauses) still hinder the field engineers from utilizing the existing automated approaches. Consequently, addressing the semantic conflicts is challenging to automate the review process of construction specifications with a sufficient level of applicability.

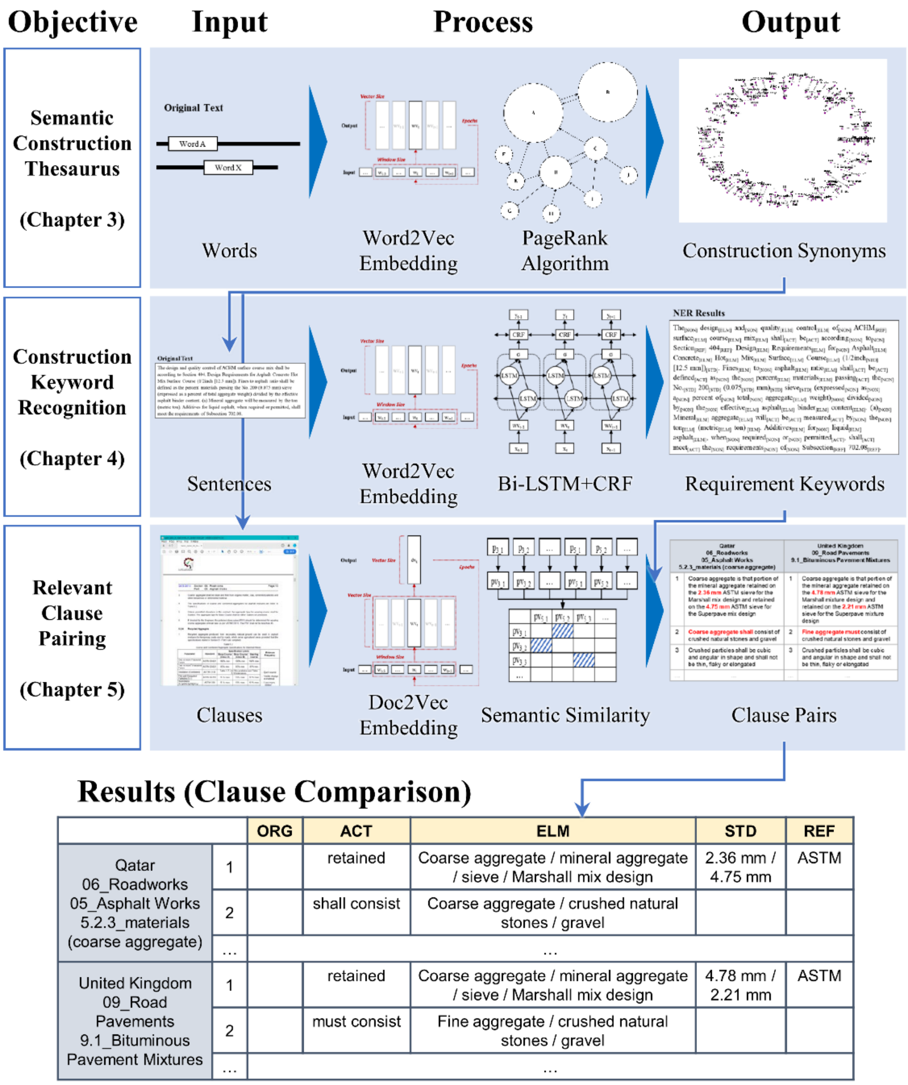
1.3. Research Objectives

This dissertation aims to develop an automated construction specification review method via semantic textual analysis. In order to address the practical limitations of existing approaches, the semantic conflicts (i.e., different vocabulary, different sentence structures, and differently organized clauses) of the construction specification are addressed with machine learning-based NLP models. The overall research framework is illustrated in Figure 1.2.

Objective 1: To develop the semantic construction thesaurus to understand different vocabulary of the specifications using Word2Vec embedding and PageRank algorithm.

Objective 2: To recognize construction keywords of qualitative requirements from natural language sentences based on the Named Entity Recognition (NER) model using Word2Vec embedding and the Bi-directional Long Short-Term Memory (Bi-LSTM) architecture with Conditional Random Field (CRF) layer.

Objective 3: To identify the most relevant clause from the standard specification for every clause in the construction specification using Doc2Vec embedding and semantic similarity calculation.



The[NON] design[ELM] and[NON] quality[ELM] control[ELM] of[NON] ACHM[REF] surface[ELM] course[ELM] mix[ELM] shall[ACT] be[ACT] according[NON] to[NON] Section[REF] 404[REF]. Design[ELM] Requirements[ELM] for[NON] Asphalt[ELM] Concrete[ELM] Hot[ELM] Mix[ELM] Surface[ELM] Course[ELM] (1/2inch[STD] [12.5 mm])[STD]. Fines[ELM] to[NON] asphalt[ELM] ratio[ELM] shall[ACT] be[ACT] defined[ACT] as[NON] the[NON] percent[ELM] materials[ELM] passing[ACT] the[NON] No.[STD] 200[STD] (0.075[STD] mm)[STD] sieve[STD] (expressed[NON] as[NON] a[NON] percent of[NON] total[NON] aggregate[ELM] weight)[NON] divided[NON] by[NON] the[NON] effective[ELM] asphalt[ELM] binder[ELM] content[ELM]. (a)[NON] Mineral[ELM] aggregate[ELM] will[ACT] be[ACT] measured[ACT] by[NON] the[NON] ton[ELM] (metric[ELM] ton) [ELM]. Additives[ELM] for[NON] liquid[ELM] asphalt[ELM], when[NON] required[NON] or[NON] permitted[ACT], shall[ACT] meet[ACT] the[NON] requirements[NON] of[NON] Subsection[REF] 702.08[REF].

Qatar 06_Roadworks 05_Asphalt Works 5.2.3_materials (coarse aggregate)		United Kingdom 09_Road Pavements 9.1_Bituminous Pavement Mixtures	
1	Coarse aggregate is that portion of the mineral aggregate retained on the 2.36 mm ASTM sieve for the Marshall mix design and retained on the 4.75 mm ASTM sieve for the Superpave mix design	1	Coarse aggregate is that portion of the mineral aggregate retained on the 4.78 mm ASTM sieve for the Marshall mixture design and retained on the 2.21 mm ASTM sieve for the Superpave mixture design
2	Coarse aggregate shall consist of crushed natural stones and gravel	2	Fine aggregate must consist of crushed natural stones and gravel
3	Crushed particles shall be cubic and angular in shape and shall not be thin, flaky or elongated	3	Crushed particles shall be cubic and angular in shape and shall not be thin, flaky or elongated
...	...	...	...

Results (Clause Comparison)

		ORG	ACT	ELM	STD	REF
Qatar 06_Roadworks 05_Asphalt Works 5.2.3_materials (coarse aggregate)	1		retained	Coarse aggregate / mineral aggregate / sieve / Marshall mix design	2.36 mm / 4.75 mm	ASTM
	2		shall consist	Coarse aggregate / crushed natural stones / gravel		
	...	...				
United Kingdom 09_Road Pavements 9.1_Bituminous Pavement Mixtures	1		retained	Coarse aggregate / mineral aggregate / sieve / Marshall mix design	4.78 mm / 2.21 mm	ASTM
	2		must consist	Fine aggregate / crushed natural stones / gravel		
	...	...				

Figure 1.2 Research framework

As a result, the output of each research objective can automate a large portion of the review process of construction specification. Specifically, the output of chapter 3 (i.e., the semantic construction thesaurus) would be applied generally over the document analysis to understand the different vocabulary. Next, the output of chapter 4 (i.e., the NER model for construction keyword recognition) would recognize every informative word from specification sentences, and the review process will compare clauses based on the recognized keywords. Lastly, the output of chapter 5 (i.e., the paired clauses of which topics are the same) would assist the review process by providing the most relevant clause for the given clause (Figure 1.3). It should be noted that the steps 1 (i.e., 'Stop at an erroneous provision') and 3 (i.e., 'Get standard specification') are not urgent to be automated since the steps account for the information need of the field engineers.

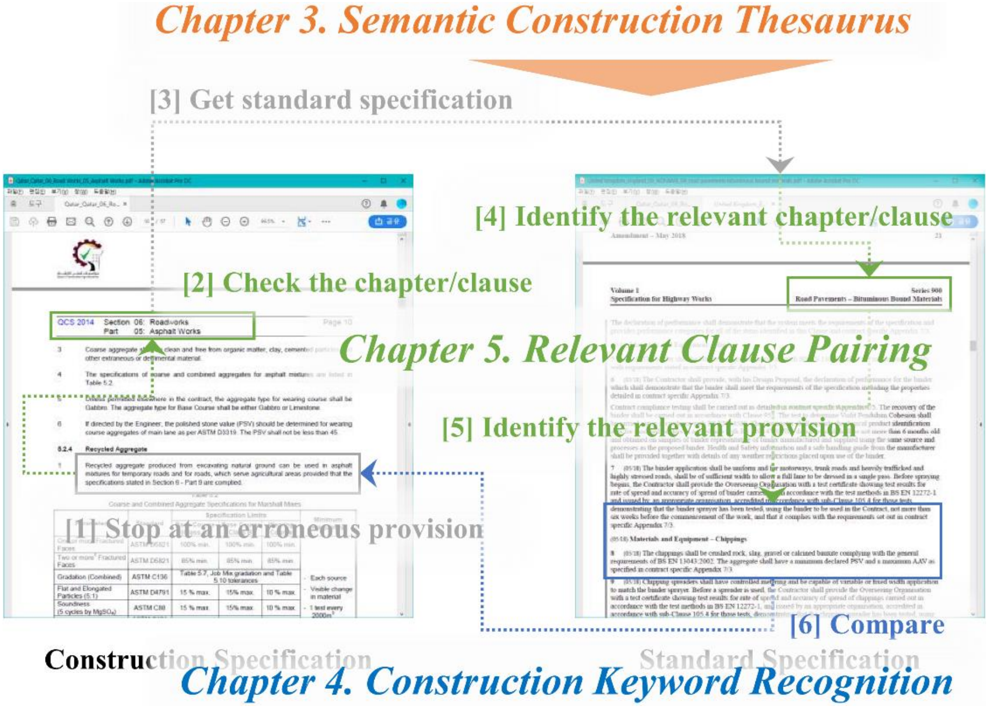

Chapter 3. Semantic Construction Thesaurus
[3] Get standard specification
[4] Identify the relevant chapter/clause
[2] Check the chapter/clause
QCS 2014 Section 06: Roadworks
Part 05: Asphalt Works
Volume 1
Specification for Highway Works
Series 900
Road Pavements – Bituminous Bound Materials
Chapter 5. Relevant Clause Pairing
[5] Identify the relevant provision
[1] Stop at an erroneous provision
[6] Compare
Construction Specification
Standard Specification
Chapter 4. Construction Keyword Recognition

Figure 1.3 Automated review process of construction specification

Eventually, the final result of the research would be the provision comparison tables that provide the recognized construction keywords for each clause pairs, as the below table of the research framework. The results can benefit the field engineers by automatically providing the information that is useful to specification review, such as which clause from the national standards is the most relevant for the erroneous clause, and whether the requirements of the paired clauses are the same or not.

1.4. Research Process and Scope

The author conducted the study under the following procedure (Figure 1.4). First of all, the author collected specifications in a PDF format and acquired text data as a TXT format (‘Data Collection’). Then, the text data was preprocessed (i.e., tokenization, stopwords removal, and lemmatization) with the most widely used techniques in NLP (‘Text Preprocessing’). Then, the preprocessed words were embedding to numeric vectors (‘Word Embedding’), and the semantic similarities between construction-related terms were calculated (‘Semantic Similarity Calculation of Construction Terms). The author grouped several words that are similar to each other and determined a pivot term for each grouped words to develop the semantic construction thesaurus (‘Pivot Term Determination’). The pivot term works as a representative of a word group; hence the computer can understand the other words in the group as the pivot term. After the author identified and assigned the informative categories that are needed to be recognized from the specification (‘Data Labeling’), the thesaurized words were embedded again (‘Thesaurized Word Embedding’). Next, the author proposed a NER model to recognize the construction keywords, which trained the thesaurized word vectors as input and the labeled data as output (‘NER Model Development’). Meanwhile, the author developed the clause corpus by extracting metadata of each clause (e.g., subtitles) and the relevant text sentences manually ('Development of Clause Corpus'). Every text data of each clause was embedded into numeric vector space ('Clause Embedding'), and relevant clauses were paired based on the similarity between clause vectors ('Relevant Clause Pairing'). Finally, the research outputs were utilized in comparing clauses automatically ('Comparative Analysis of Construction Clause'). The author evaluated the research by comparing the automated review results to the manual review results conducted by construction practitioners. The detailed descriptions and the properties of the method utilized in each research step are following in chapters 3 to 6.

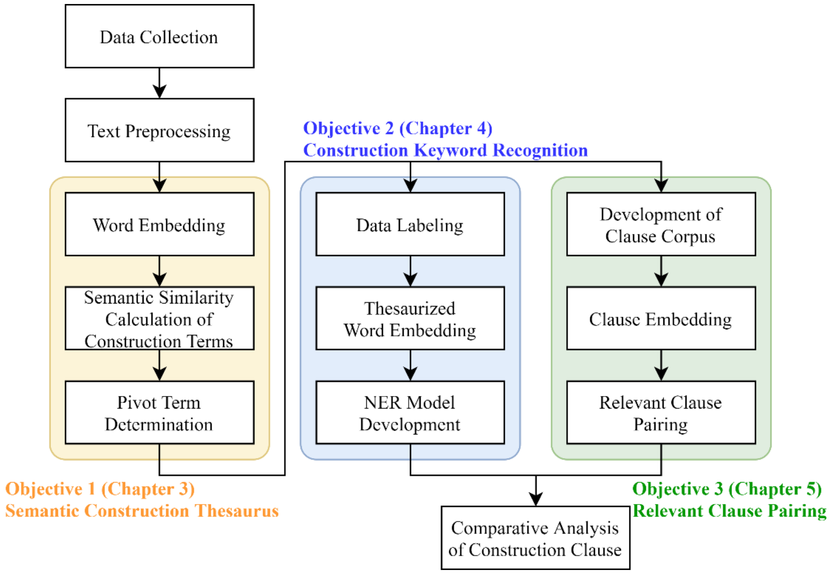


Figure 1.4 Research process

The research analyzed construction specifications written in English, the most commonly used language in international construction projects. Nevertheless, the proposed method can analyze the specification written in other languages since the language model applied in the current research learns the distributed information of input text instead of the shape of words or grammar. The proposed method's robustness to other languages is elaborated in the '3.3.1 Word Embedding' section.

The research analyzed the specifications related to road construction according to data availability. However, field engineers can utilize the proposed method regardless of the construction items or functional areas. As the author defined generic information types that are necessary for project risk management, every specification can be parsed based on these information types. The procedure of identifying the information types is discussed in more detail in the '4.2.1 Data Labeling' section.

1.5. Dissertation Outline

This dissertation consists of seven chapters. The content of each chapter is described below.

Chapter 1 Introduction: This chapter introduced the research background and motivations. The importance of construction specification review, the necessity of automation, and the limitations of the previous approaches are described. The objectives, scope, framework, and process of the research are also presented.

Chapter 2 Theoretical Background and Related Works: This chapter introduces the current review process of construction specifications via a case study of a construction project in Qatar. In addition, the previous attempts are reviewed, of which objectives were to automate document analysis to assist the manual review in the construction industry, focusing on the technical approaches and applications.

Chapter 3 Analysis of Construction Text Ambiguity: This chapter covers the first objective of this dissertation: developing the semantic construction thesaurus to understand the different vocabulary of the specifications. The concepts behind the thesaurus and PageRank algorithms are introduced with several examples from actual construction specifications. The semantic construction thesaurus is developed, and the results are discussed to consider whether the thesaurus is reasonable or not.

Chapter 4 Qualitative Requirement Recognition on Construction Clauses: This chapter covers the second objective of this dissertation: recognizing construction keywords of qualitative requirements from natural language sentences. The theoretical basis of NER architecture is introduced. The five information types which are crucial to the understanding of the critical contents of the specification are determined based on in-depth collaboration with experienced contractors. The NER model is developed to recognize the informative construction keywords automatically, and the development process and validation of the model are discussed.

Chapter 5 Identification of Relevant Clauses from Different Construction Specifications: This chapter covers the third objective of this dissertation: identifying the most relevant clause from the standard specification for every clause in the construction specification. The concept of relevant text pairing, which is proposed to identify the most relevant clause of which topic is the same, is introduced. The experimental results are followed by verification with the test dataset.

Chapter 6 Experimental Results and Discussions: Experimental Results and Discussion**:** In this chapter, the experimental design and process are described, results are presented, and the technical feasibility and in-practice applicability of this research are discussed. The reviewing results of the proposed approaches and the practitioners are compared in terms of time and cost efficiency, the accuracy of detection of erroneous provisions, and robustness to subjectivity.

Chapter 7 Conclusions: This chapter summarizes and discusses the research findings and contributions. Opportunities for further improvement and future research works are also discussed.

Chapter 2. Theoretical Background and Related Works

This chapter introduces the current review process of construction specifications via a case study of a construction project in Qatar. Besides, the previous attempts are reviewed, of which objectives were to automate document analysis to assist the manual review in the construction industry, focusing on the technical approaches and applications. The limitations of the previous approaches are also reviewed with examples.

2.1. Construction Specification

The author examined a construction specification and demonstrated the difficulty of the manual review process. Because of the data availability issue that the document is commonly confidential, the research could only investigate a construction specification that was utilized in an international construction project. The analyzed specification was QCS 2014 (Qatar Construction Specification 2014), a construction specification provided by the Qatari client in 2014 and used in a road construction project in Qatar. Since the national standard specification for road construction is absent in Qatar, the contractors reviewed the construction specification by comparing the requirements to those of the referred national standards from other regions. However, the QCS 2014 cited the standard specifications from diverse origins indiscriminately, which disturbed the contractors to review the appropriateness of provisions. The author counted every reference mentioned in the QCS 2014. The references amounted 12,995, including the UK for 5,024 of them (39%), the EU for 3,765 of them (30%), the USA for 2,491 of them (19%), the international standards for 1,196 of them (9%), and other sources for 519 of them (4%) (Figure 2.1).

It seemed logical that the document referred mostly to the UK as UK and Qatar are historical partners; however, some of the provisions were highly inappropriate to the site environment. Although Qatar has always exhibited a desert climate with the construction site being hot and dry, the provisions specifying the temperature of asphalt mixtures were referring to the national standard of the UK that presents a cold climate.

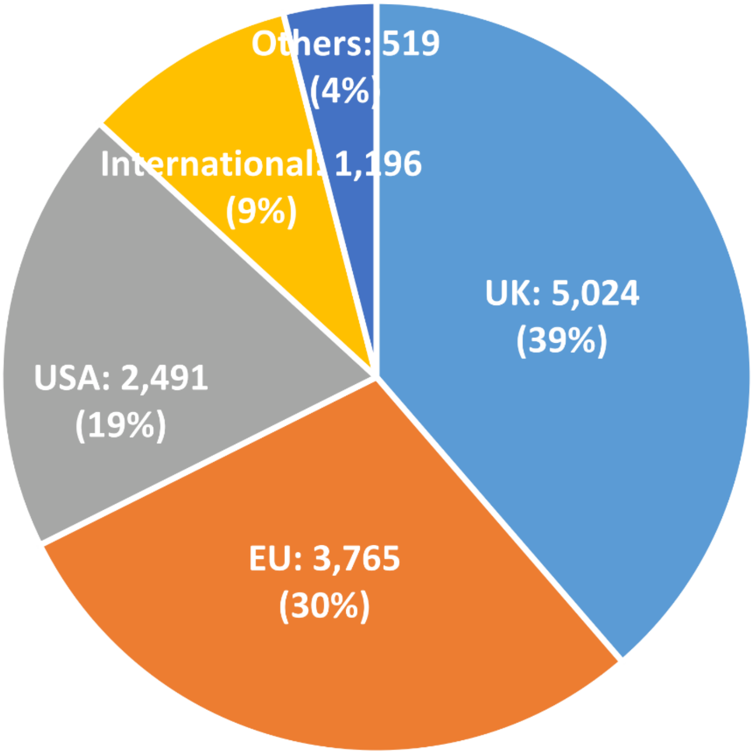


Figure 2.1 Origins of references in QCS 2014

In addition to the indiscriminate references, the large volume and complicated contents of the construction specification disturbed the review process. The QCS 2014 contained 4,790 pages, 29 major categories (i.e., chapters, such as "Concrete and Roadworks"), and 285 subcategories (i.e., subchapters, such as "Concrete Road Pavements, Concrete Plants, and Curing"). At the clause level, the categories were not mutually exclusive; for instance, the "Concrete Road Pavements" subchapter was located under the "Roadworks" chapter rather than the "Concrete" chapter. At the provision level, most sentences were too long to be understood – over four to five lines without a period – and the same construction elements or objects (e.g., coarse aggregate, curing temperature, and tack coat) were sometimes following different references. Thus, contractors may find it challenging to review, understand, and analyze all the contents of the specification manually, which might lead to overlooking errors or crucial issues and misinterpreting the provisions; this can probably result in unexpected conditions during actual construction in the field.

In reality, the contractors failed to detect erroneous provisions at the bidding stages because of the tight schedule of the bidding process, the insufficient number of available professionals, and the large volume of information. Consequently, the provisions eventually led to construction errors, which caused compensation almost equivalent to the initial cost of the project.

2.2. Automated Text Analysis in Construction Industry

In practice, the users have already manipulated the construction documents with information technologies, such as Optical Character Reader (OCR) and search function. The OCR recognizes characters by scanning the document images. The OCR contributes to the digitization of the construction industry where the documents are usually generated in hand-written since the recognized characters can be converted into text data. Once the documents are digitized, the users can search a word or phrase in the current document, and the locations of the same word or phrase would be provided.

Although these technologies have been widely used over the world for their efficiency in managing documents, the field engineers and contractors are still required to read, search, and understand the contents manually to review the text. Many researchers developed automated methods with natural language processing (NLP) to provide user needed information from not only the construction documents but also other contract documents from various industries (Kim and Chi 2019; Solihin and Eastman 2015; Zhang and El-Gohary 2014).

2.2.1. Document Interpretation

In the document level, the researchers attempted to catch the topics and interpret the documents to handle a large number of documents efficiently (Caldas and Soibelman 2003; Craig and Sommerville 2006; Kerrigan and Law 2005). They categorized the documents by construction items, such as materials, space, and physical boundaries, for which data were already available in a structured format. In other words, the results restricted to the items they focused on, and thus it was impossible to acquire data for items that were not listed. To develop more generic approaches, Al Qady and Kandil conducted a series of studies to develop document classification system based on the contents regardless of specific items (Al Qady and Kandil 2013a; b, 2015). They extract text features with Term Frequency-Inverse Document Frequency (TF-IDF) and trained the machine learning-based classifiers, including Rocchio, Support Vector Machine (SVM), k-Nearest Neighbor (kNN), and Naïve Bayes (NB). As a result, the 77 samples of construction documents were classified by topics. However, the construction specification review process required a more in-depth analysis of contents rather than just the topic

Lee and her research team attempted to predict the project feasibility by analyzing the potential risks from the bidding documents (Lee et al. 2016b; Lee and Yi 2017). They identified the type of uncertainty risks that frequently occurred in the text data and investigated 243 construction projects based on the risk types. The TF-IDF embedding and Latent Dirichlet Allocation (LDA) topic modeling were applied to extract the risk patterns from the projects. The famous classifiers, including Artificial Neural Network (ANN), SVM, kNN, and NB, were developed to classify the risk patterns presented in the data. Although the approaches to analyze the construction documents showed promising accuracies, they are inappropriate to detect the actual errors from the text. Since the analysis only focused on whether the project is feasible or not, the user needed more detailed information to review a provision is risky or not.

2.2.2. Provision Classification

To assist the risk management process practically at the provision level, the researchers have conducted analyses to detect the requirement texts from contractual documents (Le et al. 2019). They assigned the text with predetermined labels whether the statement is related to project requirements or not, and then utilized the individual statements as input and the labels as output. The results revealed that the model achieved a promising accuracy of over 90%. Moreover, several researchers proposed automated approaches to estimate the risk of the detected provisions. Some of them developed lexicon-based rules to classify the risk type of each sentence (Kim et al. 2020; Lee et al. 2019). They parsed the sentences to assign the syntactic or semantic tags to each word (e.g., 'subject,' 'relation,' and 'object') and utilized the tagged information for classification. Although these lexicon-based methods showed to be accurate in analyzing the risk of requirement texts, these approaches have a fundamental limitation in the aspect of applicability in practice. As the methods were only to be applied to the analyzed data, the user should build new lexicons, new types of tags, and new classification rules for every new analysis, which costs numerous time and human efforts.

To overcome the limitations of applicability, Zhong et al. (2020b) proposed a deep learning-based model to extract the procedural constraints from regulations (Zhong et al. 2020b). The researchers identified 13 constraint patterns between two temporal events (e.g., 'P1 before P2,' 'P1 during P2,' and 'P1 finish P2') and tried to extract the information by NER model that was developed based on the Bi-LSTM-CRF architecture. Although the research approach showed promising performance of the F1 score around 80%, the model was restricted to analyze the procedural constraints. Still, the developed models to investigate the provision risk is required further studies to be utilized in specification review. As every construction site has different environmental conditions, such provision that is classified as risky might be accepted as moderate according to the sites. Therefore, the models force the field engineers to investigate every risky provision.

2.2.3. Compliance Checking

Automatic compliance checking (ACC) is a process of automatically assessing the compliance of construction documents with applicable laws and regulations, which can address the limitation of provision risk classification approaches. ACC facilitates a detailed review of construction documents in the provision level and provides the information that what is the problem in which part of the text. Salama and El-Gohary (2013), one of the pioneers, proposed an approach for analyzing the provisions of laws and regulations in construction while understanding semantic information. Afterward, the results were enhanced by advanced applications of information transformation (Zhang and El-Gohary 2015), rule-based information extraction (Zhang and El-Gohary 2016), and ontology/deontology (Zhang and El-Gohary 2017). Eventually, a fully automated system for ACC was developed and tested with the International Building Code (Zhang and El-Gohary 2018).

However, these studies showed a fundamental improvement opportunity; the information extraction rules were built manually. In other words, the researchers with adequate domain knowledge read text data, extracted common features, determined the patterns from the features, and built the rules by listing up the patterns within the provisions for analysis. The extracted features, for example, included syntactic patterns (e.g., “subject,” “subject restriction,” “quantitative relation,” and “quantity value”) and conformance information (e.g., “obligation,” “permission,” “prohibition,” and “forbidden”). Several rules were formulated by mixing these patterns, and the rules were used to extract information from the documents (Salama and El-Gohary 2013; Zhang and El-Gohary 2016). Such approaches showed good performance when the data size is limited to a small volume; however, the longer and the more complicated a document, the more rules were required.

2.3. Limitations of Previous Research

The previous research still showed opportunities for technical improvement to automate the review process of construction specifications. Due to the different vocabulary of different specifications, the models might not be able to understand the different words that indicate the same object. Besides, since the sentence structures are different among specifications and even among chapters of the same document, the user would be required to identify new information extracting rules for new text. Moreover, the approaches cannot pair the relevant clauses from the different text, as the models understand the clauses based on the pre-defined lexicons and patterns. Consequently, despite the novelty of the existing studies, they still exist several constraints. The author demonstrated the limitations of existing studies and technical requirements for automating the review process of construction specification with several examples below.

First, as every construction project provides a unique construction specification, the vocabulary might vary among the documents, even among the chapters in the same document. For example, a construction specification might call the asphalt as “asphalt,” while another specification calls it “bituminous.” The contractors can interpret the two terms equally, whereas the computers might suffer. In another case, a general automated NLP model would tokenize a sentence “The Engineer should confirm the Job Mix Formula …” to [“the,” “engineer,” “should,” “confirm,” “the,” “job,” “mix,” “formula,” …]. There the intrinsic information of the “Job Mix Formula” under the construction industry vanished, and the text analysis might regard the term “job” as a work or a particular task. If the NLP model utilized a thesaurus with the term “Job Mix Formula,” the tokenized result can conserve the specific information of the term. These misunderstandings of provisions would make the results of the automated construction specification review vulnerable to errors. Since the previous approaches did not consider the different vocabulary and only analyzed a small population of text (i.e., one or two chapters), the models could not learn the conflict of vocabulary.

Second, the inconsistencies of sentence structure (i.e., writing style) are very crucial in discovering the erroneous provision as well as comparing two provisions automatically. Since every construction project provides a unique construction specification, the sentence structure might differ in each other. However, the previous approaches did not consider the sentence structures and developed the information extraction rules that are only applicable to the analyzed text data that shares similar sentence structures. Therefore, the results cannot be applied to other documents or even other chapters that have different sentence structures.

Lastly, the analyzed text units (i.e., document, sentence, and word) are inappropriate to text comparison for specification review. Particularly, the document-level studies analyzed the whole text all at once; hence they did not decompose the document into subcategories such as chapters, clauses, and provisions. On the other hand, the sentence-level and word-level studies analyzed a specific category from the document. Therefore, the results would be incapable of distinguishing the subject of each clause semantically (e.g., ‘Tack Coat’ and ‘Prime Coat’). Besides, the results would be incapable of matching two texts that came from different clauses (e.g., the ‘Materials’ clause in a specification and the ‘Asphalt Pavement’ clause in another).

2.4. Summary

In this chapter, the author introduced the current review process of construction specification via a case study of a construction project in Qatar. The contractors have encountered the difficulty of detecting erroneous provisions at the bidding stages because of the tight schedule, the inadequate number of experienced practitioners, and a large number of documents. Despite several information technologies existing in practice to manipulate the construction documents (e.g., OCR and search function), the field engineers and contractors are still required to read, search, and understand the contents manually to analyze the appropriateness of provisions. The previous research attempts to automate the document analysis process using NLP were reviewed, and the properties of construction specifications that limit the previous approaches were pointed out. First, the vocabulary varies among the specifications. Second, the provisions are described differently. Lastly, the analyzed text units (i.e., document, sentence, and word) are inappropriate to text comparison for specification review.

Chapter 3. Analysis of Construction Text Ambiguity

This chapter covers the first objective of this dissertation, which is to develop the semantic construction thesaurus to facilitate the computers to understand the text regardless of the different vocabulary. As described in the '2.3 Limitations of ' section, the different vocabulary of the different documents causes misunderstanding of provisions during the text comparison task for automated specification review. The author addressed the problem by developing a thesaurus that is domain-specific to the construction industry. The thesaurus represents a dictionary of replaceable words in the text of a specific domain. The proposed methods learned the usage patterns of each term based on the Word2Vec model and built a dictionary (i.e., hash list) of similarly used terms using Cosine similarity between the word vectors. To handle several terms that were recursively converted to each other, the author determined pivot terms based on the PageRank algorithm. Finally, the thesaurus was improved via the cooperation of the experienced construction experts (Figure 3.1).

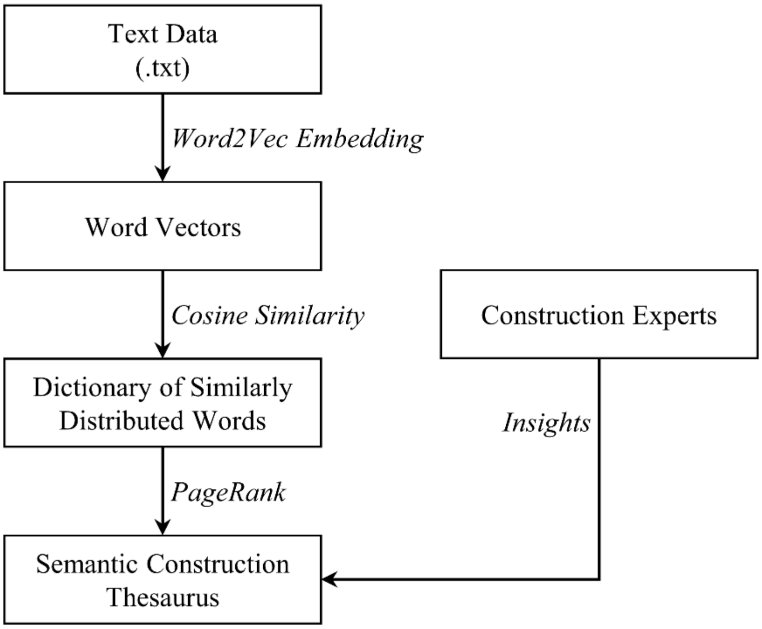
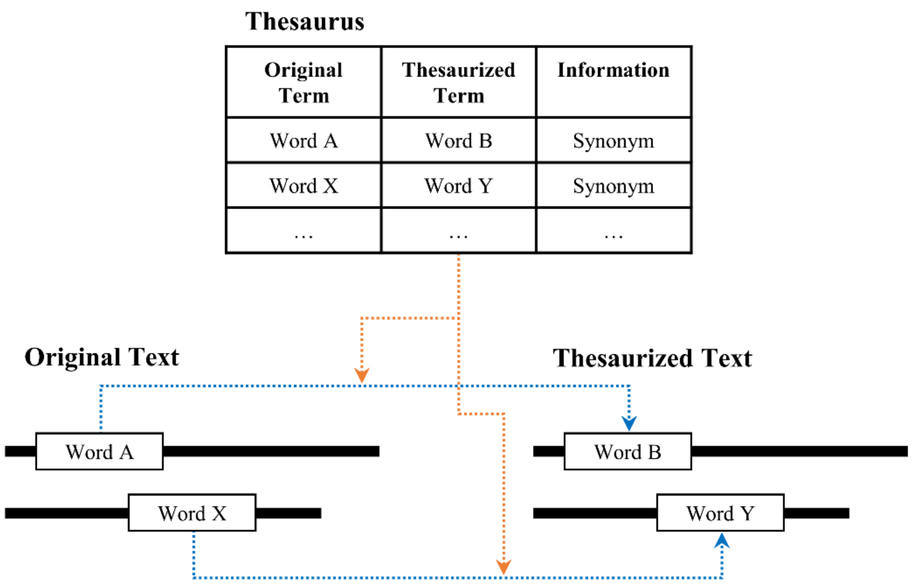


Figure 3.1 Research process of semantic construction thesaurus

3.1. Research Method: Semantic Construction Thesaurus

3.1.1. Thesaurus

Thesaurus is a dictionary that describes the relationship between terms including synonym, hypernym, and hyponym, rather than the definition of the term (Aitchison et al. 2003; Curran and Moens 2002; Jing and Croft 1994; Wielinga et al. 2001). Many of the information retrieval techniques have developed the thesaurus to expand the query and provide extensive search results (Zou et al. 2017). As the thesaurus replaces every term in the text to the pivot form that is listed itself, the text analysis would be expected to reflect the analysis intention of the user (Figure 3.2).



Original Term	Thesaurized Term	Information
Word A	Word B	Synonym
Word X	Word Y	Synonym
…	…	…

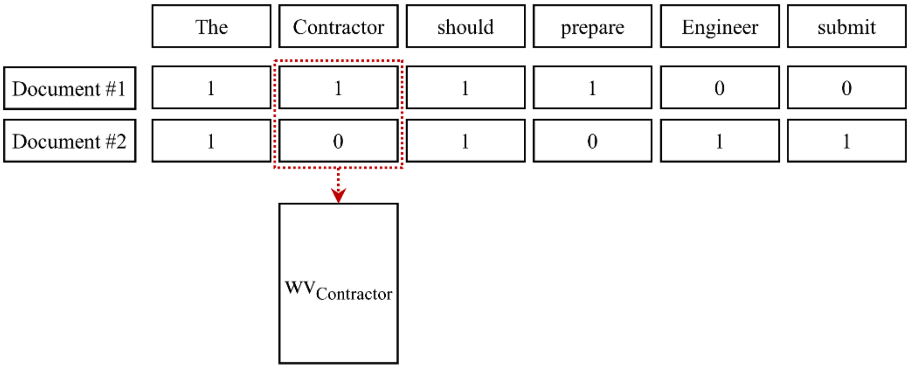
Figure 3.2 Application of thesaurus

The thesaurus is commonly built by the industrial professionals who have a profound knowledge in the terminology of a specific domain (Kim and Chi 2019) or by the existing synonym dictionaries (Zhang and El-Gohary 2015). However, these approaches might be restricted to prevalent and refined information. In this dissertation, the author advanced the construction thesaurus to be developed automatically to consider new vocabulary from new documents. Word2Vec model analyzed the usage patterns of the words (i.e., word distribution in the sentence), and figured out the semantic relationships of similarly used words. As the proposed method utilized the actual text sentences for developing the semantic construction thesaurus, the thesaurus can be applied easily to other documents.

3.1.2. Text Embedding: Word2Vec

The text data are written in the form of natural language, which the computer cannot analyze. As the computer requires numeric data as input, an embedding process is essential in text analysis. The embedding process involves mapping the data from the natural language space to the numeric vector space (Manning et al. 2008; Mikolov et al. 2013). Every set of text data would have its own location in the vector space after embedding, which means the text would be represented in a numeric vector that the computer could utilize. In other words, the text embedding process corresponds to feature extraction in machine learning model development.

One-hot encoding is the simplest and most intuitive text embedding technique, which counts each word as a unique symbol regardless of the meaning or linguistic property. For example, if only two sentences, “The Contractor should prepare” and “The Engineer should submit,” exist in the text data, the one-hot encoding would embed each word to a sparse vector with zero and one (Figure 3.3). Although this approach outstands for efficiency in the embedding process, the results with the sparse vectors impose computational cost as well as interpretational challenges.



	The	Contractor	should	prepare	Engineer	submit
Document #1	1	1	1	1	0	0
Document #2	1	0	1	0	1	1

Figure 3.3 Example of one-hot encoding

Recently, machine learning-based embedding techniques have gained popularity for addressing the limitations of the frequency-based approaches.

Among the state-of-the-art techniques, the most widely used text embedding technique is Word2Vec, developed by (Mikolov et al. 2013). Word2Vec learns the distributed representation of words within every sentence (i.e., usage patterns) and maps similarly used words to close vector space. For example, with the two sentences above, Word2Vec would locate “Contractor” and “Engineer” in a close vector space according to the similar distribution of adjacent words (i.e., “the” and “should”).

There are two architectures of Word2Vec: Continuous Bag of Words (CBOW) and skip-gram (Le and Mikolov 2014; Mikolov et al. 2013). CBOW tries to predict the current word from the surrounding words by controlling a specific size of the window (i.e., the number of surrounding words) (Figure 3.4(a)). For example, with the sample sentence above, “the” and “should” are the surrounding words of “Contractor” with a window size of 1. The CBOW model finds the most robust projection matrix that receives n-surrounding word vectors and predicts the target word vector (i.e., Wordt in the figure) by adjusting the window size and the projection matrix repeatedly. On the other hand, the skip-gram model tries to predict the surrounding words based on the current word (Figure 3.4(b)). According to the developer’s note, the CBOW is faster while the skip-gram infers text better (Google Code Archive 2013). This study developed the Word2Vec model with the skip-gram architecture for text embedding, since the embedding quality would significantly affect the performance of the NER model.

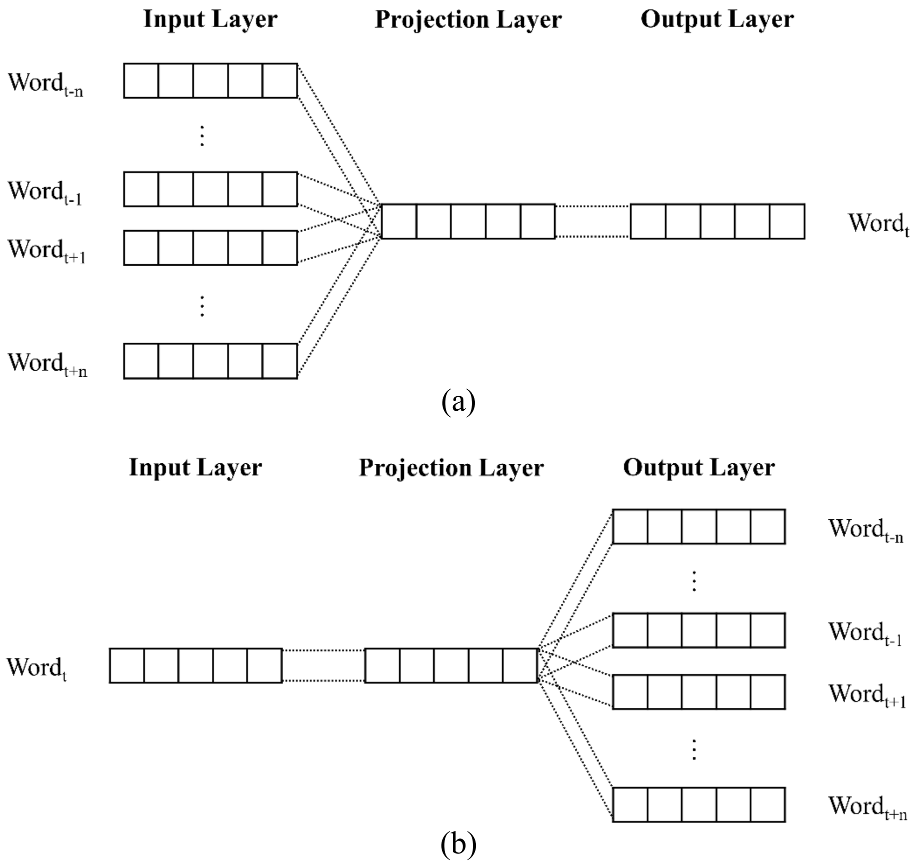


Figure 3.4 Word2Vec architecture: (a)CBOW, (b)skip-gram

3.1.3. Word Weighting: PageRank

PageRank is an algorithm that the search engine of Google uses, which assigns the weight for each document based on the relative importance to the other linked documents. The weighting algorithm is based on the logic that the more critical document gets the more number of inflows from other documents (Figure 3.5) (Kleinberg 1999; Page et al. 1999). Each circle indicates a document (i.e., web page), and the size of each circle indicates the importance of the document. The arrows indicate that the origin includes a hyperlink to the destination. The document 'A' would be considered very important since it gets all of the hyperlinks from the document 'C,' 'D,' and 'E,' which are hyperlinked by lots of other documents. Besides, since the document 'B' is interconnected with 'A,' it is almost as important as the document 'A.'

The author adopted the PageRank algorithm to discover the critical terms (i.e., the pivot terms) among the similarly used terms. The pivot term indicates a term that is apprehended to be a good alternative for many other terms (i.e., massive inflow) but offers only a few alternatives to be replaced (i.e., small outflow). Each term stands for a document in the PageRank algorithm, where the similarity represents the hyperlinks.

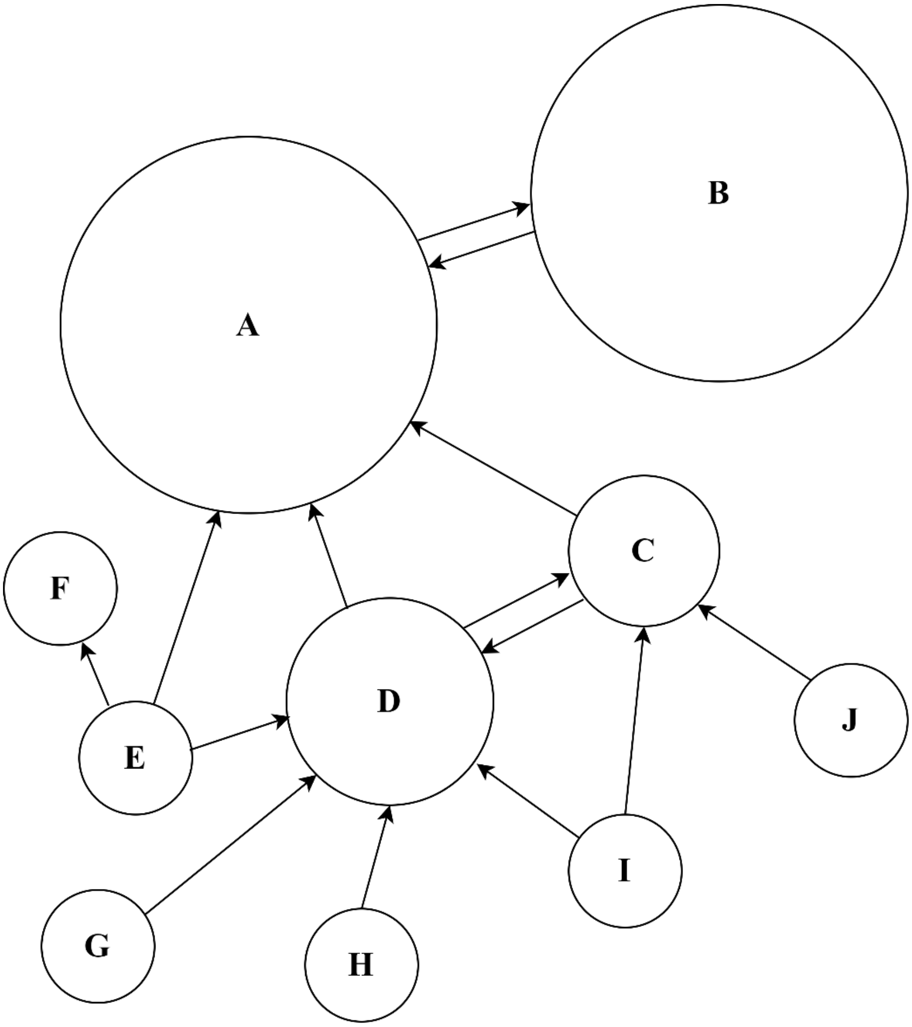


Figure 3.5 PageRank algorithm

3.2. Data Preparation

3.2.1. Data Collection

A total of 56 construction specifications were collected for analysis; two of them were practical specifications used in construction projects performed by Korean contractors: road construction projects in Qatar (2010 and 2014). The remaining 54 specifications were national (or regional) standards, which could be fundamentally applied to any road construction project in the country (or region). Since most developed countries have the standard specifications well organized, the research team preferentially collected the latest specifications from the United States of America (USA), the United Kingdom (UK), Canada (CAN), and Australia (AUS). The specifications were collected from national or government websites (Table 3.1).

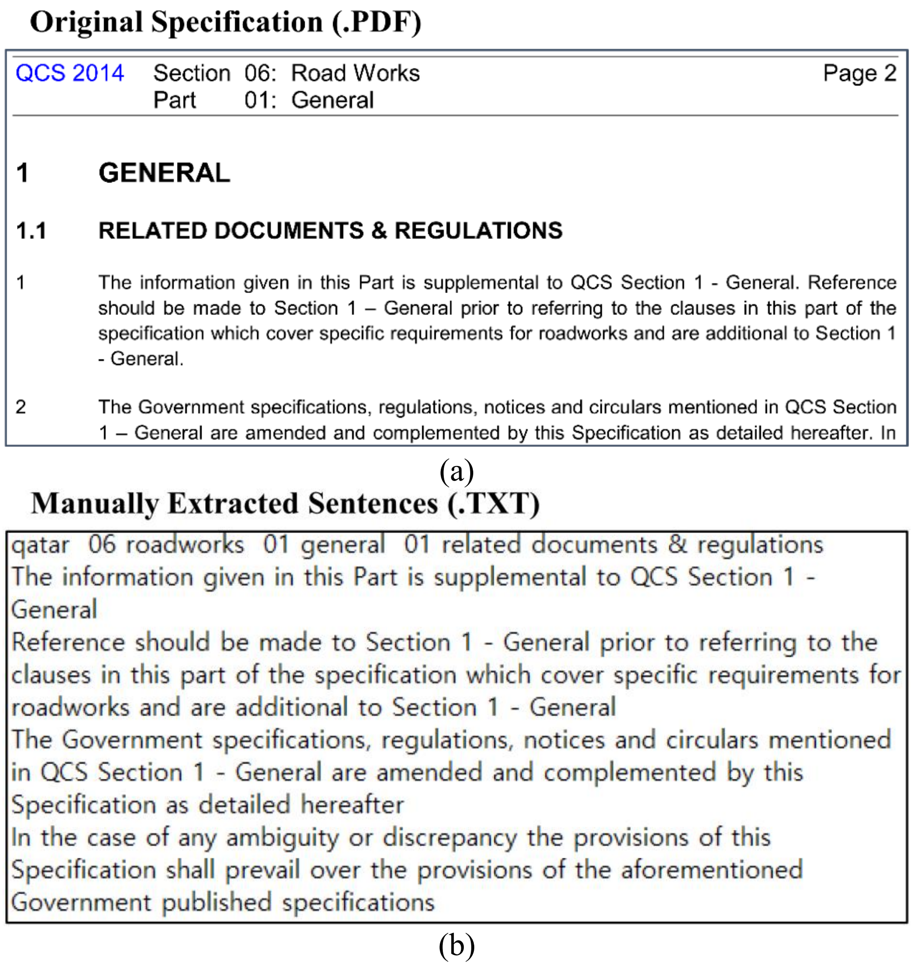
Table 3.1 Data collection

Index	Country	State	Last Edited	URL
1	Qatar	-	2010	-
2		-	2014	-
3	USA	Alabama	2018	https://www.dot.state.al.us/
4		Alaska	2017	http://www.dot.state.ak.us/
5		Arizona	2008	https://azdot.gov/
6		Arkansas	2014	https://www.arkansashighways.com/
7		California	2015	http://www.dot.ca.gov/
8		Colorado	2017	https://www.codot.gov/
9		Connecticut	2018	http://www.ct.gov/

10		Delaware	2016	https://deldot.gov/
11		Florida	2018	http://www.fdot.gov/
12		Georgia	2013	http://www.dot.ga.gov/
13		Idaho	2018	https://itd.idaho.gov/
14		Indiana	2018	https://www.in.gov/
15		Kentucky	2012	https://transportation.ky.gov/
16		Louisiana	2016	http://wwwsp.dotd.la.gov/
17		Maine	2014	https://www1.maine.gov/
18		Maryland	2008	http://roads.maryland.gov/
19		Massachusetts	1995	https://www.mass.gov/
20		Michigan	2012	https://www.michigan.gov/
21		Minnesota	2018	http://www.dot.state.mn.us/
22		Mississippi	2017	http://sp.mdot.ms.gov/
23		Missouri	2017	http://www.modot.org/
24		Montana	2014	https://www.mdt.mt.gov/
25		Nevada	2014	https://www.nevadadot.com/
26		New Hampshire	2016	https://www.nh.gov/
27		New Jersey	2007	http://www.newjersey.gov/
28		New Mexico	2014	http://dot.state.nm.us/
29		New York	2018	https://www.dot.ny.gov/
30		North Carolina	2018	https://connect.ncdot.gov/
31		North Dakota	2014	https://www.dot.nd.gov/
32		Ohio	2018	http://www.dot.state.oh.us/
33		Oklahoma	2009	https://ok.gov/
34		Oregon	2018	https://www.oregon.gov/
35		Pennsylvania	2016	https://www.penndot.gov/
36		Rhode Island	2013	http://www.dot.ri.gov/
37		South Dakota	2015	http://www.sddot.com/
38		Tennessee	2015	https://www.tn.gov/
39		Texas	2014	https://www.txdot.gov/
40		Utah	2017	https://www.udot.utah.gov/
41		Vermont	2018	http://vtrans.vermont.gov/
42		Virginia	2016	http://www.virginiadot.org/
43		Washington	2018	https://www.wsdot.wa.gov/
44		West Virginia	2017	https://transportation.wv.gov/
45		Wyoming	2010	http://www.dot.state.wy.us/
46	UK	England	2018	https://www.gov.uk/
47	CAN	Alberta	2013	http://www.transportation.alberta.ca/
48		British Columbia	2016	https://www2.gov.bc.ca/gov/
49		New Brunswick	2015	https://www2.gnb.ca/

50		Newfoundland and Labrador	2013	https://www.tw.gov.nl.ca/
51		Nova Scotia	2014	https://novascotia.ca/tran/
52		Ontario	2018	http://www.raqsa.mto.gov.on.ca/
53		Prince Edward Island	2019	https://www.princeedwardisland.ca/
54	AUS	Northern Territory	2017	https://dipl.nt.gov.au/
55		Tasmania	2017	https://www.cbos.tas.gov.au/
56		Western Australia	2018	https://www.mainroads.wa.gov.au

As the collected specifications were in the PDF format, so the computer could not modify or analyze the contents. Therefore, every PDF file was converted into the TXT format, which allows the text to be modified (Zou et al. 2017). Initially, the author utilized open-source software products for automatic file conversion, including “pdftotext” of XpdfReader developed by Blyph & cog, LLC, an online platform “https://pdftotext.com/,” and Python library “pdftotext” (Noonburg 2017). However, the conversion results were inadequate for further text analysis because the converted TXT files included too many incorrect text recognitions, space errors, unnecessary punctuation marks, and meaningless symbols (Figure 3.6(a)). As the quality of the data had an immediate effect on the analysis results, the conversion processes were performed manually, i.e., the drag-copy-paste approach was conducted to extract every sentence from the PDF file one at a time (Figure 3.6(b)). As a result, a total of 2,527 clauses (i.e., 19,346 sentences) was prepared for analysis, which was from six regions, including two construction specifications (i.e., QCS 2010 and QCS 2014) and four national standard specifications, which are comparable to each other (Table 3.2).



Original Specification (.PDF)

QCS 2014 Section 06: Road Works Page 2
Part 01: General

1 GENERAL

1.1 RELATED DOCUMENTS & REGULATIONS

1 The information given in this Part is supplemental to QCS Section 1 - General. Reference should be made to Section 1 – General prior to referring to the clauses in this part of the specification which cover specific requirements for roadworks and are additional to Section 1 - General.

2 The Government specifications, regulations, notices and circulars mentioned in QCS Section 1 – General are amended and complemented by this Specification as detailed hereafter. In

(a)

Manually Extracted Sentences (.TXT)

qatar 06 roadworks 01 general 01 related documents & regulations
The information given in this Part is supplemental to QCS Section 1 - General
Reference should be made to Section 1 - General prior to referring to the clauses in this part of the specification which cover specific requirements for roadworks and are additional to Section 1 - General
The Government specifications, regulations, notices and circulars mentioned in QCS Section 1 - General are amended and complemented by this Specification as detailed hereafter
In the case of any ambiguity or discrepancy the provisions of this Specification shall prevail over the provisions of the aforementioned Government published specifications

(b)

Figure 3.6 Example of data format conversion: (a)original PDF, (b)manually extracted TXT

Table 3.2 Data exploration

Country	State	Last Edited	Number of Clauses	Number of Sentences
Qatar	-	2010	462	4,786
Qatar	-	2014	611	7,097
Australia	Tasmania	2017	181	1,181
UK	England	2018	528	3,940
USA	Alabama	2018	475	2,466
USA	Arkansas	2014	208	1,175

3.2.2. Text Preprocessing

Text preprocessing is one of the essential steps in NLP, which handles the text data in natural language to ensure its quality. The preprocessing consists of tokenization, stopword removal, and stemming (or lemmatization), according to the purpose of analysis (Manning et al. 2008; Zou et al. 2017).

Tokenization parses the sentence to a sequence of words to utilize each word as a minimum unit for the analysis. For example, the sentence "The specification includes provisions about a technical requirement" would be tokenized into eight words, i.e., "the," "specification," "includes," "provisions," "about," "a," "technical," and "requirement," and the model might treat the sentence as a group of those eight tokens. This research also applied a multi-gram (i.e., n-grams) approach to tokenization, which counts several adjacent tokens that frequently occurred in the text simultaneously;

for example, “should be,” “job mix formula,” and “asphalt plant.” The multi-gram tokenization is known for its effectiveness in downsizing the feature dimension of the data, which improves the quality of text analysis results (Joulin et al. 2016; Wang and Manning 2012). The author counted n-grams that occurred at least ten times, with n of 2 to 5 over the whole documents (Figure 3.7). The most frequent n-gram was “shall be” with 10,285 times of occurrence, and the number of n-grams that occurred at least ten times was counted as 13,948, which showed an extremely long tail. The author designated the minimum occurrence of n-gram as 100, which returns the top of 481 n-grams since too infrequent n-grams might make confuse the meaning of individual words adversely.

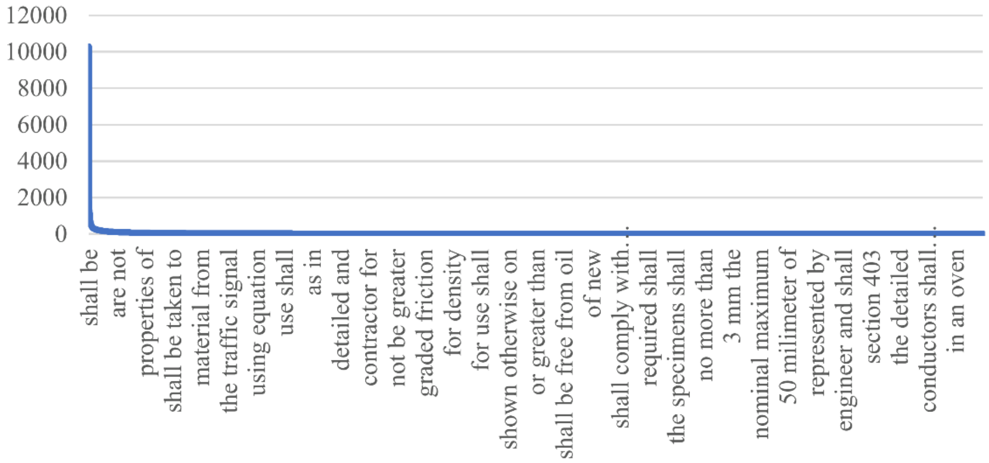


Figure 3.7 Most frequent n-grams (n: 2 to 5)

Stopword removal is to remove every stopword in the text, which indicates such word that is not necessary for text analysis since the word occurs in most of the documents and thus plays a small role as a feature of a specific document. For example, when analyzing two sentences, "The Contractor should prepare" and "The Engineer should submit," the word "the" is needless to distinguish the sentences, which would be a stopword. The NLTK (Natural Language Toolkit), which is the most commonly used python packages to process and analyze natural language data, provides a stopword list that can be utilized in general situations. However, the list includes lots of false-positive stopwords such as the modals (e.g., "should," "shall," and "must," prepositions (e.g., "before," "after," and "between"), conjunctions (e.g., "while," "until," and "than"), determiners (e.g., "all," "one," and "any"), adjectives (e.g., "same," "equal," and "further"), and adverbs (e.g., "once," "off," and "over"). Although these words might be regarded as less valuable in common cases, they are the cores of the qualitative criteria of the provisions, which are indispensable in construction specification analysis. Therefore, the author customized the stopword list only to include "a," "be," "is," "are," "was," "were," "the," "this," "these," "that," "those," and "of."

Stemming converts every word to its stem (i.e., a base part of the word), which never changes when the ending of the word changes. For example, the stem of “pave,” “pavement,” “paving,” and “paved” would be “pav.” The stem is a grammatical element that might not have any meaning, which indicates a shared part of a group of words. The stemming facilitates the text analysis to consider the words in different forms but have an identical stem.

Lemmatization converts every word to its lemma (i.e., a root form of the word). For example, the lemma of “pave,” “pavement,” “paving,” and “paved” would be “pave,” rather than “pav.” Since the lemmatization does not change the base form of the word, it is preferable when the analysis purpose is to remove inflectional endings only and to return the base form of the word. In order to keep the semantic basis of words, the author utilized the WordNet lemmatization algorithm that is provided by the NLTK. In other words, if there are two words, “contract” and “contractor,” the stemming would convert the words to the same stem (i.e., “contract”), while the lemmatization would convert “contract” to “contract,” and “contractor” to “contractor,” since the two words indicate different instances.

The research followed four simple data cleaning steps before the text preprocessing. The data cleaning steps included (1) converting every text to lowercase, (2) removing noise characters from the text, (3) overlaying “LINK,” “REF,” and “NUM” to the Uniform Resource Locators (URL), reference names, and numbers, respectively, and (4) integrating unit notations

(Figure 3.8). First, (1) converting every text to lowercase aimed to enable the model to recognize several cases well, such as “Contractor” and “contractor.” Second, (2) removing noise characters was necessary because plenty of noise characters were generated during data conversion (i.e., PDF to TXT) and encoding-decoding process of the TXT data, such as “\\r\\n.” Third, (3) overlaying “LINK,” “REF,” and “NUM” to the related words was to recognize the information as information type itself, instead of the exact qualitative value. Those types of information would be essential when comparing and reviewing steps of analysis, but they might be noises under text embedding and feature extraction. Fourth, (4) integrating unit notations handled the differently notated units among different specifications. For example, “%” and “percent” would be equally interpreted as “Percent.”

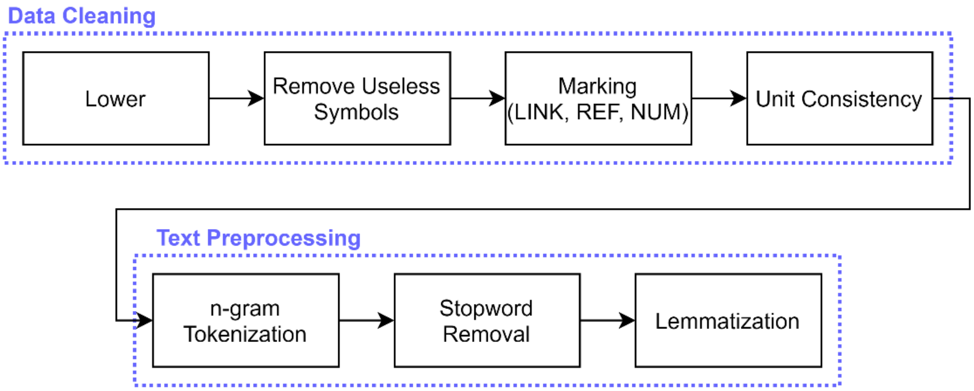


Figure 3.8 Text preprocessing steps

3.3. Development of Semantic Construction Thesaurus

This research utilized the Word2Vec model and Cosine similarity to find the most similarly distributed words for every term. The Word2Vec model embedded every word to the numeric vector, then the Cosine similarities between the word vectors were calculated. The results would be a dictionary of similarly used words, of which example is provided in Table 3.3.

Table 3.3 Example of dictionary of similarly used words

Word	**1st**	**2nd**	**3rd**
w_1	w_2	w_4	w_5
w_2	w_1	w_8	
w_3	w_1	w_{10}	
w_4	w_1		
w_5	w_1	w_4	
w_6	w_4	w_5	
w_7	w_1	w_2	w_8
w_8	w_2		
w_9	w_1		
w_{10}	w_3	w_4	

The items of the dictionary can operate as nodes and edges of the word link graph. Each word in the dictionary would be a node in the word link graph. The relationships between the key-word (i.e., hash) and the value-words (i.e., elements of the data block) would be the edges that link the nodes.

The data provided in Table 3.3 would be represented to a linked graph, as illustrated in Figure 3.9.

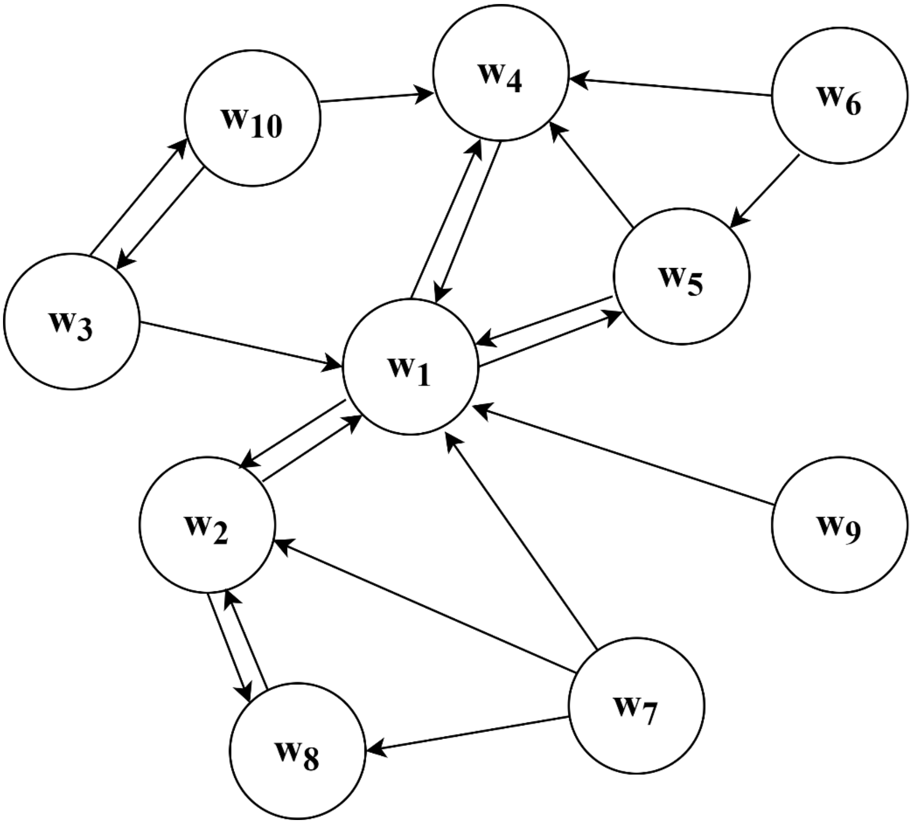


Figure 3.9 Example of word link graph

As the w1 and w4 in Figure 3.9 shows, several words might be converted to each other recursively. To address this problem, the author proposed a simple but powerful algorithm based on the concept of PageRank.

3.3.1. Word Embedding

The author developed the Word2Vec model based on the skip-gram architecture, as described in the '3.1.2 Text Embedding: Word2Vec' section. The Word2Vec model trained 346,950 words (8,692 terms) from all of the 19,346 sentences that were the manually extracted text data. The author set the hyperparameters of the Word2Vec model based on the empirical studies conducted by the author (Table 3.4).

Table 3.4 Hyperparameters of Word2Vec model

Hyperparameter	**Value**	**Description**
Vector Size	200	The dimension of word vector
Window Size	10	The number of adjacent words used to learn the word distribution
Minimum Count	50	The minimum frequency of each word to learn the distribution
Epochs	100	The number of iterations to learn the training data

The vector size implies the dimension of word vectors. The window size indicates the number of surrounding words that are considered to learn the usage pattern of a word. Too rarely occurred words, of which frequent was less than the minimum count, were discounted during the training. The epochs represent the number of iterations for the model trained a set of data. The skeleton of the Word2Vec architecture is provided in Figure 3.10.

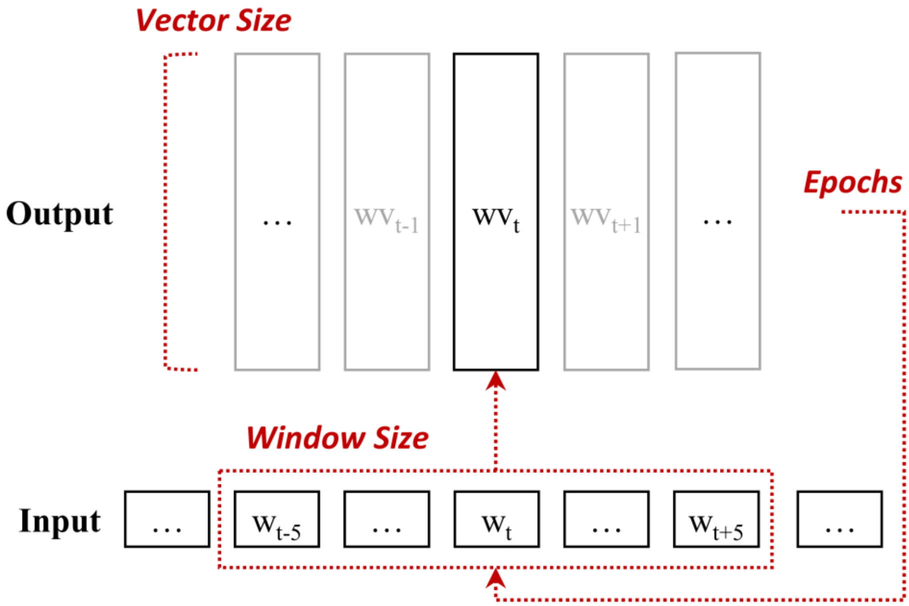


Figure 3.10 Word2Vec embedding architecture

Due to the Word2Vec architecture that learns the distributed representations rather than a lexicon, the proposed method can be utilized in analyzing documents written in other languages. The Word2Vec model can be trained again with the text data in other languages with a few human efforts for data preparation (Chung et al. 2017; Le and Mikolov 2014; Mikolov et al. 2000; Moon et al. 2019).

3.3.2. Pivot Term Determination based on Semantic Similarity

A similar word dictionary was developed based on the Cosine similarity between the word vectors of the Word2Vec model. The author calculated the Cosine similarity between the word vectors selected the ten most similar words for each word and dropped the dissimilar words of which the similarity showed less than 0.5. Finally, the similar word dictionary included the pairs of similar words, which facilitates the text analysis method to replace a word with its most similar word.

The similar word dictionary showed a problem that several words would be converted to each other recursively. For example, the words “tyr,” “tir,” and “pneum” (i.e., the lemmas of “tyre,” “tire,” and “pneumatic,” respectively) turned out to be forming a recursive network (Figure 3.11). Each node indicates a word, each edge indicates the relationship between two words, and the distance between two nodes is inversely proportional to the Cosine similarity between the two words.

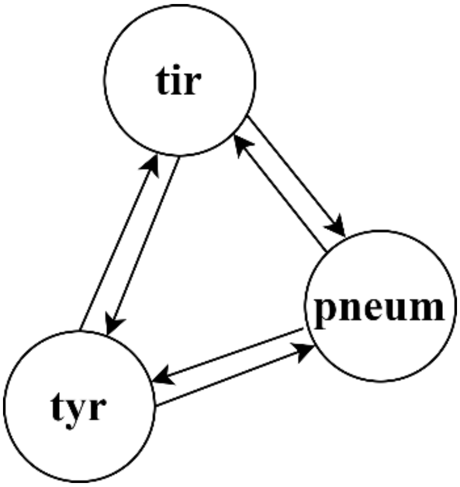


Figure 3.11 Sample of recursive word replacement

In order to address the recursive replacement problem, the authors proposed a link analysis approach of which concept is mainly based on the PageRank algorithm. Each word would be considered as a document in the PageRank, and the relationship between the words would be considered as a hyperlink in the PageRank. In this dissertation, the inflow of a term indicates the number of other terms that have the term as one of the similar terms. For example, in Figure 3.9, "w_3" and "w_4" generate inflows to "w_1." In reverse, the outflow of a term indicates the number of terms that the term showed to be similar, of which similarity is larger than 0.5. For example, "w_1" has three outflows to "w_2," "w_4," and "w_5." Such a term that acquires massive inflow and small outflow would be regarded as an important term (i.e., pivot term).

In other words, the pivot terms can be determined based on the number of links and the flow margins. The flow margin is calculated as Equation 3.1 to 3.4.

$$g(w', W_w) = \frac{n(W_w) - index(w' | sorted(W_w))}{n(W_w)} \quad (3.1)$$

$$inflow_w = \sum_{w' \in W_{w,IN}} g(w', W_{w,IN}) * c(w, w') \quad (3.2)$$

$$outflow_w = \sum_{w' \in W_{w,OUT}} g(w', W_{w,OUT}) * c(w, w') \quad (3.3)$$

$$f_w = inflow_w - outflow_w \quad (3.4)$$

where w and w' indicate an individual word, and W indicates a set of words. Particularly, $W_{w,IN}$ indicates a set of words that includes the word w as one of the most similar words and $W_{w,OUT}$ indicates the set of the most similar words of the word. $g(w', W_w)$ is a weight function of w' from the list of sorted the elements of W_w based on the Cosine similarity to w. f_w means the flow margin of a word, and the function $c(w, w')$ returns the Cosine similarity between the input word vectors. The algorithm of determining the pivot term is provided as below:

(1) If the flow margin of a word is positive, keep the word.

(2) If the flow margin is negative, replace the word with other words:

(2-1) If the flow margin of the most similar word is positive, replace the original word with the most similar word.

(2-2) Else, do (2-1) with the next most similar word.

(3) If there is no word of which flow margin is positive, replace the original word with the most similar word, and do (1).

3.4. Results of Semantic Construction Thesaurus

3.4.1. Results of Word Embedding

The Word2Vec model returned a unique vector for 1,409 terms, which can be inferred that most sentences from specifications widely share analogous terms. As Word2Vec is an unsupervised embedding model, the author evaluated the embedding results in a qualitative approach; randomly selected several words, and investigated the most similar words of each word. As the words "American Association of State Highway and Transportation Officials (AASHTO)" and "American Society for Testing Materials (ASTM)" are the most frequently occurred references, the author expected that the most similar word for "AASHTO" would be "ASTM." However, a tri-gram "in-accordance-with" turned out to be the most similar word for "AASHTO," while "ASTM" was the second most similar, due to the plenty of clauses that have the phrase of "… in accordance with AASHTO …." Similarly, the most similar word for "Contractor" was a tri-gram "by-the-contractor." "Bituminous" and "Failure" showed reasonable results of which the most similar words were "asphalt" and "event," respectively. Meanwhile, "Weather" had "event" for the most similar word, and "condition," which seemed to have to be the most similar, was the second (Table 3.5). In summary, the Word2Vec model delivered a fine performance in learning the distributed representation of specification text, despite a few miss-graded similarities.

Table 3.5 Samples of Word2Vec embedding result

Word	Most Similar Words
AASHTO	in-accordance-with, ASTM, standard
Contractor	by-the-contractor, proposed, his
Bituminous	asphalt, mixture, surfacing
Failure	event, load, defect
Weather	event, condition, bed

3.4.2. Semantic Word Similarity

The cosine similarities between word vectors were calculated, and the most five similar words, of which the similarity is larger than 0.5, were listed each word to build a similar word dictionary. The author post-processed the dictionary to eliminate several inappropriate records. First, 965 records that have empty data were removed. The empty data indicates that no similar words exist, of which the Cosine similarity exceeds 0.5. Second, seven rules of which key-word is a single alphabet that seemed to be derived from chapter names (e.g., “B. Asphalt Plants” were removed. Exceptionally, the author took the key-word “a” as the data block of the alphabet showed prepositions (e.g., “an,” “the,” “of”). As a result, the research secured 374 records of a similar word dictionary (Table 3.6).

Table 3.6 Sample of similar word dictionary

Term	**Most Similar Terms**				
	1st	**2nd**	**3rd**	**4th**	**5th**
asphalt	bind	mix	mixt	bitumen	liquid
aggreg	coars	mixt	crush	min	blend
mix	the-mix	mixt	asphalt	batch	-
pav	concret	-	-	-	-

The records of the similar word dictionary are illustrated in the form of word network with the direction from the original term to similar words. Since plenty of the records show the recursive replacement problems, the necessity of determining pivot terms would resonate (Figure 3.12).

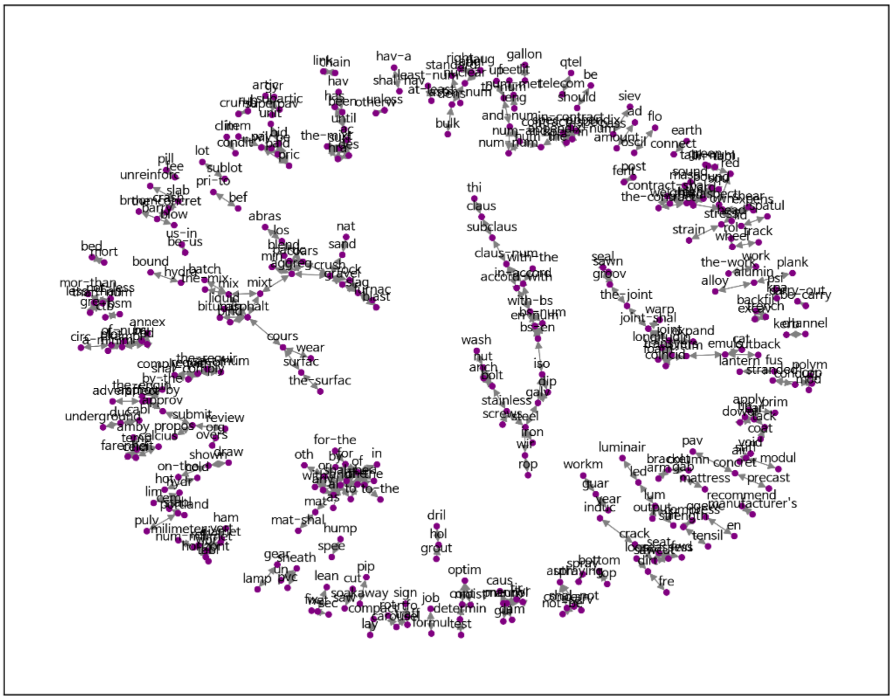


Figure 3.12 Word network of similar word dictionary

3.4.3. Semantic Construction Thesaurus

The semantic construction thesaurus was developed by analyzing every link between the terms and the similar words based on a simple concept of the PageRank algorithm. The replacement rules of the semantic construction thesaurus (i.e., each term to the pivot) are illustrated in the form of a word network (Figure 3.13). The thesaurus rules will replace every word at the start points with the word at the endpoint throughout the text. Note that the words are in the lemmatized forms.

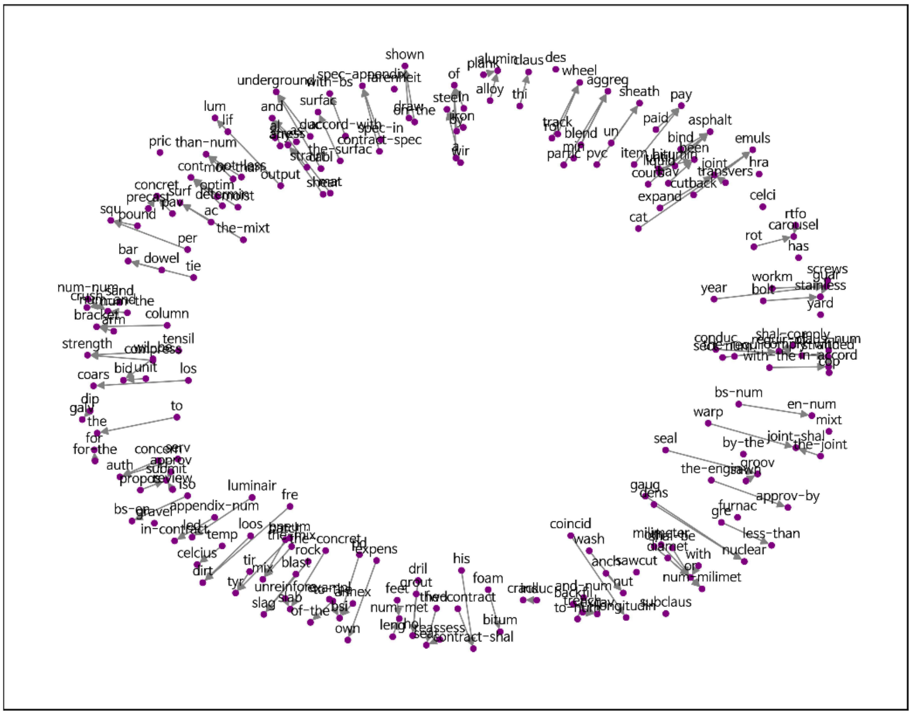


Figure 3.13 Semantic construction thesaurus

As the word replacement rules of the developed semantic construction thesaurus have no correct answers, the author evaluated the results by investigating several randomly selected rules. Table 3.7 provides the samples of the semantic construction thesaurus.

Table 3.7 Word replacement rules of semantic construction thesaurus

Index	Word		Pivot Term	
	Lemma	Original Token	Lemma	Original Token
1	temp	temperature	celcius	celcius
2	tir	tire	tyr	tyre
3	propos	propose	submit	submit
4	approv	approval	submit	submit
5	iron	iron	steel	steel
6	bitumin	bituminous	asphalt	asphalt
7	liquid	liquid	asphalt	asphalt
8	item	item	pay	payment
9	accord-with	accordance-with	with-bs	with-BS
10	shal-comply	shall comply	require-of	requirement-of

The records with the index number of 1, 2, 3, 5, and 6 turned out to be reasonable. Although a few conversions would be recognized better if the converting direction is opposed (e.g., “temp” lemma to “celcius” lemma), the direction is not a problem for the computer. The point is that the computer recognizes the two words as highly similar. However, some records operate to replace the word with the word that is appeared just behind (i.e., records of

4, 7, 8, 9, and 10). For example, the “accord-with” lemma and the “with-bs” lemma are come from the sentence like “… in accordance with the BS EN ….” These misinterpretations of the construction thesaurus are caused by the Word2Vec model. The Word2Vec model is widely known for understanding the extremely close words to be similar, which is respectable in general cases since the close words would co-occur in high frequency. The author removed several conversion records manually based on the in-depth collaboration with the construction practitioners to consider the in-practice insights. Finally, the construction thesaurus included 208 conversion records.

3.5. Summary

In this chapter, the first objective of this dissertation was covered, which is to develop the semantic construction thesaurus to facilitate the computers to understand the text regardless of the different vocabulary. The thesaurus indicates a dictionary of the relationship between terms, which replaces every term with the pivot form. First, the author collected a total of 56 construction specifications and manually converted 2,527 clauses (i.e., 19,346 sentences) from the corpus into the TXT format that is required in text analysis. Then, the research followed several steps for data cleaning and text preprocessing. Next, the Word2Vec embedding model with CBOW architecture learned the distributed representation of 346,950 words (i.e., 8,692 terms) from all of the 19,346 sentences, and returned a unique vector for 1,409 terms that occurred with frequent of larger than the minimum threshold. The author constructed a similar word dictionary based on the Cosine similarity between the word vectors of the Word2Vec model. Since the similar word dictionary showed a problem that several words would be converted to each other recursively, the authors proposed a link analysis approach of which concept is mainly based on the PageRank algorithm. Such a term that acquires massive inflow and small outflow would be regarded as an important term (i.e., pivot term). Finally, the author developed the thesaurus with 208 replacement rules.

Chapter 4. Qualitative Requirement Recognition on Construction Clauses

This chapter covers the second objective of this dissertation to recognize construction keywords automatically regardless of sentence structure. As described in the '2.3 Limitations of ' section, the different sentence structures of different specifications obstruct comparing two provisions from different documents automatically. The author addressed the problem with the machine learning-based NER model that can recognize construction keywords from sentences. First, five information categories were defined based on in-depth collaborations with experienced construction practitioners, which are essential to understand the construction specification. Then, the researchers manually labeled every word token to the pre-defined categories for training the model. Next, each word token was mapped to an identical numeric vector by Word2Vec, which converted text data into a computer-readable vector format. Finally, the RNN model was developed, including Bi-LSTM and CRF layers, which predicted the category of each word (Figure 4.1). The theoretical backgrounds, the development process, validation, and discussions of the NER model are described in this chapter.

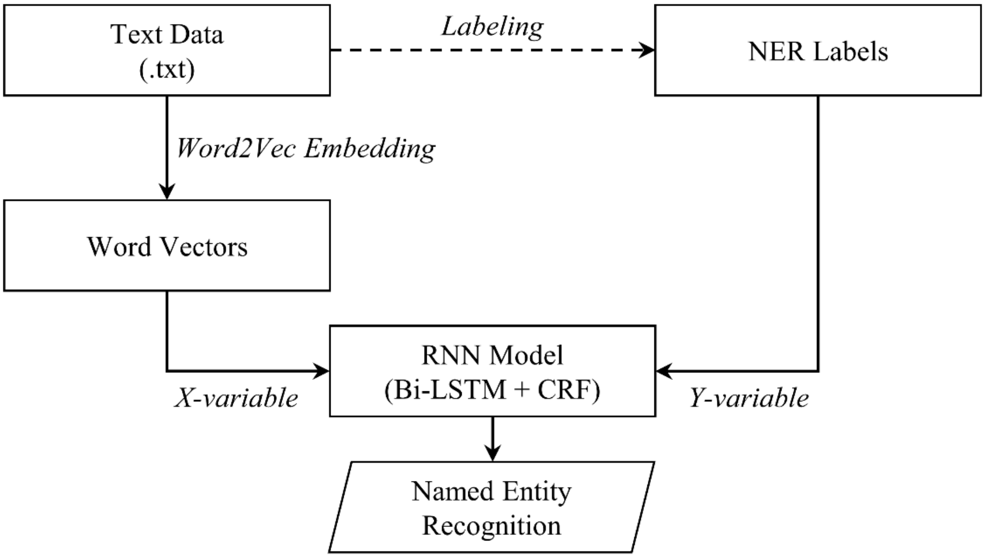


Figure 4.1 Research process of construction keyword recognition

4.1. Research Method: Construction Keyword Recognition

4.1.1. Named Entity Recognition

NER is a subfield of machine learning-based information extraction methodologies, which recognizes each word with pre-defined labels, such as name, location, and an object (McCallum and Li 2003; Sang and De Meulder 2003). The categories of the labels, called “named entities,” are defined by researchers in order to comprehend the text data based on the categories. For example, in the construction industry, the NER model was developed to extract construction instances, such as regulatory items (Zhang and El-Gohary 2016), bridge defect information (Liu and El-Gohary 2017) and accident information (Kim and Chi 2019).

NER can be implemented in two ways for which feature of the text data is used, i.e., syntactic features or semantic features. First, NER that uses syntactic features is known to show good performance for small and clean datasets since syntactic expressions in a sentence needed to determine the category of each target word can be easily extracted from such datasets (Newman et al. 2006). For instance, words associated with the “name” category would have the first character to be capitalized, words associated with the “location” category would appear right after a preposition, such as “in,” “on,” or “to,” and words associated with the “object” category would be nouns in many cases. This approach shows satisfactory accuracy if the text data are in a clean and standard format. In research, Zhang and El-Gohary (2016) defined which objected to extract from text sentences (e.g., subject, subject restriction, quantitative relation, quantity value) and developed extraction rules based on “phrase structure grammar.” Liu and El-Gohary (2017) proposed a NER model to extract bridge damage information, such as bridge elements, deficiency types and causes of deficiency, and maintenance activities. Kim and Chi (2019) calculated the conditional probability of each word's role in the sentence and extracted the accident keywords, including hazard object, hazard position, work process, and accident result.

Although such syntactic approaches provided promising results in the information extraction, they all required domain knowledge and ontologies (i.e., the relationship between text words) to build extraction rules. Besides, the approaches also required new extraction rules as the sentence structures might diverse among the documents. On the other hand, NER using semantic features is well known for its robustness and expandability compared to the syntactic approaches. These approaches identify the text features (i.e., usage patterns of each word) automatically and acquire the semantic information based on a machine learning algorithm (Cucerzan and Yarowsky 1999; Ratinov and Roth 2009). Therefore, the approaches enable the model to be less limited to the sentence structure (i.e., writing styles). For instance, if a set of data consists of two sentences, “The Contractor should prepare the equipment” and “The document should be submitted by the Engineer,” the semantic NER model would consider usage patterns and classify “Contractor” and “Engineer” into the same category, even though the syntactic roles are different. In this research, a semantic NER model was developed to deal with the inconsistencies of sentence structures among different construction specifications.

4.1.2. Recurrent Neural Network

RNN is one of the Deep Neural Networks (DNN) of which the networks are connected in a series structure (Nallapati et al. 2016). Due to the serialized networks, the RNN model can use the sequential information of the input data, and thus outperform other machine learning models in the analysis of serialized data, such as NER. Figure 4.2 illustrates the basic architecture of RNN, where t, x, h, and y indicate sequence step, layer input, layer output, and output class, respectively.

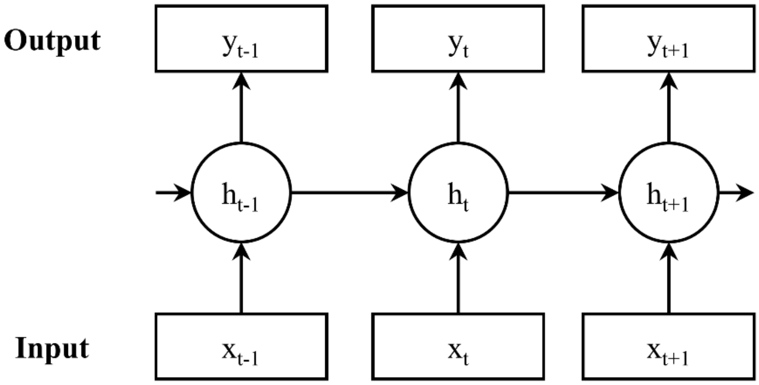
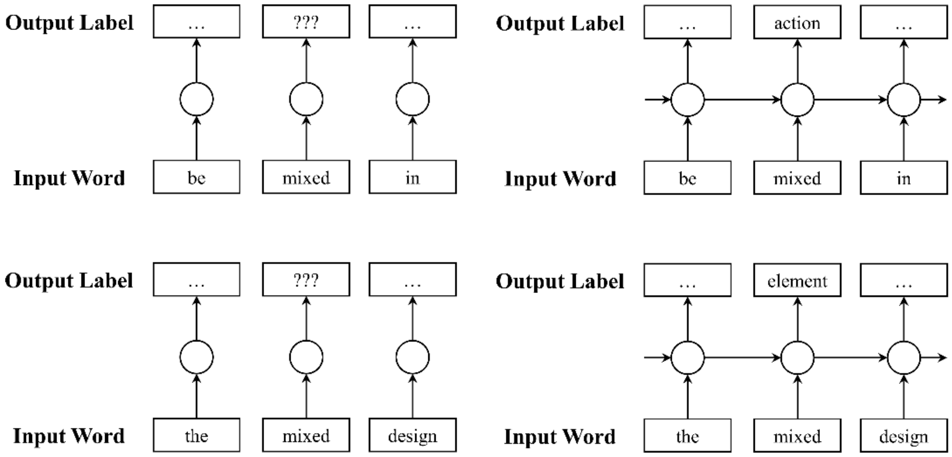


Figure 4.2 RNN model architecture

Unlike structured data, text data (i.e., the sequence of words) have context information between words. Hence, the text sometimes has a different meaning, even with the same spelling. For instance, a term “mixed” is used as a verb (i.e., action to do) in a sentence “… should be mixed in …”; however, in another case, the same term is used as one of a noun phrase (i.e., construction element) in a sentence “… a sampling of the mixed design ….” The conventional model that does not have the serialized structure cannot differentiate these examples since the model input is the same as “mixed” (Figure 4.3(a)). However, the RNN model can address the problem by getting the input data as the sequence of words. In other words, the RNN model can classify the category of the term “mixed” with higher accuracy by considering the context information of the text data (Figure 4.3(b)).



(a) (b)

Figure 4.3 Example of word classification result:

(a)conventional classification model, (b)RNN model

Although the RNN model is competent to analyze text data, it still has a limitation called a “vanishing gradient problem.” The longer the network serializes, the smaller the gradient becomes, which is vital for delivering past information forward and obtaining updates by backpropagation; hence, the learning ability of the model decreases severely. The LSTM concept has addressed such limitations (Hochreiter and Schmidhuber 1997). The LSTM networks contain a unique structure (i.e., forget gate and input gate) within the hidden layer of RNN (Figure 4.4). The forget gate (i.e., f_t), plays a role in forgetting the past information, whereas the input gate (i.e., i_t), plays a role in remembering the current information. Therefore, the model can conserve essential information for a longer distance by forgetting unnecessary signals while reinforcing necessary signals (Wu et al. 2016).

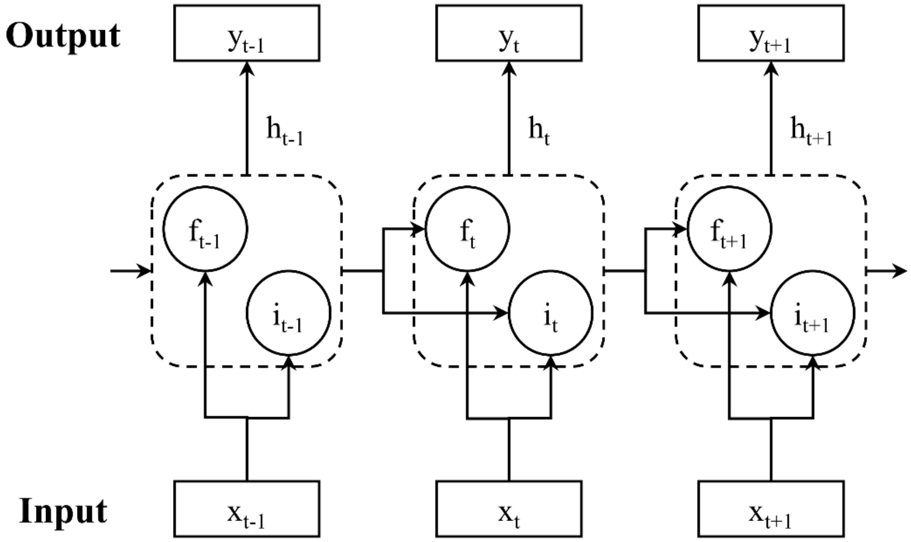


Figure 4.4 LSTM model architecture

4.1.3. Bi-directional Long Short-Term Memory

In this research, Bi-LSTM architecture that considers the sequential information in both forward and backward ways was used to develop the NER model. The Bi-LSTM contains two LSTM layers (i.e., forward and backward), and the output sequences of the two layers that were combined by the concatenating function (σ) (Figure 4.5) (Cui et al. 2018).

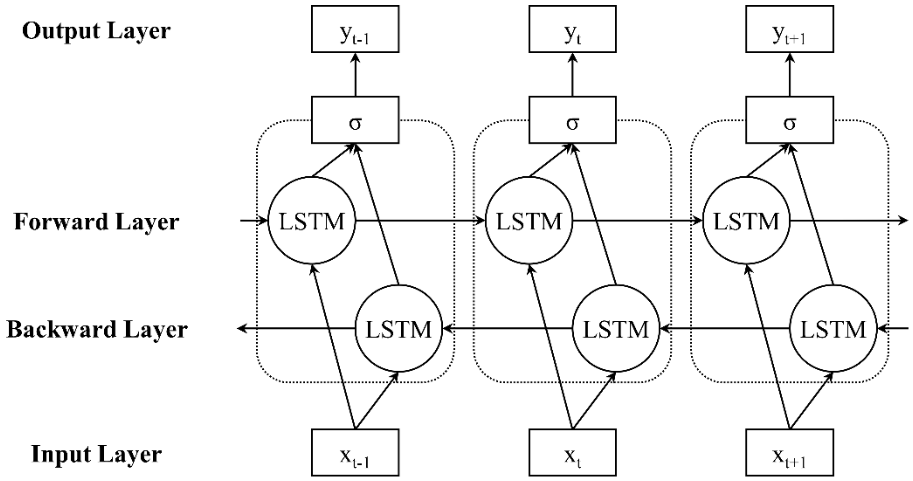


Figure 4.5 Bi-LSTM model architecture

For a detailed structure, the LSTM unit at time *t* contains the input vector x_t and the layer output h_t, where f_t, i_t, and o_t indicate the forget gate, the input gate, and the output gate, respectively (Figure 4.6). The cell state considers the cell input state (i.e., $\widetilde{C_t}$), the cell output state (i.e., C_t), and the previous cell output state, (i.e., C_{t-1}). Ws and Us are weight matrices, and bs are bias vectors. σ_g indicates the activation function (Equation 4.1 to 4.6).

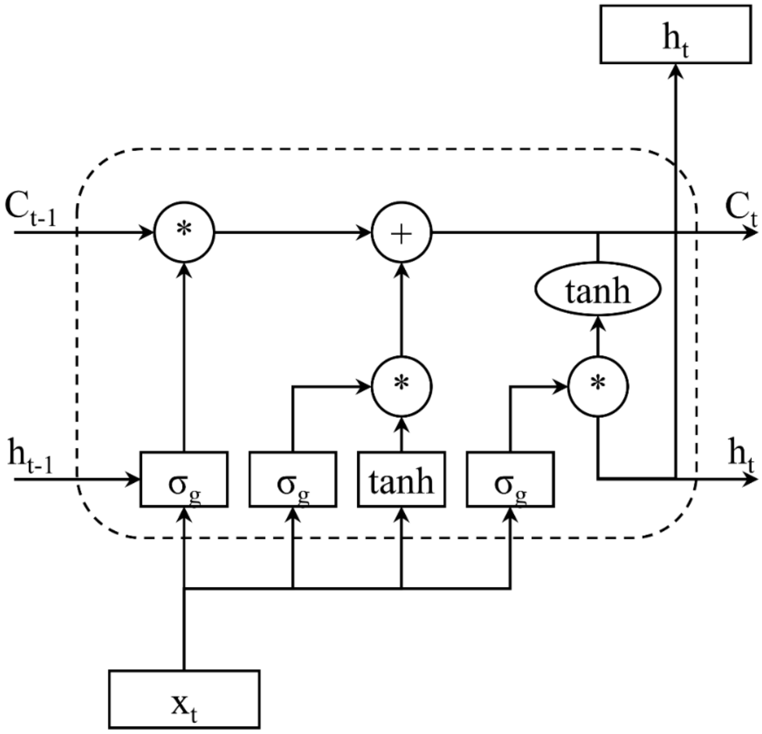


Figure 4.6 LSTM unit structure

$$f_t = \sigma_g\big(W_f \ \cdot x_t + U_f h_{t-1} + b_f\big) \tag{4.1}$$

$$i_t = \sigma_g\big(W_f \ \cdot x_t + U_i h_{t-1} + b_i\big) \tag{4.2}$$

$$o_t = \sigma_g(W_o \ \cdot x_t + U_o h_{t-1} + b_o) \tag{4.3}$$

$$\widetilde{C_t} = \tanh(W_C \cdot x_t + U_C \cdot h_{t-1} \ + b_C) \tag{4.4}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t, \tag{4.5}$$

$$h_t = o_t * \tanh(C_t) \tag{4.6}$$

4.1.4. Conditional Random Field

Conditional Random Field (CRF) is one of the statistical modeling methods, which is specialized in taking context (i.e., the adjacent samples) into account compared to the conventional statistical models (Lafferty et al. 2001). It had shown good performance in the analysis of sequenced data before the RNN was introduced. Recently, the RNN models commonly apply the CRF as the last layer of prediction to avoid label bias problem by considering the labels of the adjacent samples (Huang et al. 2015; Lample et al. 2016).

CRF requires two sequences of random variables, which are jointly distributed. If $G = (V, E)$ be a graph between ***X*** (i.e., a random variable of input sequence) and ***Y*** (i.e., a random variable of output sequence), the pair of (***X***, ***Y***) is called a conditional random field. In case, every component of ***Y*** obeys the Markov property when conditioned on ***X***, which is provided in Equation 4.7, where *w~v* means that *w* and *v* are neighbors. Then, the joint distribution over the ***Y*** given ***X*** for every sample k can be formulated as Equation 4.8, where **x** and **y** indicate the input sequence and output sequence, respectively (Lafferty et al. 2001). The notation $\mathbf{y}|_S$ means the set of components of y associated with the subgraph S for the given feature functions f_k and g_k the CRF model learns the parameters $\Lambda =$

$(\lambda_1, \lambda_2, \ldots; \mu_1, \mu_2, \ldots)$ from training data. After training the parameters, the CRF model can return the most likelihood class y^* as Equation 4.9 (Kim and Chi 2019). The CRF model selects a label that maximizes the probability among all possible sequences of labels (Figure 4.7).

$$p(\boldsymbol{Y}_v | \boldsymbol{X}, \boldsymbol{Y}_w, w \neq v) = p(\boldsymbol{Y}_v | \boldsymbol{X}, \boldsymbol{Y}_w, w \sim v) \quad (4.7)$$

$$p_\Lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{z(\mathbf{x})} \exp\left(\sum_{e \in E,k} \lambda_k f_k(e, \mathbf{y}|_e, \mathbf{x}) + \sum_{v \in V,k} \mu_k g_k(v, \mathbf{y}|_v, \mathbf{x})\right) \quad (4.8)$$

$$y^* = \arg\max_y p_\Lambda(\mathbf{y}|\mathbf{x}) \quad (4.9)$$

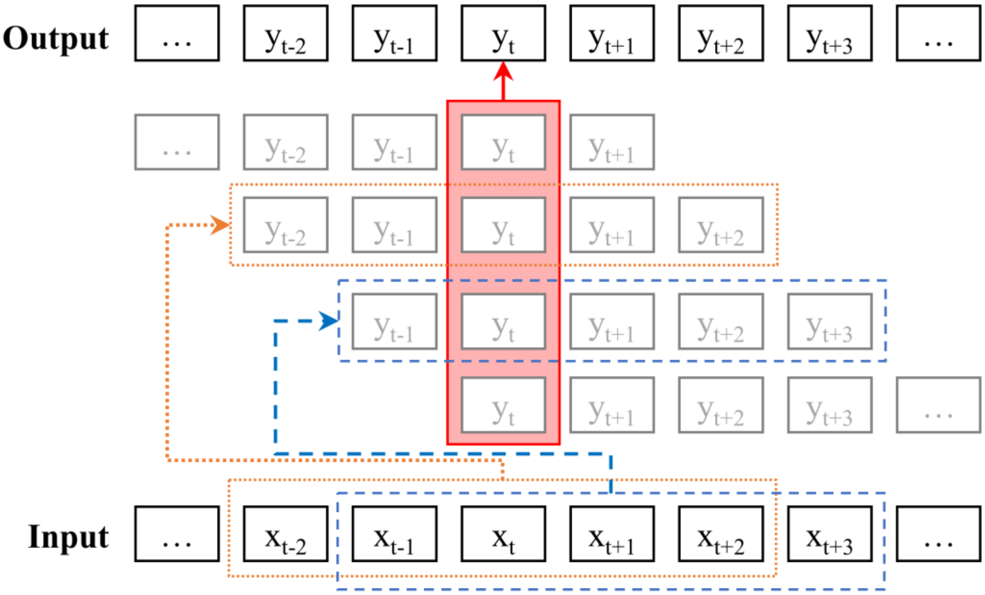


Figure 4.7 CRF framework

4.2. Development of NER Model for Construction Keyword Recognition

4.2.1. Data Labeling

The authors collaborated with fourteen experienced practitioners to understand the information needs for risk management, and identified four types of questions that should be answered by the specification. If these types of questions are not clearly answered or such contents do not satisfy the local standards during the planning phase before construction, the contractors can face risks during the actual construction while causing rework, cost overrun, or project delay.

(1) Who was responsible for? For instance, who was responsible for the maintenance of equipment: engineer or contractor?

(2) What should be done for when? An example is that the pavement surface should be cleaned at least three days before pouring the cement.

(3) How should the construction be? An example is that the edges should be slopped at gradients, “not exceeding 10%.”

(4) Which reference should be followed? For example, the sampling of aggregates shall be done in accordance with “AASHTO T2, T248, ASTM C50” or equivalent as applicable.

Table 4.1 Personal information of consultants

Index	**Department**	**Position**	**Work Experience (Year)**
1	Design	Director	30
2	Design	Director	30
3	Design	Department Head	15
4	Design	Department Head	15
5	Engineering	Department Head	20
6	Engineering	Department Head	20
7	Engineering	Department Head	20
8	Engineering	Deputy Department Head	15
9	Engineering	Deputy Department Head	15
10	Engineering	Deputy Department Head	15
11	Engineering	Deputy Department Head	15
12	Engineering	Deputy Department Head	15
13	Engineering	Manager	10
14	Engineering	Manager	10

The contractors finally concurred that the following information types are crucial to answer to the identified questions: (1) persons and organizations in charge, (2) activities required, (3) construction and installation items, (4) quality standards and criteria, and (5) relevant references (Table 4.2). The useful and crucial contents for risk management can be understood based on the determined information types. The authors confirmed that every informative keyword from the specification can be assigned to one of those five categories.

Table 4.2 Identification of Information Types

Questions for Risk Management	**Information**	**User-needed Information Type**
(1) Who was responsible for?	Who	(1) Persons and organizations in charge
(2) What should be done for when?	What	(3) Construction and installation items
	When	(4) Quality standards and criteria
(3) How should be the construction be?	How	(2) Activities required
	How	(4) Quality standards and criteria
(4) Which reference should be followed?	Which reference	(5) Relevant references

The organization category explained subjects, participants, and stakeholders, such as “contractor,” “engineer,” and “designer.” The action category covered information about “how the standard should be met,” so it usually included words that corresponded to verb phrases, such as “must submit,” “have to approve,” and “shall test.” The term “Element” referred to the construction element that was utilized at the site, and it was usually mentioned as the object of the standards. The element category included formulas, the name of materials and equipment, and documents. The standard category refers to the actual criteria that the organizations must follow or that the construction elements should satisfy. The standard category was usually related to numerical values, such as “one month,” “a week,” and “38 mm.”

The reference category was composed of every document, specification, and code referenced in the text, including “AASHTO,” “ASTM,” and “BS EN ISO.” To consider the remaining words that were not assigned to one of the first five meaningful categories, the last category “None” was added to the named entities. Various meaningless words, such as “and,” “to,” and “for,” were assigned to the “None” category. Table 3 provides examples of each NER category.

The information types determined from the case study in Qatar referred to five named entities, “Organization (ORG),” “Action (ACT),” “Element (ELM),” “Standard (STD),” and “Reference (REF),” which were used as informative word categories in the NER model. The organization category explained subjects, participants, and stakeholders, such as “contractor,” “engineer,” and “designer.” The action category covered information about “how the standard should be met,” so it usually included words that corresponded to verb phrases, such as “must submit,” “have to approve,” and “shall test.” The term “Element” referred to the construction element that was utilized at the site, and it was usually mentioned as the object of the standards. The element category included formulas, the name of materials and equipment, and documents. The standard category refers to the actual criteria that the organizations must follow or that the construction elements should satisfy. The standard category was usually related to numerical values, such as “one month,” “a week,” and “38 mm.” The reference category was composed of every document, specification, and code referenced in the text, including “AASHTO,” “ASTM,” and “BS EN ISO.” To consider the remaining words that were not assigned to one of the first five meaningful categories, the last category “None” was added to the named entities. Many meaningless words were assigned to the “None” category, such as “and,” “to,” and “for.” Table 4.3 provides examples of each NER category.

Table 4.3 Examples of NER categories

Information Type	**Category**	**Examples**
Organization	ORG	contractor, engineer, designer
Action	ACT	must submit, have to approve, shall test
Element	ELM	formula, certification, design value
Standard	STD	one month, a week, 38 mm
Reference	REF	AASHTO, ASTM, BS EN ISO
None	NON	and, to, for

Six construction practitioners were involved in manually assigning word labels to 4,659 sentences of construction specifications according to the defined categories to be used for training and testing the NER model. They read every sentence and assigned the appropriate category (i.e., ORG, ACT, ELM, STD, and REF) to every word in the sentence. Every labeled sentence was cross-checked by them to assure that the data was consistently labeled. Such words that do not play a role in understanding the context were labeled as the “None” category (i.e., NON).

4.2.2. NER Model Development

The Word2Vec model that was developed before as ‘3.4.1 Results of Word Embedding’ section was updated with the semantic construction thesaurus. As a result of applying the thesaurus, 49 terms were replaced, and the total number of terms turned into 8,643. The Word2Vec model trained the thesaurized text data under the same hyperparameters (Table 3.4). Then, this research proposed an automatic information extraction model based on NER with Word2Vec Embedding, Bi-LSTM, and CRF (Figure 4.8). The model utilizes various text features, including (1) the numerical values from each word vector via Word2Vec, (2) the bi-directional order of the words in each sentence via Bi-LSTM, and (3) the bi-directional order of the labels in each sentence via CRF. The framework of developing the NER model includes input data adjustment (i.e., sentence padding), text embedding via Word2Vec, and RNN based on Bi-LSTM, dense, and CRF layers (Figure 4.9).

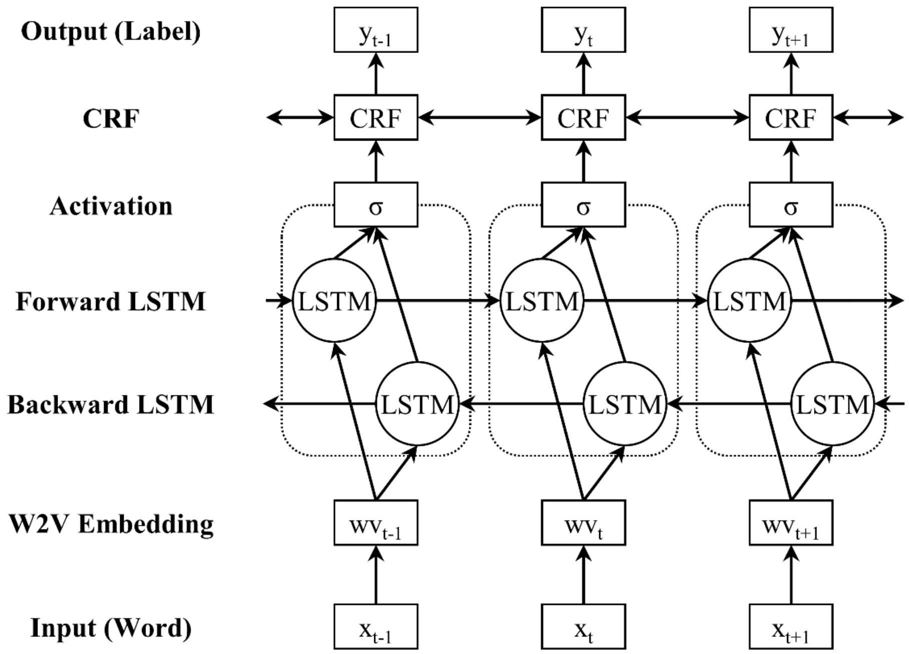


Figure 4.8 NER model architecture

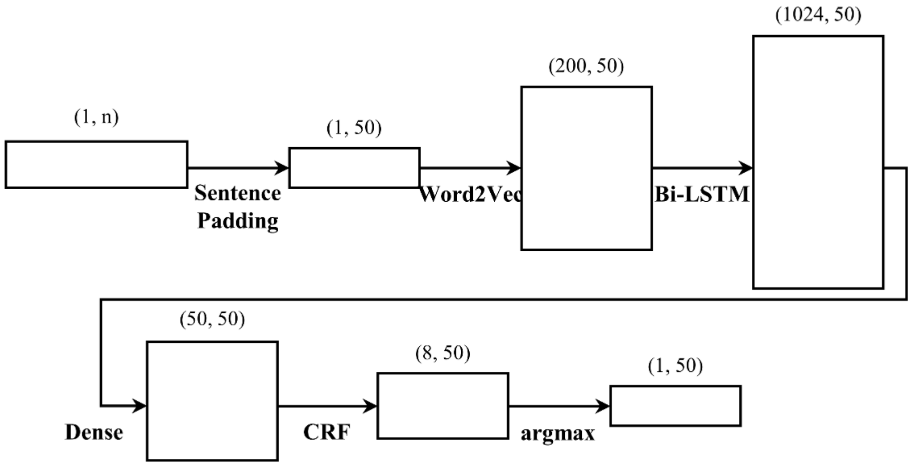
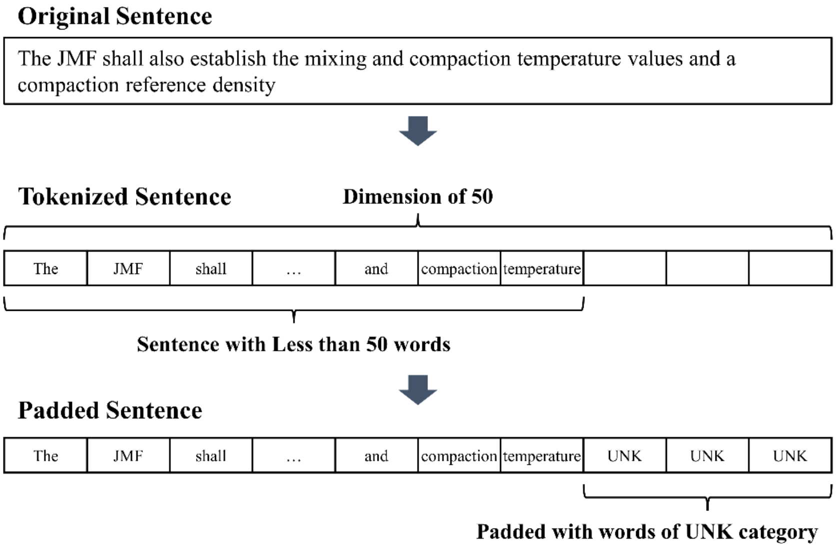
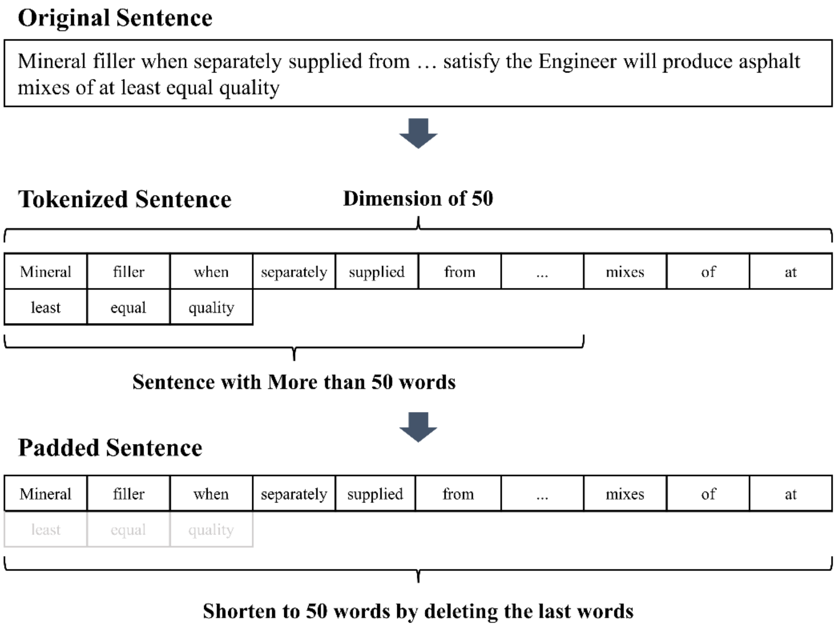


Figure 4.9 NER model framework

The first step in developing the NER model was sentence padding. Since the input data (i.e., sentences) has a different length (i.e., the number of words in each sentence are different) and the RNN model requires the input data to have the same length, every sentence was tokenized into a sequence of words and padded to the length of 50 (i.e., each sentence should consist of 50 words). Sentences with less than 50 words were extended with new tokens assigned to the 'Unknown (UNK)' category (Figure 4.10(a)), and, conversely, sentences with more than 50 words were shortened by deleting the last words (Figure 4.10(b)). The maximum length of the sentence was set to 50 words according to the experiential knowledge that 50 words were sufficient to contain meaningful contexts in construction specifications.


Original Sentence
The JMF shall also establish the mixing and compaction temperature values and a compaction reference density
Tokenized Sentence
Dimension of 50
The
JMF
shall
…
and
compaction
temperature
Sentence with Less than 50 words
Padded Sentence
The
JMF
shall
…
and
compaction
temperature
UNK
UNK
UNK
Padded with words of UNK category

(a)



(b)

Figure 4.10 Sentence padding: (a)sentence less than 50 words, (b)sentence more than 50 words

Then, the NER model was developed based on RNN architecture, including Bi-LSTM, Dense, and CRF layers. The Dense layer operates for dimension reduction, converting the Bi-LSTM output (i.e., a sparse matrix of which shape is 1,024 by 50) to a dense matrix for better representation. The hyperparameters of the NER model are provided in Table 4.4.

Table 4.4 Hyperparameters of NER model

Layer	Hyperparameter	Value	Description
Input	Maximum Sentence Length	50	The determined number of words in each sentence (i.e., input length)
LSTM	Units	1,024	Dimension of Bi-LSTM units
Dense	Units	50	Dimension of dense matrix

The NER model was trained using 70% of the data (3,261 sentences), of which 90% (2,935 sentences) were used to train the model and 10% (326 sentences) were used to validate the model. Next, the author tested the model by the remaining 30% of the data (1,398 sentences). The training, validation, and testing datasets were partitioned randomly, irrespective of the origin of each sentence (i.e., which document each sentence came from). Therefore, the model could be robust to the sentence structure by learning a range of text information with different expressions. The model trained under a dropout of 0.2, batch size of 32, and epochs of 200, which were determined by grid search. Furthermore, the model trained thirty different sets of randomly split training, validation, and testing data to avoid overtraining on a specific dataset.

4.3. Results of Construction Keyword Recognition

4.3.1. Results of Thesaurized Word Embedding

The Word2Vec model that trained the thesaurized text data returned a unique vector for 1,388 terms, rather than 1,409. The total amount of unique terms decreased after applying the semantic construction thesaurus since several terms were converged to their pivot term. The decreased amount (i.e., 21 terms) was different from the number of covered terms of the thesaurus (i.e., 208 terms) since the Word2Vec model trained only a part of text data that included NER labels.

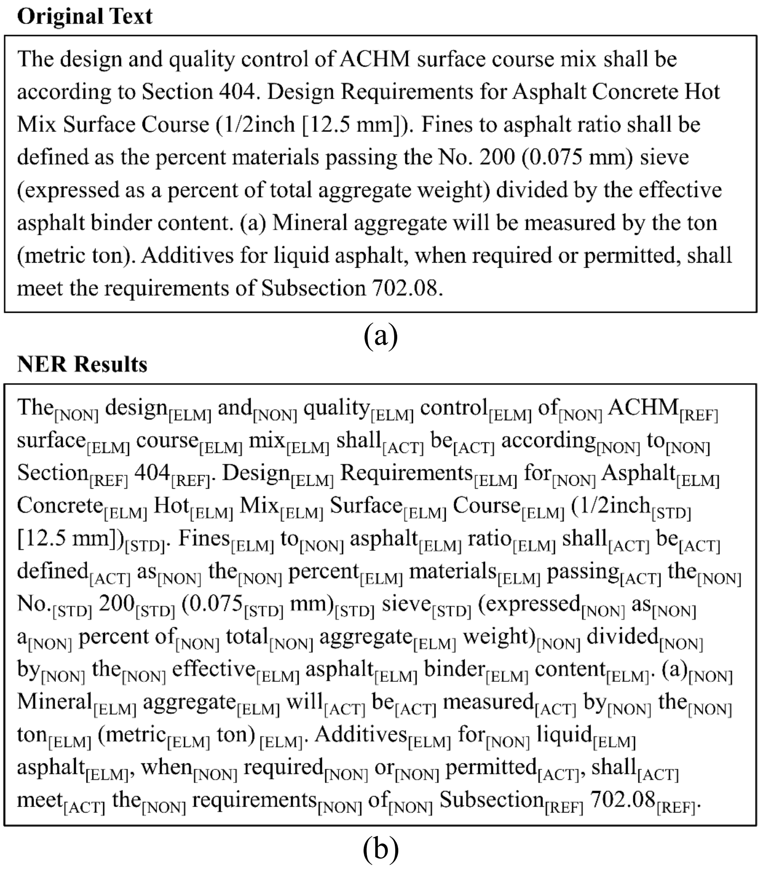
The Word2Vec model delivered better performance in embedding the words. For example, the most similar word for “AASHTO” appeared to “ASTM,” which was “in-accordance-with” without the thesaurus. Although the results of “Bituminous,” “Contractor,” and “Failure” did not change dynamically, the word “condition” emerged to be the most similar word for “Weather” as expected (Table 4.5).

Table 4.5 Samples of thesaurized Word2Vec embedding result

Word	Most Similar Words	
	Original Text	**Thesaurized Text**
AASHTO	in-accordance-with, ASTM, standard	ASTM, standard, text
Bituminous	asphalt, mixture, surfacing	asphalt, mixture, material
Contractor	by-the-contractor, proposed, his	by-the-contractor, proposed, subcontractor
Failure	event, load, defect	event, load, stress
Weather	event, condition, bed	condition, event, public

4.3.2. NER Model Validation

Figure 4.11 showed examples of the NER results that every word of the original text (Figure 4.11(a)) was automatically tagged by one of the six-word categories (Figure 4.11(b)). The sentences were sampled randomly from the validation set. As indicated in the figure, every word was assigned one-by-one to the same category. Although there were some misclassifications, e.g., “total” was assigned to “NON,” “percent” was assigned to “ELM,” and “sieve” was assigned to “STD,” other 75 identification results of the total 85 appeared to be consistent with common practical knowledge.



Original Text

The design and quality control of ACHM surface course mix shall be according to Section 404. Design Requirements for Asphalt Concrete Hot Mix Surface Course (1/2inch [12.5 mm]). Fines to asphalt ratio shall be defined as the percent materials passing the No. 200 (0.075 mm) sieve (expressed as a percent of total aggregate weight) divided by the effective asphalt binder content. (a) Mineral aggregate will be measured by the ton (metric ton). Additives for liquid asphalt, when required or permitted, shall meet the requirements of Subsection 702.08.

(a)

NER Results

The$_{[NON]}$ design$_{[ELM]}$ and$_{[NON]}$ quality$_{[ELM]}$ control$_{[ELM]}$ of$_{[NON]}$ ACHM$_{[REF]}$ surface$_{[ELM]}$ course$_{[ELM]}$ mix$_{[ELM]}$ shall$_{[ACT]}$ be$_{[ACT]}$ according$_{[NON]}$ to$_{[NON]}$ Section$_{[REF]}$ 404$_{[REF]}$. Design$_{[ELM]}$ Requirements$_{[ELM]}$ for$_{[NON]}$ Asphalt$_{[ELM]}$ Concrete$_{[ELM]}$ Hot$_{[ELM]}$ Mix$_{[ELM]}$ Surface$_{[ELM]}$ Course$_{[ELM]}$ (1/2inch$_{[STD]}$ [12.5 mm])$_{[STD]}$. Fines$_{[ELM]}$ to$_{[NON]}$ asphalt$_{[ELM]}$ ratio$_{[ELM]}$ shall$_{[ACT]}$ be$_{[ACT]}$ defined$_{[ACT]}$ as$_{[NON]}$ the$_{[NON]}$ percent$_{[ELM]}$ materials$_{[ELM]}$ passing$_{[ACT]}$ the$_{[NON]}$ No.$_{[STD]}$ 200$_{[STD]}$ (0.075$_{[STD]}$ mm)$_{[STD]}$ sieve$_{[STD]}$ (expressed$_{[NON]}$ as$_{[NON]}$ a$_{[NON]}$ percent of$_{[NON]}$ total$_{[NON]}$ aggregate$_{[ELM]}$ weight)$_{[NON]}$ divided$_{[NON]}$ by$_{[NON]}$ the$_{[NON]}$ effective$_{[ELM]}$ asphalt$_{[ELM]}$ binder$_{[ELM]}$ content$_{[ELM]}$. (a)$_{[NON]}$ Mineral$_{[ELM]}$ aggregate$_{[ELM]}$ will$_{[ACT]}$ be$_{[ACT]}$ measured$_{[ACT]}$ by$_{[NON]}$ the$_{[NON]}$ ton$_{[ELM]}$ (metric$_{[ELM]}$ ton) $_{[ELM]}$. Additives$_{[ELM]}$ for$_{[NON]}$ liquid$_{[ELM]}$ asphalt$_{[ELM]}$, when$_{[NON]}$ required$_{[NON]}$ or$_{[NON]}$ permitted$_{[ACT]}$, shall$_{[ACT]}$ meet$_{[ACT]}$ the$_{[NON]}$ requirements$_{[NON]}$ of$_{[NON]}$ Subsection$_{[REF]}$ 702.08$_{[REF]}$.

(b)

Figure 4.11 Samples of NER results: (a)original text, (b)NER tagged text

Table 4.6 is a confusion matrix of the NER classification results, which explained both original and predicted categories (i.e., ORG, ACT, ELM, STD, REF, and NON). There were 30,109 tokens in the testing set of 1,398 sentences (i.e., 30% of the total labeled data). Although the labeled volume of each category was unbalanced, the classification performance was stable and satisfactory.

Table 4.6 Confusion matrix of NER results

		Actual Categories						
		NON	ORG	ACT	ELM	STD	REF	TOTAL
Predicted Categories	NON	10,360	10	382	527	63	26	11,368
	ORG	14	571	0	28	0	2	615
	ACT	366	0	4,409	82	15	1	4,873
	ELM	694	12	86	9,273	138	22	10,225
	STD	62	0	25	135	1,764	13	1,999
	REF	10	0	2	36	0	981	1,029
	TOTAL	11,506	593	4,904	10,081	1,980	1,045	30,109

Table 4.7 explains the numerical classification performance of each category. F1 score is one of the indexes that measure the accuracy of a model, system, or test (Manning et al., 2008). F1 score simultaneously considers two well-known performance indexes, i.e., precision and recall, by calculating their harmonic means. The developed NER model resulted in the average F1 score of 0.917, indicating that the model was ready for use in real-world applications. Besides, the precision and the recall of the model were calculated as 0.919 and 0.914, respectively, indicating that the model predicted the actual category as accurate as well as extracted actual category information as it was. The training results with 30 different sets of randomly split training, validation, and testing data confirmed the robustness of the model, not overfitted to the specific data set. The results showed an average

of 0.912 F1 scores with a minimum of 0.900 and a maximum of 0.926 (Table 4.8).

Table 4.7 Classification performance for each category

Category	Precision	Recall	F1 Score
NON	0.900	0.911	0.906
ORG	0.963	0.928	0.945
ACT	0.899	0.905	0.902
ELM	0.920	0.907	0.913
STD	0.891	0.882	0.887
REF	0.939	0.953	0.946
AVG	0.919	0.914	0.917

Table 4.8 F1 scores of 30 randomly split data sets

Iteration	F1 Score	Iteration	F1 Score	Iteration	F1 Score
1	0.907	11	0.914	21	0.919
2	0.900	12	0.919	22	0.921
3	0.904	13	0.926	23	0.910
4	0.914	14	0.912	24	0.907
5	0.913	15	0.914	25	0.920
6	0.911	16	0.913	26	0.924
7	0.904	17	0.901	27	0.908
8	0.910	18	0.917	28	0.911
9	0.904	19	0.918	29	0.916
10	0.905	20	0.918	30	0.914

In detail, the classification results of ORG and REF showed the highest accuracy (i.e., 0.945 and 0.946, respectively) despite the least volume of the training set. These results likely occurred because the two categories were written in an extremely structured format. For instance, the ORG category included words such as “Engineer,” “Contractor,” and “Manager,” which would be placed near to causative verbal phrases such as “should submit to,” “must prepare,” and “is responsible for.” Besides, the REF category included words such as “ASTM C 535” (i.e., American Society for Testing Materials), “AASHTO T 245” (i.e., American Association of State Highway and Transportation officials), and “BS EN 12697” (i.e., British and European Standards), which would be placed after prepositional phrases, such as “in accordance with,” “according to,” and “as determined by.” However, the model showed relatively less accurate performance for the categories of ACT and STD (i.e., 0.902 and 0.887, respectively). The ACT category included various verbal phrases, such as “be based on,” “will produce,” and “shall be in accordance with” expressed with multiple verbs and prepositions; thus, the category was difficult to be differentiated by different usage purposes. For instance, the word “in” should be assigned to the ACT category in the case of “… shall be in a good condition …,” but the NON category is more appropriate in the case of “… supply adequate in order to execute ….”

Meanwhile, the STD category included words such as “55%,” “175 °C,” and “2.36 mm,” of which the format and usage patterns of the text data were very close to the words included in the ELM category, such as “60/70 penetrations bitumen,” “a 3 m long straightedge,” and “MC-70 liquid asphalt.” For example, “No. 200 sieve” should be assigned to the ELM category, but the model incorrectly assigned it to the STD category.

4.3.3. Evaluation of Impact of Thesaurus

In order to evaluate the impact of applying the semantic construction thesaurus, the experiment developed two NER models; one with thesaurized Word2Vec embedding, and another without the thesaurus. Every parameter was exactly the same, and the results showed that there was only a slight difference between the performance of the two models. After training the models with 30 randomly split data sets, the averaged F1 scores of the thesaurized model and the other were 0.913 and 0.912, respectively. Despite the theoretical background that the thesaurized text would affect the performance of NLP, the results might be due to the following limitations. First, the semantic construction thesaurus covered 208 terms for replacing the term to its pivot, and in effect, only 21 terms were affected by the thesaurus during developing the Word2Vec model. Thus, the impact of replacing each term to its pivot was insignificant. Second, although the hyperparameters significantly affect the model performance, the experiment utilized the same set of hyperparameters. The impact of applying the semantic construction thesaurus might appear if the hyperparameters were optimized.

4.4. Summary

In this chapter, the second objective of this dissertation was covered, which is to recognize construction keywords automatically regardless of sentence structure. Theoretical backgrounds for the proposed methods (e.g., NER, RNN, Bi-LSTM, and CRF) were introduced. The authors collaborated with fourteen experienced practitioners to understand the information needs for risk management. The practitioners acknowledged that the following information types are crucial to answer to the identified questions: (1) persons and organizations in charge (i.e., ORG), (2) activities required (i.e., ACT), (3) construction and installation items (i.e., ELM), (4) quality standards and criteria (i.e., STD), and (5) relevant references (i.e., REF). Six construction practitioners were involved in manually assigning word labels to 4,659 sentences of construction specifications, which were utilized the labeled data for training, validation, and testing the NER model. The input data was thesaurized based on the semantic construction thesaurus that was developed in 'Chapter 3 Analysis of Construction Text .' The developed NER model trained 70% of input data and was tested with the remaining 30%, presenting the average F1 score of 0.917. Being trained 30 different sets of randomly split data, the NER model proved its robustness, not overfitted to the specific data set.

Chapter 5. Identification of Relevant Clauses from Different Construction Specifications

This chapter covers the third and the last objective of this dissertation, which is to propose a relevant clause pairing approach that enables the comparative analysis for different specifications. As described in the '2.3 Limitations of ' section, the differently organized clauses obscure the process of text comparison for specification review. The author addressed the problem by proposing an appropriate analyzed text unit of the construction specification (i.e., clause) and pairing the relevant clauses based on the semantic features. First, the author developed a clause corpus (i.e., a set of text data) by manually extracting text data as described in the '3.2.1 Data Collection' section. Then, all of the clauses were embedded to numeric vector space by the Doc2Vec model that learned the semantic features of clauses. Lastly, the relevant clause pairs were identified based on the cosine similarity between Doc2Vec vectors (Figure 5.1). As the proposed methods (i.e., Doc2Vec and Cosine similarity calculation) are based on unsupervised learning, the approach identifies the most relevant clauses with no need of human efforts on feature extraction or data labeling. In other words, the approach would work well regardless of the differently organized clauses from the different specifications.

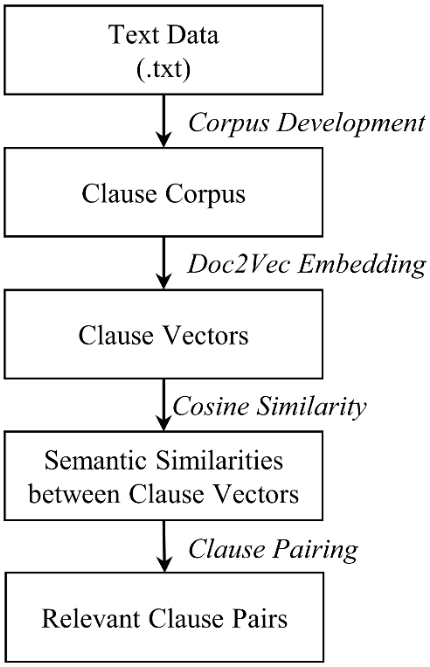


Figure 5.1 Research process of relevant clause pairing

5.1. Research Method: Relevant Clause Pairing

5.1.1. Analyzed Unit of Text Relevance: Clause

Text relevance is a measure of how similar the subjects of two texts are focused on. Identifying the most relevant clause is crucial to the automated specification review, as the qualitative requirements should be reviewed within the same subject area. Many researchers have attempted to develop similar case retrieval systems that identify the most similar text (i.e., document or sentence) (Fan and Li 2013; Al Qady and Kandil 2014). Despite the promising performance of the developed systems, they restricted to analyze the text in document-level or sentence-level. This limitation is critical for the specification review process because of the following reasons. First, since the document-level text retains too manifold information, even if the most relevant document was provided, the user should investigate every sentence to find the most relevant requirement. On the other hand, since the sentence-level text provides too specific information, the review process should struggle to figure out which construction item the sentence describes. In order to address these problems, the author suggested the analysis text unit with a clause. The clause-level text, which consists of several continuous sentences, seems to retain the proper amount of information for text comparison.

5.1.2. Text Embedding: Doc2Vec

Because a clause consists of several sentences (i.e., many words), the Word2Vec model that handles words can not be applied to the embedding process of clauses. The author reviewed several prominent embedding techniques for longer text data.

Term Frequency (TF) counts the frequency of each word in each document and considers it as a document vector (Manning et al. 2008). For example, two sentences, “The Contractor should prepare” and “The Engineer should submit,” would be mapped to [1_{the}, $1_{\text{Contractor}}$, 1_{should}, 1_{prepare}, 0_{Engineer}, 0_{submit}] and [1_{the}, $0_{\text{Contractor}}$, 1_{should}, 0_{prepare}, 1_{Engineer}, 1_{submit}]. A slightly enhanced text embedding technique is Term Frequency-Inverse Document Frequency (TF-IDF), which normalizes the common terms that are spread throughout almost every document and provide less importance to those words (Joulin et al. 2016; Zhou and El-Gohary 2016), such as “a,” “an,” and “the.” For example, two documents above would be mapped to [0.5_{the}, $1_{\text{Contractor}}$, 0.5_{should}, 1_{prepare}, 0_{Engineer}, 0_{submit}] and [0.5_{the}, $0_{\text{Contractor}}$, 0.5_{should}, 0_{prepare}, 1_{Engineer}, 1_{submit}], respectively, so that the meaningful terms (e.g., “Contractor” and “Engineer”) could be used to characterize each document more effectively. While these frequency-based approaches can conduct text embedding efficiently (i.e., just counting the frequency of each term), they have a critical limitation in that they do not take into account the context information (i.e., the order or sequence of words).

Doc2Vec is a machine learning-based text embedding technique that represents longer text (i.e., sentence, paragraph, or document) into a dense numeric vector (Lau and Baldwin 2016; Le and Mikolov 2014; Lee et al. 2016a). Since the architecture similar to Word2Vec, the Doc2Vec model also learns the distributed representation of words within every sentence. The Doc2Vec model provides two kinds of learning architectures that are similar to the architectures of Word2Vec: Paragraph Vector with Distributed Memory (PV-DM) and Paragraph Vector with Distributed Bag of Words (PV-DBOW). In PV-DM, the model (1) initializes a document vector, (2) appends it to the word vectors from the document, (3) averages both of the document vector and the word vectors except one word as a context vector, and (4) adjusts the values of each vector, so the context vector becomes similar with the vector of the excepted word, repeatedly with other words (Figure 5.2(a)). In PV-DBOW), the model tries to predict every word vector using only the document vector as a context vector (Figure 5.2(b)). The PV-DBOW architecture is known to be faster on learning and predict the document vector better on short input text (i.e., a few dozens of words).

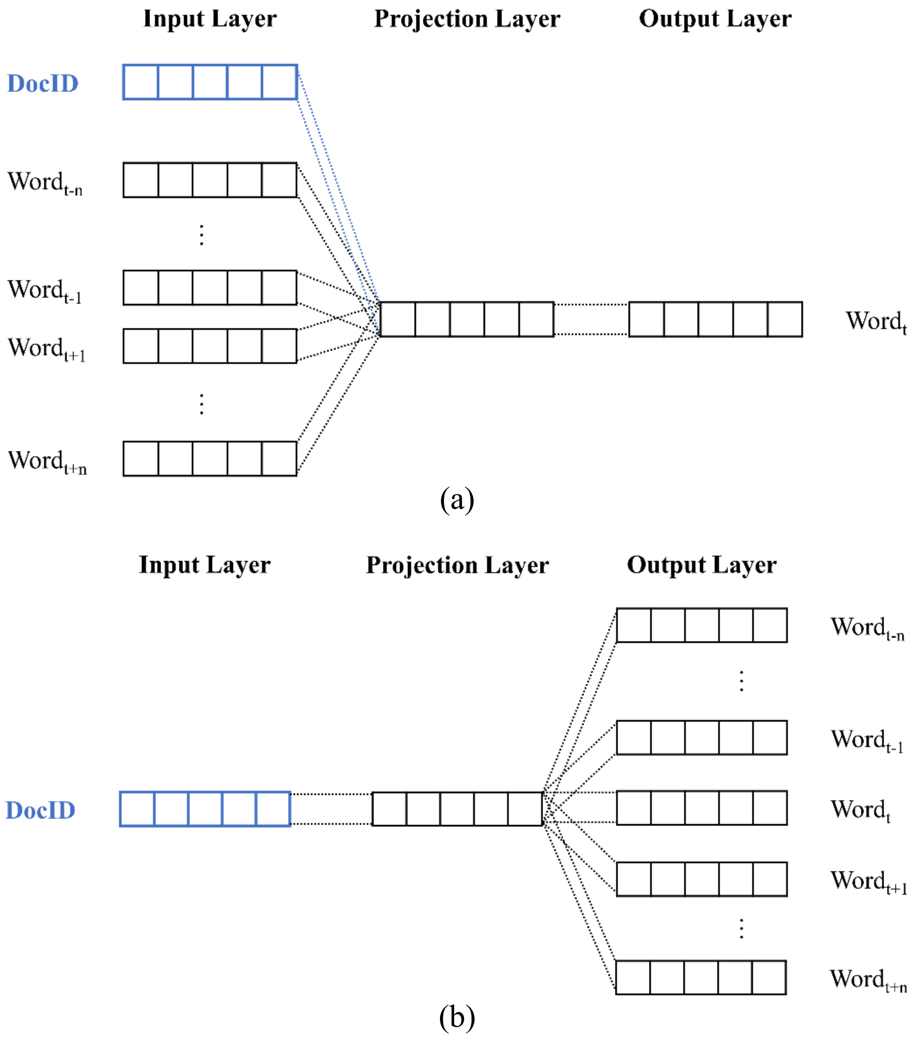


Figure 5.2 Doc2Vec architecture: (a)PV-DM, (b)PV-DBOW

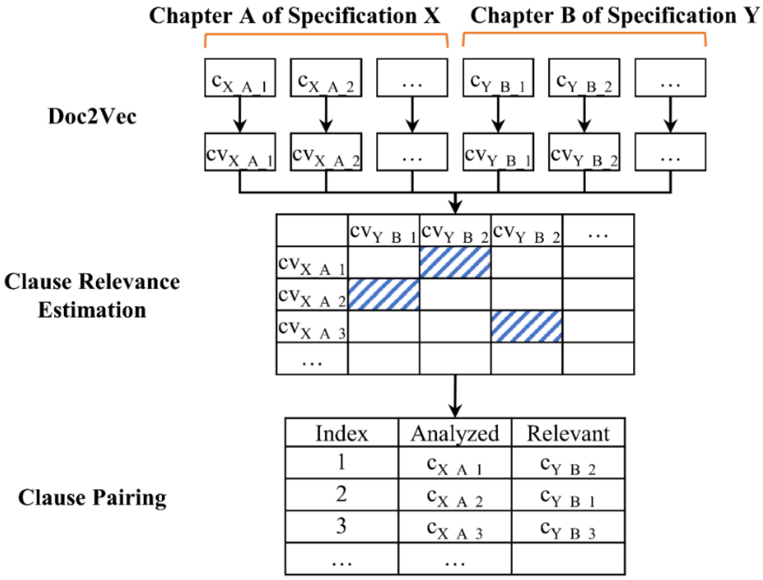
5.1.3. Cosine Similarity

Cosine similarity computes the distance between two vectors based on the inner value of the angle, not the straight distance (Croft et al. 2010). Equation 5.1 describes the detailed mathematics of the cosine similarity, where A and B indicate a vector, respectively, and n indicates the dimension of the vectors. Cosine similarity is commonly utilized in NLP for its satisfying representation of text similarity. As the text vectors are represented in a virtual vector space with considerably high dimensionality (i.e., usually 50 to 500, in this dissertation, 200), the quantitative distances between the vectors would not have any meaning on its own element values. Instead of the other similarity methods (e.g., Euclidean, Mahalanobis, and Manhaton) that are based on the distance between two vectors, the Cosine approach provides the angle similarity that can be interpreted as topic similarity of the text.

$$Cosine\ \ Similarity = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (5.1)$$

5.2. Relevant Clause Pairing Framework

The overall framework of the relevant clause pairing process is provided in Figure 5.3. If a field engineer wants to analyze a chapter “A” from a construction specification “X” with a chapter B from a national standard specification “Y,” every clause should be embedded by Doc2Vec model first. Next, the clause relevance that is mainly based on the Cosine similarity between each pair of clause vectors would be calculated. Finally, the most relevant clause (i.e., blue cells in Figure 5.3) from the national standard (i.e., “Y” in Figure 5.3) would be identified for every clause of the analyzed construction specification (i.e., “X” in Figure 5.3)



Index	Analyzed	Relevant
1	$c_{X\_A\_1}$	$c_{Y\_B\_2}$
2	$c_{X\_A\_2}$	$c_{Y\_B\_1}$
3	$c_{X\_A\_3}$	$c_{Y\_B\_3}$
…	…	

Figure 5.3 Relevant clause pairing framework

5.2.1. Development of Clause Corpus and Clause Embedding

The author extracted a total of 2,527 clauses from six specifications, as mentioned in the '3.2.1 Data Collection' section and developed a clause corpus. The corpus provides information on clauses, including the originated specification, chapter number, chapter name, and text sentences, which would be used in the clause pairing process.

The Doc2Vec model was developed by training the corpus of 2,527 clauses based on the PV-DM architecture. The hyperparameters of the model were settled based on the empirical experiments conducted by the author (Table 5.1).

Table 5.1 Hyperparameters of Doc2Vec model

Hyperparameter	Value	Description
Vector Size	500	The dimension of the document vector
Window Size	10	The number of adjacent words used to learn the text distribution
Minimum Count	30	The minimum frequency of each word to learn the distribution
Epochs	200	The number of iterations to learn the training data

The vector size implies the dimension of clause vectors. The window size indicates the number of neighboring words that are considered to learn the text distribution of the clause. Too rarely occurred words, of which frequent was less than the minimum count, were discounted during the training. The epochs represent the number of iterations for the model trained a set of data. The skeleton of the Doc2Vec model is provided in Figure 5.4.

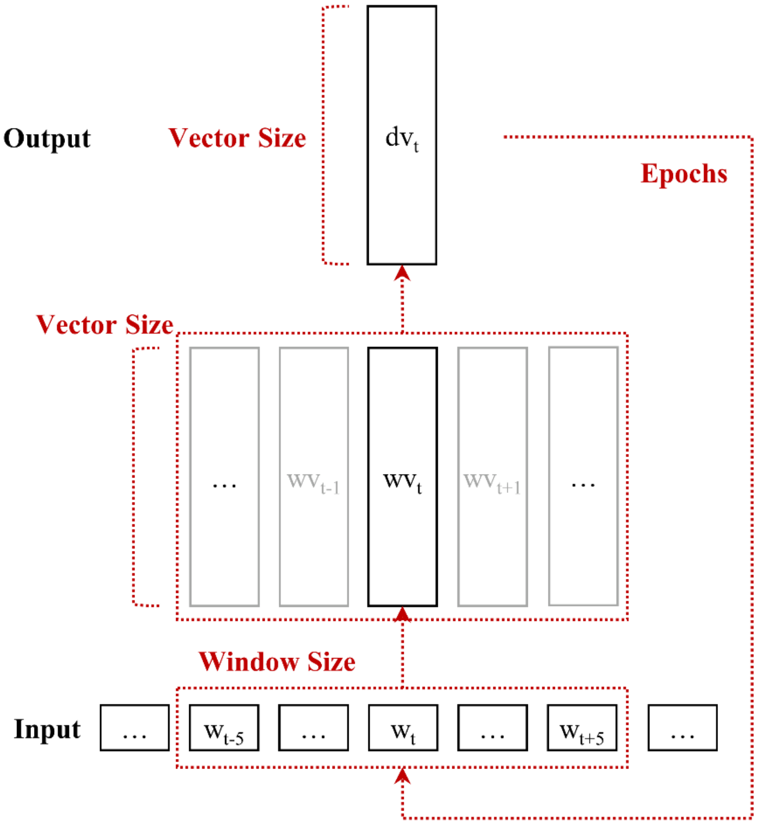


Figure 5.4 Doc2Vec embedding architecture

5.2.2. Estimation of Semantic Relevance of Clauses

Since the vector space of the Doc2Vec model consists of the rational number that includes both of the positive and negative numbers, the text vectors that included some negative numbers might return a negative similarity. To ensure the intuitive of the similarity, the author regularized the Cosine similarity to clause relevance as Equation 5.2, limiting the results between 0 and 1.

$$Clause\ \ Relevance = \frac{(Cosine\ \ Similarity+1)}{2} \qquad (5.2)$$

5.3. Results of Relevant Clause Pairing

5.3.1. Results of Clause Embedding

The Doc2Vec model returned a unique vector for 2,527 clauses. Likewise to the evaluation issue of the Word2Vec, the Doc2Vec embedding results were evaluated in the qualitative approach, since the model is based on unsupervised learning architecture. The author randomly sampled several sentences and investigated the similarity between each sampled sentence and a new sentence of which a few words were replaced to other words.

For example, a sampled sentence “Coarse aggregate shall be clean and free from organic matter” was embedded to a 500-dimension vector of [0.479, 0.438, 0.097, …, -0.580]. The author prepared two experimental sentences: a sentence of which one word was replaced (i.e., “coarse” to “fine”), and a sentence of which the meaning was the same but differently written (i.e., “No organic matter is allowed in coarse aggregate”). The clause relevance between the sampled sentence and the experimental sentences was calculated as 0.917 and 0.984, respectively. Although the first experimental sentence shared most of the words with the sampled sentence, the requirements were totally different; one corresponds to coarse aggregate, and another corresponds to fine aggregate. Since the sampled sentence and the second experimental sentence (i.e., the same meaning) showed larger relevance than the first experiment, the Doc2Vec model seemed to learn the meaning of each clause successfully.

For another example, a sampled sentence “When cement is used ad mineral filler, it shall meet the requirements of ASTM C150” was embedded to a vector of [0.160, 1.106, 1.342, …, -0.206]. Similar to the previous experiment, the author prepared two experimental sentences: a sentence that the word “mineral filler” was replaced with “asphalt binding,” and a new sentence “The mineral filler with cement should follow ASTM C150.” The clause relevance between the sampled sentences and the experimental sentences were calculated as 0.915 and 0.985, respectively, which also supported that the Doc2Vec model is developed finely.

5.3.2. Identification of Relevant Clauses

The author analyzed the construction specification that was used in 2014 at the Qatar construction site (i.e., QCS 2014) by comparing the clauses with other specifications. The Qatar specification that was written in 2010 and the national standard specification from Connecticut, USA, were selected as the relevant specifications. Particularly, the author utilized the section 5 (i.e., ‘Asphalt Work’) of Chapter 6 (i.e., ‘Road Works’) from the QCS 2014 and the QCS 2010, and Section 4 (i.e., ‘Bituminous Concrete Materials’) of

Chapter 19 (i.e., ‘Material Section’) from the Connecticut, USA. Note that this dissertation notated the clauses in the format of ‘COUNTRY_STATE_ YEAR_CHAPTER ID_SECTION ID.’ For example, the analyzed construction specification is called ‘Qatar_Qatar_2014_06_05,’ and the selected relevant specifications are called ‘Qatar_Qatar_2010_06_05’ and ‘United States_Connecticut_2018_19_04,’ respectively.

First, a total of 77 clauses from the ‘Qatar_Qatar_2014_06_05’ were paired with the most relevant clause from the “Qatar_Qatar_2010_06_05.” Due to the absence of a national standard in Qatar, the author regarded ‘Qatar_Qatar_2010_06_05’ as a national standard, and analyzed ‘Qatar_ Qatar_2014_06_05’ (Table 5.2).

Table 5.2 Result of relevant clause pairing (‘Qatar_2014’ and ‘Qatar_2010’)

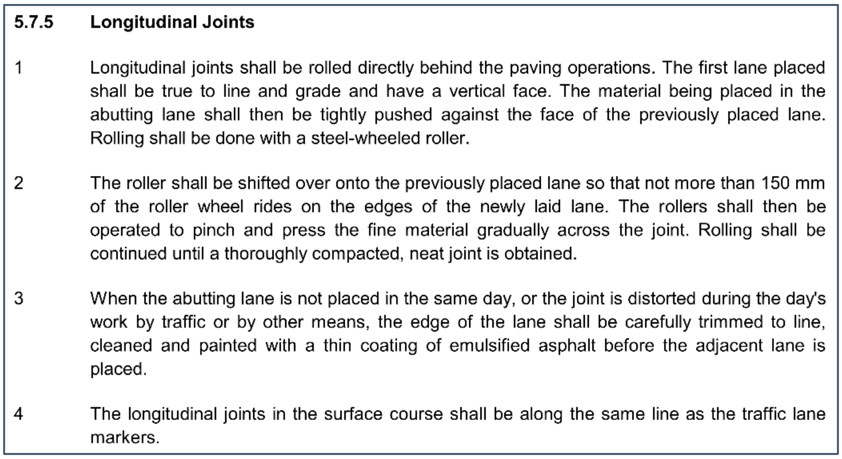
Index	**Clause ID (Qatar_Qatar_2014_06_05)**	**Clause ID (Qatar_Qatar_2010_06_05)**	**Clause Relevance**	**Paired**	**Evaluation (Correct or Not)**
1	01_01	01_01	0.982	O	O
2	01_02	08_01	0.767	X	X
3	01_03	02_04	0.763	X	O
4	01_04	01_03	0.744	O	X
5	01_05	01_04	0.891	O	O
6	02_01	04_04	0.716	X	X
7	02_02	02_01	0.924	O	O
8	02_03	02_01	0.916	O	O
9	02_04	02_02	0.811	X	X
10	02_05	02_03	0.940	O	O
11	02_06_01_01	02_01	0.761	X	O
12	02_06_01_02	08_06	0.716	X	O
13	02_07	12_02	0.832	X	X
14	02_08	13_02	0.933	O	O
15	02_09	01_05	0.971	O	O
16	02_10	01_06	0.928	O	O
17	03_01	07_02	0.926	O	O
18	03_02	07_01	0.735	X	O
19	03_03	11_01	0.790	X	O
20	04	08_01	0.816	X	X
21	05	08_06	0.823	X	X
22	06	11_01	0.754	X	O

23	07_01	09_01	0.927	O	O
24	07_02	09_02	0.870	O	O
25	07_03	09_03	0.962	O	O
26	07_04	09_04	0.976	O	O
27	07_05	09_05	0.957	O	O
28	07_06	09_06	0.975	O	O
29	07_07	09_07	0.967	O	O
30	07_08	09_08	0.963	O	O
31	07_09	09_09	0.943	O	O
32	07_10	09_10	0.969	O	O
33	08	10	0.934	O	O
34	09_01	12_01	0.945	O	O
35	09_02	12_03	0.967	O	O
36	09_03	12_04	0.949	O	O
37	09_04	12_05	0.947	O	O
38	09_05	12_06	0.936	O	O
39	10_01	13_01	0.963	O	O
40	10_02	13_03	0.962	O	O
41	10_03	13_04	0.964	O	O
42	10_04	13_05	0.932	O	O
43	10_05	13_06	0.979	O	O
44	11_01	11_05	0.978	O	O
45	11_02	11_07	0.862	O	O
46	11_03	11_05	0.763	X	O
47	11_03_01	11_06	0.868	O	O
48	11_03_02	11_05	0.727	X	O

49	12	11_07	0.920	O	O
50	13_01	08_01	0.956	O	O
51	13_02	08_02	0.980	O	O
52	13_03	08_03	0.949	O	O
53	13_04	08_04	0.953	O	O
54	13_05	08_05	0.943	O	O
55	13_06	08_06	0.918	O	O
56	14	05	0.952	O	O
57	15_01	06_01	0.949	O	O
58	15_02	06_02	0.926	O	O
59	15_03	06_03	0.927	O	O
60	99_01	04_04	0.722	X	O
61	99_02	08_01	0.751	X	O
62	99_03	11_04	0.737	X	O
63	99_03_02	12_05	0.760	X	O
64	99_03_03	04_12	0.790	X	O
65	99_04	11_01	0.774	X	O
66	99_05	02_02	0.767	X	O
67	99_05_02	04_14	0.714	X	O
68	99_05_03	11_07	0.722	X	O
69	99_05_04	08_01	0.765	X	O
70	99_05_05	04_05	0.745	X	O
71	99_05_06	09_01	0.748	X	O
72	99_05_07	08_01	0.762	X	O
73	99_05_08	08_01	0.813	O	X
74	99_05_09	04_12	0.767	X	O

75	99_06	12_04	0.683	X	O
76	99_06_02	07_01	0.744	X	O
77	99_06_03	03	0.750	X	O

Since the same client wrote the two specifications for the same construction project, most of the clauses shared semantic properties; thus, the results might generally show high scores for clause relevance. For example, both of the specifications included the same clause of ‘Longitudinal Joints’, of which every word was same (Figure 5.5). The relevance between the two clauses (i.e., ‘07_05’ clause from ‘Qatar_Qatar_2014_06_05’ and ‘09_05’ clause from ‘Qatar_Qatar_2010_06_05’) showed to be 0.957, not 1, due to the embedding architecture of PV-DM. Since the developed Doc2Vec model learned the distributed representation of each clause including the clause ID (i.e., the ‘DocID’ of PV-DM architecture in Figure 5.2), the same text from different documents were mapped to different (but extremely close) vectors, which made the relevance score not be 1.



5.7.5 Longitudinal Joints

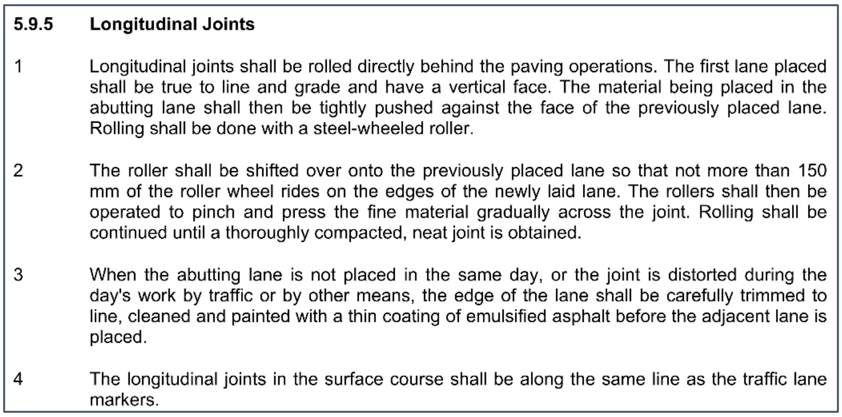
1 Longitudinal joints shall be rolled directly behind the paving operations. The first lane placed shall be true to line and grade and have a vertical face. The material being placed in the abutting lane shall then be tightly pushed against the face of the previously placed lane. Rolling shall be done with a steel-wheeled roller.

2 The roller shall be shifted over onto the previously placed lane so that not more than 150 mm of the roller wheel rides on the edges of the newly laid lane. The rollers shall then be operated to pinch and press the fine material gradually across the joint. Rolling shall be continued until a thoroughly compacted, neat joint is obtained.

3 When the abutting lane is not placed in the same day, or the joint is distorted during the day's work by traffic or by other means, the edge of the lane shall be carefully trimmed to line, cleaned and painted with a thin coating of emulsified asphalt before the adjacent lane is placed.

4 The longitudinal joints in the surface course shall be along the same line as the traffic lane markers.

(a)



5.9.5 Longitudinal Joints

1 Longitudinal joints shall be rolled directly behind the paving operations. The first lane placed shall be true to line and grade and have a vertical face. The material being placed in the abutting lane shall then be tightly pushed against the face of the previously placed lane. Rolling shall be done with a steel-wheeled roller.

2 The roller shall be shifted over onto the previously placed lane so that not more than 150 mm of the roller wheel rides on the edges of the newly laid lane. The rollers shall then be operated to pinch and press the fine material gradually across the joint. Rolling shall be continued until a thoroughly compacted, neat joint is obtained.

3 When the abutting lane is not placed in the same day, or the joint is distorted during the day's work by traffic or by other means, the edge of the lane shall be carefully trimmed to line, cleaned and painted with a thin coating of emulsified asphalt before the adjacent lane is placed.

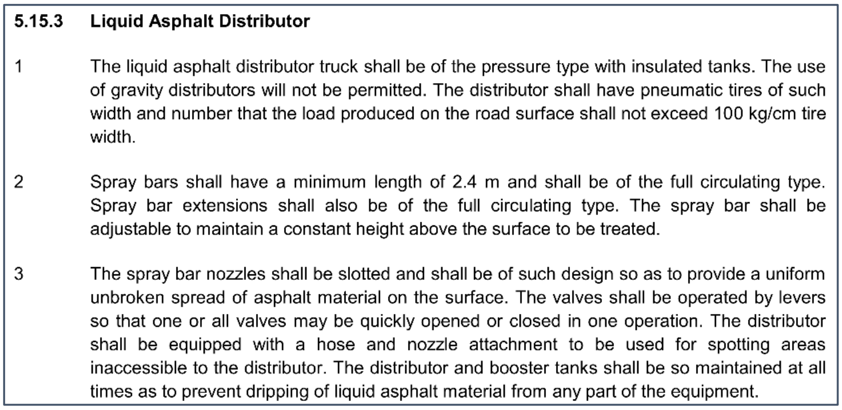
4 The longitudinal joints in the surface course shall be along the same line as the traffic lane markers.

(b)

Figure 5.5 Sample of relevant clause pairing (the same clauses):

(a)‘Qatar_2014,’ (b)‘Qatar_2010’

For another example, both of the specifications included clauses of ‘Liquid Asphalt Distributor’ and ‘Liquid Bitumen Distributor,’ respectively, of which every word was same except for the terms “asphalt” and “bitumen” (Figure 5.6). The relevance between the two clauses (i.e., ‘15_03’ clause from ‘Qatar_Qatar_2014_06_05’ and ‘06_03’ clause from ‘Qatar_Qatar_2010_ 06_05’) showed to be 0.927.



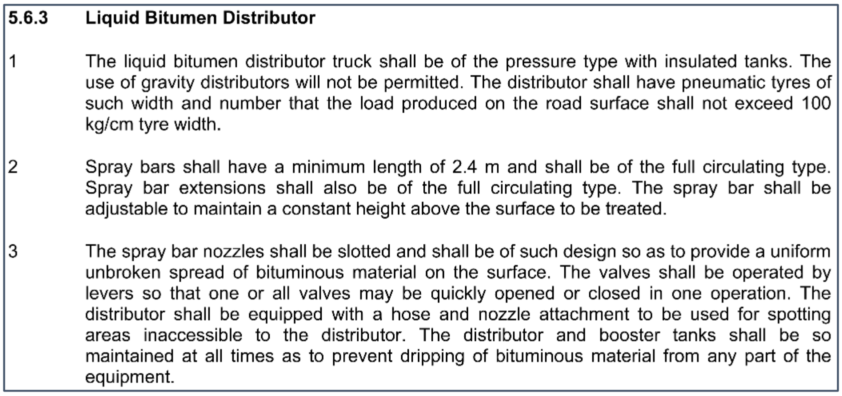
5.15.3 Liquid Asphalt Distributor

1 The liquid asphalt distributor truck shall be of the pressure type with insulated tanks. The use of gravity distributors will not be permitted. The distributor shall have pneumatic tires of such width and number that the load produced on the road surface shall not exceed 100 kg/cm tire width.

2 Spray bars shall have a minimum length of 2.4 m and shall be of the full circulating type. Spray bar extensions shall also be of the full circulating type. The spray bar shall be adjustable to maintain a constant height above the surface to be treated.

3 The spray bar nozzles shall be slotted and shall be of such design so as to provide a uniform unbroken spread of asphalt material on the surface. The valves shall be operated by levers so that one or all valves may be quickly opened or closed in one operation. The distributor shall be equipped with a hose and nozzle attachment to be used for spotting areas inaccessible to the distributor. The distributor and booster tanks shall be so maintained at all times as to prevent dripping of liquid asphalt material from any part of the equipment.

(a)



5.6.3 Liquid Bitumen Distributor

1 The liquid bitumen distributor truck shall be of the pressure type with insulated tanks. The use of gravity distributors will not be permitted. The distributor shall have pneumatic tyres of such width and number that the load produced on the road surface shall not exceed 100 kg/cm tyre width.

2 Spray bars shall have a minimum length of 2.4 m and shall be of the full circulating type. Spray bar extensions shall also be of the full circulating type. The spray bar shall be adjustable to maintain a constant height above the surface to be treated.

3 The spray bar nozzles shall be slotted and shall be of such design so as to provide a uniform unbroken spread of bituminous material on the surface. The valves shall be operated by levers so that one or all valves may be quickly opened or closed in one operation. The distributor shall be equipped with a hose and nozzle attachment to be used for spotting areas inaccessible to the distributor. The distributor and booster tanks shall be so maintained at all times as to prevent dripping of bituminous material from any part of the equipment.

(b)

Figure 5.6 Sample of relevant clause pairing (the similar clauses):

(a)'Qatar_2014,' (b)'Qatar_2010'

The author determined the threshold of clause relevance as 0.8. That is, only if the clause relevance of the most similar pair exceeds the threshold, the model will return the pair is corresponding correctly. Consequently, the automated relevant clause pairing showed a promising accuracy of 89.6% with ‘Qatar_Qatar_2014_06_05’ and ‘Qatar_Qatar_2010_06_05’ (Table 5.3).

Table 5.3 Confusion matrix of clause pairing (‘Qatar_2014’ and ‘Qatar_2010’)

		Actual Pairing Results		
		Relevant	**No Relevant**	**Total**
Predicted Pairing Results	**Relevant**	44	3	47
	No Relevant	5	25	30
	Total	49	28	77

Next, the clauses from the ‘Qatar_Qatar_2014_06_05’ were paired with the most relevant clause from the ‘USA_Connecticut_2018_19_04’ (Table 5.4). The specification was acknowledged to be relevant to the ‘Qatar_ Qatar_2014_06_05’ by the construction practitioners who were involved in the collaboration in ‘4.2 Development of NER Model for Construction Keyword Recognition’ section.

Table 5.4 Result of clause pairing (‘Qatar_2014’ and ‘United States_Connecticut_2018’)

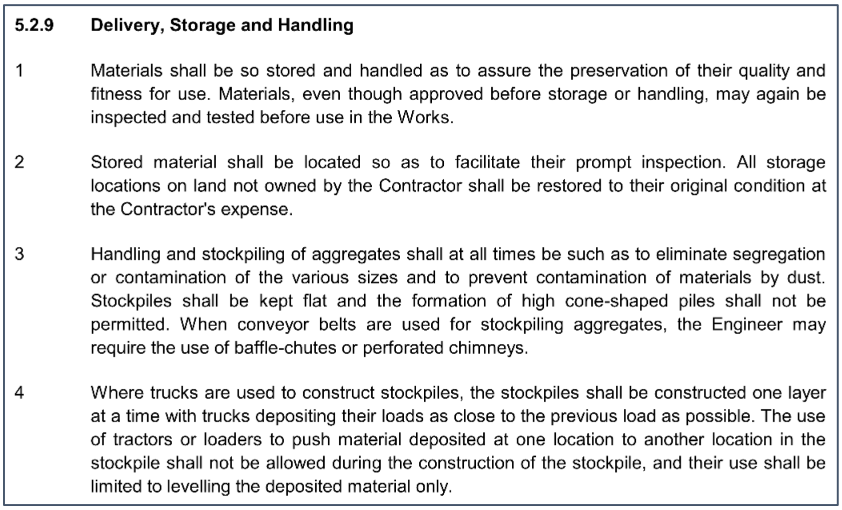
Index	**Clause ID (Qatar_Qatar_2014_06_05)**	**Clause ID (United States_Connecticut_ 2018_19_04)**	**Clause Relevance**	**Paired**	**Evaluation (Correct or Not)**
1	01_01	01_07_02	0.841	X	X
2	01_02	01_04_02_01	0.711	X	O
3	01_03	03_02_02	0.691	X	O
4	01_04	01_06_02_02	0.743	X	O
5	01_05	01_05_02_02	0.835	O	O
6	02_01	01_02_01	0.691	X	O
7	02_02	02_02_01_04	0.745	X	O
8	02_03	01_01_02	0.811	O	O
9	02_04	03_02_02	0.842	X	X
10	02_05	01_05_03_03	0.724	X	O
11	02_06_01_01	01_06_02	0.818	O	O
12	02_06_01_02	01_06_01	0.770	X	O
13	02_07	01_05_03_03	0.759	X	O
14	02_08	01_05_03_02	0.817	O	O
15	02_09	01_04_01_03	0.839	O	O
16	02_10	01_05_02_02	0.751	O	X
17	03_01	03_02_03_02	0.763	O	X
18	03_02	01_10_03	0.724	O	X
19	03_03	01_10_03	0.693	X	O
20	04	01_10_01	0.822	X	X

21	05	03_02_02	0.755	X	O
22	06	03_02_02	0.812	X	X
23	07_01	01_04_04_02	0.808	X	X
24	07_02	01_04_04_02	0.745	X	O
25	07_03	01_06_02	0.729	X	O
26	07_04	01_05_03_01	0.824	X	X
27	07_05	01	0.752	X	O
28	07_06	02_02_01_01	0.732	X	O
29	07_07	01_06_02	0.784	X	O
30	07_08	01_10_01	0.752	X	O
31	07_09	01_03_01	0.802	X	X
32	07_10	01_04_01_03	0.760	X	O
33	08	01_10_04	0.727	X	O
34	09_01	01_04_01_04	0.827	X	X
35	09_02	01_02_02	0.783	X	O
36	09_03	01_04_01_04	0.788	X	O
37	09_04	01_08	0.852	O	O
38	09_05	01_06_02	0.749	X	O
39	10_01	01_04_01_03	0.870	X	X
40	10_02	01_06_02	0.773	X	O
41	10_03	01_08	0.773	X	O
42	10_04	01_08	0.847	X	X
43	10_05	01_04_01_03	0.873	O	O
44	11_01	01_10_04	0.775	X	O
45	11_02	01_10_02	0.816	X	X
46	11_03	01_10_04	0.805	X	X

47	11_03_01	01_05_01_01	0.796	X	O
48	11_03_02	03_01	0.718	X	O
49	12	01_02_02	0.759	O	X
50	13_01	01_10_01	0.907	X	X
51	13_02	01_07_02	0.807	X	X
52	13_03	01_04_04_02	0.786	X	O
53	13_04	01_05_02_02	0.731	X	O
54	13_05	02_02_01_04	0.710	X	O
55	13_06	03_02_02	0.860	X	X
56	14	01_04_01_03	0.663	X	O
57	15_01	01_04_04_01	0.794	X	O
58	15_02	01_06_02	0.699	X	O
59	15_03	01_04_01_01	0.718	X	O
60	99_01	01_04_04_01	0.760	X	O
61	99_02	01_05_03_01	0.800	O	O
62	99_03	03_02_03_04	0.810	O	O
63	99_03_02	01_05_03_01	0.843	O	O
64	99_03_03	03_02_03_01	0.798	O	X
65	99_04	03_02_01	0.786	X	O
66	99_05	03_02_03_01	0.708	X	O
67	99_05_02	01_01_01	0.680	X	O
68	99_05_03	02_02_01	0.707	X	O
69	99_05_04	03_02_03_01	0.720	X	O
70	99_05_05	01_10_07	0.770	X	O
71	99_05_06	01_04_04_01	0.721	X	O
72	99_05_07	01_10_02	0.755	X	O

73	99_05_08	02_02_03	0.758	X	O
74	99_05_09	01_04_04_01	0.666	X	O
75	99_06	01_04_03	0.720	X	O
76	99_06_02	03_02_01	0.840	O	O
77	99_06_03	01_05_02_02	0.790	X	O

The two specifications were acknowledged by experts to be relevant, however, almost clauses showed to be irrelevant, and only a few clause pairs seemed to be relevant. For example, both of the specifications included clauses related to delivery, storage, and handling (Figure 5.7). Although the relevance between the two clauses (i.e., ‘02_09’ clause from ‘Qatar_Qatar_ 2014_06_05’ and ‘01_04_01_03’ clause from ‘United States_Connecticut_ 2018_19_04’) showed to be 0.839, most of the requirements seemed different. With the determined relevance threshold of 0.8, the relevant clause pairing showed a fine performance of 74.0% accuracy (Table 5.5).



5.2.9 Delivery, Storage and Handling

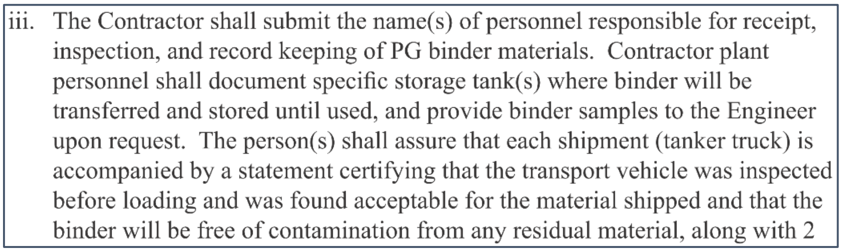
1 Materials shall be so stored and handled as to assure the preservation of their quality and fitness for use. Materials, even though approved before storage or handling, may again be inspected and tested before use in the Works.

2 Stored material shall be located so as to facilitate their prompt inspection. All storage locations on land not owned by the Contractor shall be restored to their original condition at the Contractor's expense.

3 Handling and stockpiling of aggregates shall at all times be such as to eliminate segregation or contamination of the various sizes and to prevent contamination of materials by dust. Stockpiles shall be kept flat and the formation of high cone-shaped piles shall not be permitted. When conveyor belts are used for stockpiling aggregates, the Engineer may require the use of baffle-chutes or perforated chimneys.

4 Where trucks are used to construct stockpiles, the stockpiles shall be constructed one layer at a time with trucks depositing their loads as close to the previous load as possible. The use of tractors or loaders to push material deposited at one location to another location in the stockpile shall not be allowed during the construction of the stockpile, and their use shall be limited to levelling the deposited material only.

(a)



iii. The Contractor shall submit the name(s) of personnel responsible for receipt, inspection, and record keeping of PG binder materials. Contractor plant personnel shall document specific storage tank(s) where binder will be transferred and stored until used, and provide binder samples to the Engineer upon request. The person(s) shall assure that each shipment (tanker truck) is accompanied by a statement certifying that the transport vehicle was inspected before loading and was found acceptable for the material shipped and that the binder will be free of contamination from any residual material, along with 2

(b)

Figure 5.7 Sample of relevant clause pairing: (a)‘Qatar_2014,’ (b)‘United States_Connecticut_2018’

Table 5.5 Confusion matrix of clause pairing (‘Qatar_2014’ and ‘United States_Connecticut_2018’)

		Actual Pairing Results		
		Relevant	**No Relevant**	**Total**
Predicted Pairing Results	**Relevant**	11	15	26
	No Relevant	5	46	51
	Total	49	28	77

In conclusion, the averaged accuracy of the relevant clause pairing would be 81.8%, which means that the user can receive the pairs of corresponding clauses that are paired based on semantic text features. The current accuracy might insufficient to be utilized in the field for the moment. However, the proposed approaches demonstrated the possibility of addressing the differently organized clauses from different documents.

5.4. Summary

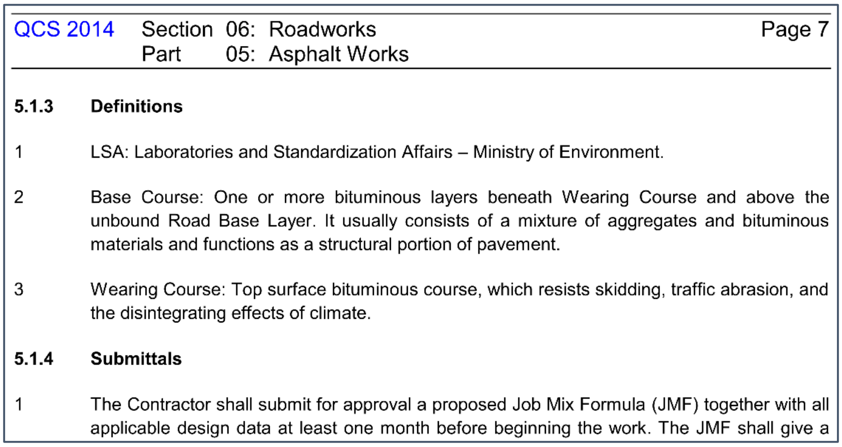
In this chapter, the third objective of this dissertation was covered, which is to propose a relevant clause pairing approach that enables the comparative analysis for different specifications. The author proposed a concept of text relevance (i.e., a measure of how similar the subjects of two texts are being focused on) and a unit of text analysis as a clause (i.e., a group of several continuous sentences) to acquire appropriate information for text comparison. Then, the author extracted a total of 2,527 clauses to develop a corpus and trained the Doc2Vec model based on the PV-DM architecture. With the threshold of clause relevance of 0.8, the averaged accuracy of the relevant clause pairing that is proposed in this dissertation would be 81.8%, which means that the user can receive the pairs of corresponding clauses that are paired based on semantic text features.

Chapter 6. Experimental Results and Discussions

This chapter summarizes the experimental design, process, results, and discussion to confirm the technical feasibility and in-practice applicability of this study. The experiment aimed to compare the proposed approaches and the human for reviewing a part of the construction specification. The comparison is mainly focused on the time efficiency, robustness to subjectivity, and accuracy of detecting erroneous provisions.

6.1. Experimental Design

The author conducted experiments to validate the practical usefulness of the proposed approaches by asking the construction practitioners to review a part of the QCS 2014. Especially, sub-clauses from '5.1.3.1' to '5.2.10.6' (i.e., a total of 58) from 'Qatar_Qatar_2014_06_05,' which were related to material issues of asphalt works, were selected for the experiment because they seemed to be obvious to be reviewed. Although the proposed method provides information in clause-level (e.g., "5.1.3 Definitions" in Figure 6.1), the experiment conducted the review in sub-clause-level (e.g., "5.1.3.2 Base Course: One or more …" in Figure 6.1), and thus reflected the practical review process.



QCS 2014 Section 06: Roadworks Page 7
Part 05: Asphalt Works

5.1.3 Definitions

1 LSA: Laboratories and Standardization Affairs – Ministry of Environment.

2 Base Course: One or more bituminous layers beneath Wearing Course and above the unbound Road Base Layer. It usually consists of a mixture of aggregates and bituminous materials and functions as a structural portion of pavement.

3 Wearing Course: Top surface bituminous course, which resists skidding, traffic abrasion, and the disintegrating effects of climate.

5.1.4 Submittals

1 The Contractor shall submit for approval a proposed Job Mix Formula (JMF) together with all applicable design data at least one month before beginning the work. The JMF shall give a

Figure 6.1 Sample of analyzed specification

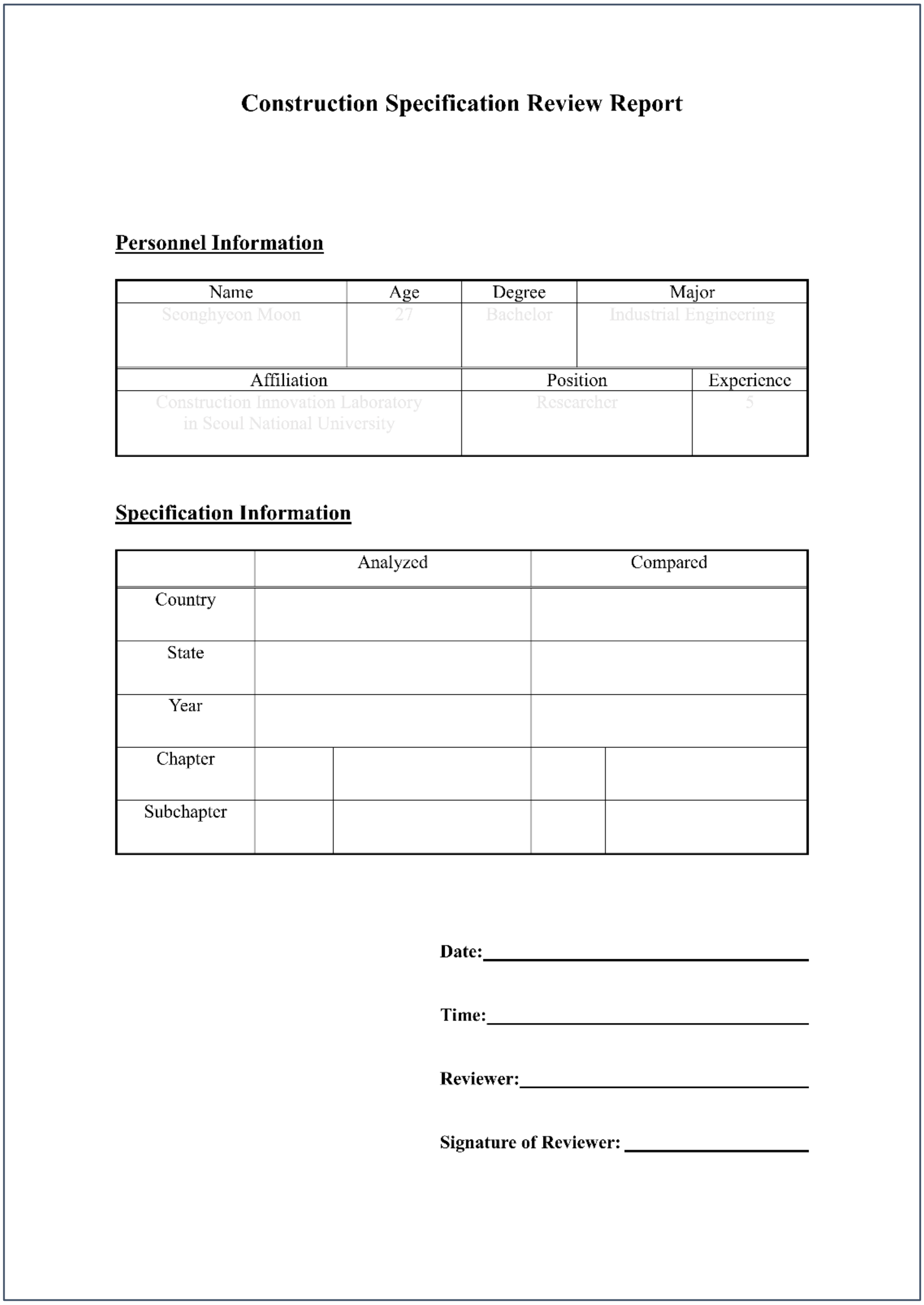
A total of 12 experienced practitioners in the construction industry were involved in the experiments (Table 6.1). The participants consisted of four field engineers (i.e., index of 1, 2, 5, and 7 in Table 6.1) and eight researchers (i.e., index of 3, 4, 6, 8, 9, 10, 11, and 12 in Table 6.1). The average work experience of the participants was six years, when most of the practitioners are required to review a construction specification for the first time in practice. Therefore, the experiment results would demonstrate the usefulness in practice remarkably.

Table 6.1 Personal information of experiment participants

Index	**Work Experience (Year)**	**Specialty**
1	20	Construction Engineering
2	8	Construction Engineering
3	6	Construction Management
4	6	Construction Management
5	5	Architectural Engineering
6	5	Architectural Engineering
7	5	Construction Engineering
8	4	Architectural Engineering
9	4	Construction Management
10	3	Civil & Environmental Engineering
11	3	Civil & Environmental Engineering
12	3	Civil & Environmental Engineering

The experiment provided the specifications of ‘Qatar_Qatar_2010_06_05’ and ‘United States_Connecticut_2018_19_04’ to the participants as relevant specifications, as described in ‘5.3.2 Identification of Relevant Clauses’ section. During the review process, the participants were randomly divided into two groups: the control group and the experiment group. The control group reviewed the provisions of QCS 2014 manually, while the experiment group reviewed the same provisions with the automatic construction specification review method that is proposed in this dissertation.

The report format for the experimental result is provided below. The cover page is to acquire the personal information and the experimental setups (Figure 6.2). The participants would fill the second page with review results (Figure 6.3).



Construction Specification Review Report

Personnel Information

Name	Age	Degree	Major
Seonghyeon Moon	27	Bachelor	Industrial Engineering

Affiliation	Position	Experience
Construction Innovation Laboratory in Seoul National University	Researcher	5

Specification Information

	Analyzed		Compared	
Country				
State				
Year				
Chapter				
Subchapter				

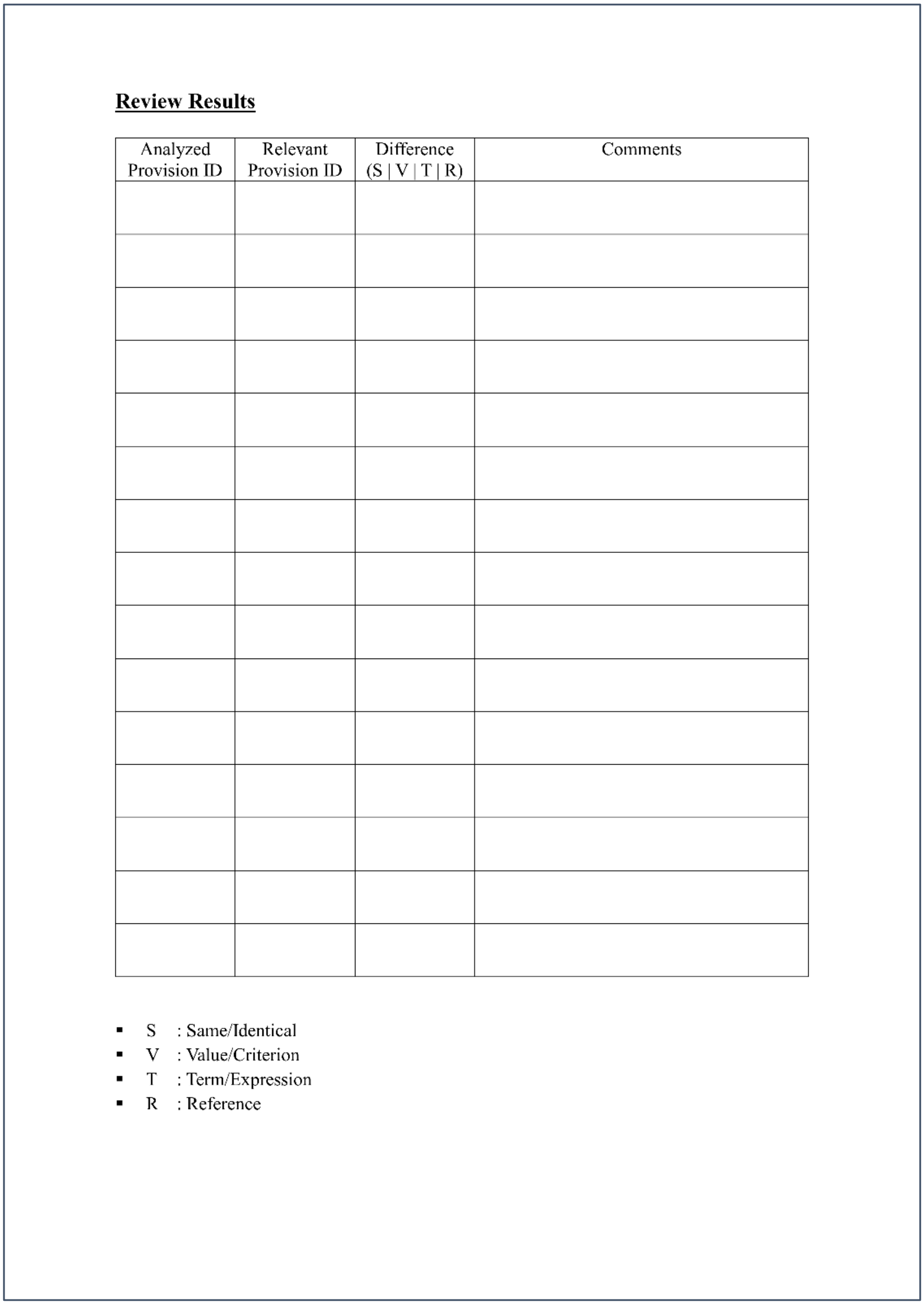
Date:________________________________

Time:________________________________

Reviewer:_____________________________

Signature of Reviewer: ___________________

Figure 6.2 Report cover of experimental result



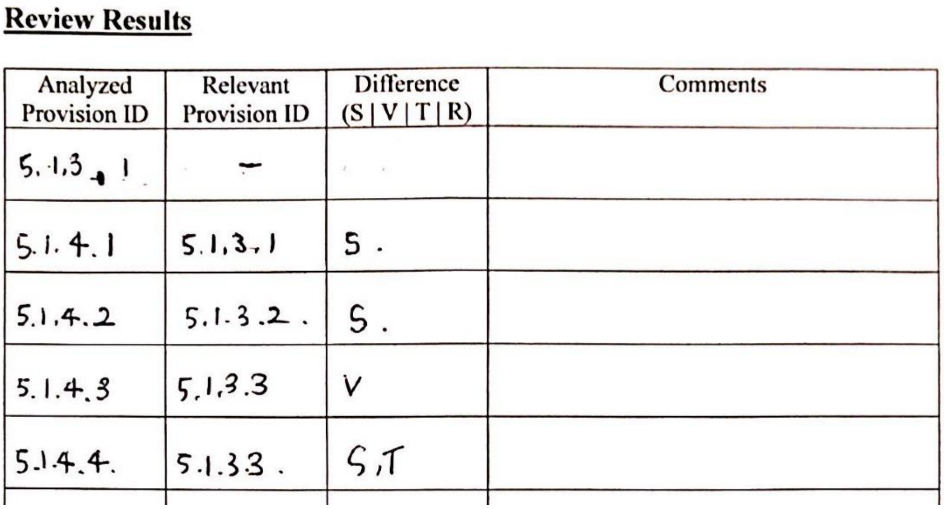
Review Results

Analyzed Provision ID	Relevant Provision ID	Difference (S | V | T | R)	Comments

- S : Same/Identical
- V : Value/Criterion
- T : Term/Expression
- R : Reference

Figure 6.3 Report format of experimental result

The experiment required the participants to review the sub-clauses for two types of information. The first information is that which sub-clause is the most relevant to the current sub-clause of the construction specification. If such sub-clauses that correspond to the current sub-clause exist, the participant should write the ID of the current sub-clause on the first column of the second page and the ID of the relevant sub-clause on the second column. If no clause is appropriate to be paired, the second column would be left empty. Another information is only for the sub-clauses that have relevant sub-clauses; whether the paired sub-clauses are the same or different on the qualitative requirements. If the contents are the same or identical, the third column should be filled with 's,' which indicates 'same/identical.' If the contents are different, the third column should be filled with 'v,' which indicates 'value/criteria.' The other two categories are for optional; 't' stands for 'term/expression' where the sub-clauses utilized different terminologies of expressions, and 'r' stands for 'reference' where the sub-clauses referred to different references. Figure 6.4 illustrates the report of the experimental results with examples. Since there is no relevant sub-clause in the relevant specification for the sub-clause '5.1.3.1,' the second column was empty. Since the sub-clause '5.1.4.1' from the construction specification describes the same requirements with the sub-clause '5.1.3.1' from the relevant specification, they were categorized to ‘s.’ The sub-clause ‘5.1.4.3’ from the construction specification and the sub-clause ‘5.1.3.3’ from the relevant specification seemed to be different, and they were categorized to ‘v.’



Review Results

Analyzed Provision ID	Relevant Provision ID	Difference (S | V | T | R)	Comments
5.1.3.1	-		
5.1.4.1	5.1.3.1	S.	
5.1.4.2	5.1-3.2.	S.	
5.1.4.3	5.1.3.3	V	
5.1.4.4.	5.1.3.3.	S,T	

Figure 6.4 Example of experimental result

6.2. Experimental Results

The author collected 17 experimental results of specification review; 9 reports were compared by ‘Qatar_Qatar_2010_06_05’ and the remaining eight reports were compared by ‘Unite States_Connecticut_2018_19_04.’ The author estimated (1) the time spent for the review process and (2) the review performance of each experiment, regarding the result of the first participant whose work experience is the longest as the correct answers.

6.2.1. Review of QCS 2014 against QCS 2010

To the review of ‘Qatar_Qatar_2014_06_05’ and ‘Qatar_Qatar_2010 _06_05,’ the correct answers figured out 66 pairs for 58 sub-clauses. 44 sub-clauses were paired to 52 relevant sub-clauses, while 14 sub-clauses were not paired to any sub-clause. 40 pairs of sub-clauses showed the same requirements and 12 pairs that showed different requirements. Except for the participant whose experimental result was used as the correct answer, four participants experimented manually (i.e., control group). In comparison, the remaining four participants conducted using the automated method (i.e., experiment group).

The experimental results of the control group were estimated as Table 6.2. They spent 62.75 minutes to review the 58 sub-clauses, and returned disappointing performances with the average precision, recall, and F1 score of 0.619, 0.557, and 0.586, respectively. Considering the analyzed clauses were covered in about ten pages, and the whole construction specification is about 5,000 pages, a field engineer who endeavored to review every requirement would need more than 500 hours under the arithmetical assumption.

Table 6.2 Experimental results of control group (‘Qatar_2014’ and ‘Qatar_2010’)

Participant ID	**Duration (minutes)**	**Precision**	**Recall**	**F1 Score**
5	60	0.623	0.576	0.598
6	66	0.717	0.652	0.683
7	73	0.500	0.470	0.484
10	60	0.636	0.530	0.579
Average	**62.75**	**0.619**	**0.557**	**0.586**

The experiment group spent only 48.75 minutes (i.e., reduced 22.3% of time) to review the same amount of provisions, and the performance was accurate: the average precision, recall, and F1 score of 0.755, 0.705, and 0.728, respectively (i.e., increased 24.2% of F1 score) (Table 6.3).

Table 6.3 Experimental results of experimental group ('Qatar_2014' and 'Qatar_2010')

Participant ID	Duration (minutes)	Precision	Recall	F1 Score
9	63	0.685	0.561	0.617
2	48	0.818	0.818	0.818
4	44	0.800	0.788	0.794
11	40	0.717	0.652	0.683
Average	**48.75**	**0.755**	**0.705**	**0.728**

6.2.2. Review of QCS 2014 against Connecticut of United States

To the review of 'Qatar_Qatar_2014_06_05' and 'United States_ Connecticut_2018_19_04,' the correct answers figured out a pair for every sub-clause. 25 sub-clauses showed to have relevant pairs, while the remaining 33 sub-clauses resulted in being irrelevant to any other sub-clauses. Only 5 pairs of sub-clauses retained the same requirements, and 20 pairs seemed different. The control group and the experiment group included 3 participants for each.

The results of the control group were estimated as Table 6.4. They spent 124.33 minutes to review the given document, and the average precision, recall, and F1 score were 0.521, 0.563, 0.541, respectively.

Table 6.4 Experimental results of control group ('Qatar_2014' and 'United States_Connecticut_2018')

Participant ID	**Duration (minutes)**	**Precision**	**Recall**	**F1 Score**
4	196	0.629	0.672	0.650
6	99	0.441	0.448	0.444
8	78	0.493	0.569	0.528
Average	**124.33**	**0.521**	**0.563**	**0.541**

The results of the experiment group were estimated as Table 6.5. The automated review program assisted the participants to reduce the working hours remarkably, which took only 39.33 minutes (i.e., reduced 68.4% of the time) on average. Although the experimental results showed similar performances to those of the control group with the average precision, recall, and F1 score of 0.531, 0.540, and 0.535, the reduced time might compensate for the insufficient improvement of performance. Besides, the performance variation of the participants was also decreased, and the automated review results are more consistent than the manual results.

Table 6.5 Experimental results of experimental group ('Qatar_2014' and 'United States_Connecticut_2018')

Participant ID	Duration (minutes)	Precision	Recall	F1 Score
3	33	0.500	0.500	0.500
5	50	0.541	0.569	0.555
12	35	0.552	0.552	0.552
Average	**39.33**	**0.531**	**0.540**	**0.535**

6.3. Evaluation of Automated Specification Review

The beneficiaries of the automated construction specification review would be the field engineers who have less or no experience in specification review and the contractors who lack time and employees to review such a plentiful document. In order to evaluate the proposed approach, this dissertation discussed the experimental results in the aspect of time efficiency, accuracy of detecting erroneous provisions, and robustness to subjectivity, which are crucial for site risk management (Table 6.6).

Table 6.6 Evaluation indicators

Validating Indicator	Description
Time Efficiency	The working hours required in reviewing all of the given text
Accuracy of Detecting Erroneous provisions	How close are the results of the proposed approaches to the results of the experienced practitioner
Robustness to Subjectivity	How consistent the review result is

(1) Time Efficiency

First, the proposed approach showed considerable performance in enhancing the time efficiency of the specification review process. According to the experimental results of 'Qatar_Qatar_2014_06_05' and 'Qatar_Qatar_2010_06_05,' the manual review process required more than 20% of working hours without the support of the automated methods. Besides, according to the experimental results of ‘Qatar_Qatar_2014_06_05’ and ‘United States_Connecticut_2018_19_04,’ the automated construction specification review can easily bypass the non-relevant provisions. Reducing the required time for reviewing the specification would facilitate the contractors to focus more on some risky provisions despite the tight schedule of the bidding process.

(2) Accuracy of Detecting Erroneous Provisions

Second, the proposed approach provided higher accuracy on pairing the relevant clauses and determining the differences of each pair, which facilitate to identify erroneous provisions that are not appropriate to the site condition. In the experimental results of ‘Qatar_Qatar_2014_06_05’ and ‘Qatar_Qatar_2010_06_05,’ the less experienced participants showed disappointing performance (i.e., 0.586 of F1 score) compared to the result of the most experienced expert when conducted the review manually. However, the experiment group acquired 0.728 of the F1 score for the correct answers, which indicates the unskilled engineers can perform more similarly to the fully experienced engineer for 0.728 with the support of the proposed method. Although the result of the most experienced professional, which was used as a correct set in the experiment, might contain errors and mistakes, the experiment demonstrated that the developed method could narrow the gap between practitioners. Besides, although most of the sub-clauses from 'Qatar_Qatar_2014_06_05' and 'United States_Connecticut_2018_19_04' did not seem to be relevant, the user can easily bypass the non-relevant provisions, as mentioned above. The developed method might show better performance based on the ontology (i.e., relationships between elements) of the construction specification.

(3) Robustness to Subjectivity

Lastly, the automated review demonstrated to avoid the subjectivity of the reviewer during the review process suggesting consistent results. The manual review produced conflicting results among reviewers due to the difference of experience and capability. For example, the results for the sub-clause '5.1.4.4' of the construction specification were incompatible. One participant answered there is no relevant sub-clause, another answered as the sub-clause '5.1.3.3' from the 'Qatar_Qatar_2010_06_05' should be paired, but the requirements are different, and the other answered as the sub-clause '5.1.3.3' from the 'Qatar_Qatar_2010_06_05' is precisely the same. Meanwhile, the automated review results relatively consistent, returning similar pairs of sub-clauses. The final decision making of the review is for the reviewer, and the proposed method can be utilized as technical support. What is important is that the automated specification review can suggest consistent results to the user and makes the review to be robust to subjectivity. The consistency of results is crucial in risk management as the review errors can be predictable, and the further direction of improvement can be determined.

6.4. Industrial Applications

The experimental results demonstrated the necessity and practical usefulness of the proposed method for automatic specification review. By utilizing the automated method of semantic text comparison, the users can address the semantic conflicts of the specifications (i.e., different vocabulary, different sentence structures, and differently organized clauses), which enables an adequate review of the project requirements.

The developed method facilitates the contractors to review specifications in the early phases of the construction project, which improves the risk management process. Once provided a construction specification from the client, the contractor would convert the document into TXT format, input the data to the automated review program, and select a set of relevant standard specifications to be compared. Then, the program would analyze every provision against the most relevant provision. If erroneous provisions are detected, of which qualitative requirements are different from the national standards, the client and contractor will discuss to correct the provisions to reduce project risk.

The developed method can be used during construction phases repeatedly. Since construction projects commonly last years and the site condition can be changed during the project, the field engineers should review and analyze the requirements frequently. Occasionally, the client might provide a new construction specification with modified or updated clauses. As the program automatically identifies the most relevant clause from other specifications, the field engineers can easily find other clauses that are relevant to the construction specification.

The client who writes project requirements as a construction specification can be a beneficiary of the proposed method. The client can similarly describe the requirements to the relevant clauses of other specifications with much lower efforts to searching. Besides, the proposed method facilitates a preliminary review of the construction specification.

6.5. Summary

In this chapter, the experimental design, process, results, and discussion was covered to confirm the technical feasibility and in-practice applicability of this study. The author conducted experiments to validate the practical usefulness of the proposed approaches by asking the construction practitioners to review a part of the QCS 2014. A total of 12 experienced practitioners (i.e., four field engineers and eight researchers in the construction industry) were involved in the experiments. Being provided the specifications of 'Qatar_Qatar_2010_06_05' and 'United States_Connecticut _2018_ 19_04"'as relevant specifications, the participants were randomly divided into two groups: the control group and the experiment group. The control group reviewed the provisions of QCS 2014 manually, while the experiment group reviewed the same provisions with the automatic construction specification review method. As a result of the first experiment (i.e., 'Qatar_Qatar_2014_06_05' and 'Qatar_Qatar_2010_06_05'), the control group spent 62.75 minutes and returned the average precision, recall, and F1 score of 0.619, 0.557, and 0.586, respectively. The experiment group spent only 77.7% of the time (i.e., 48.75 minutes) to review the same amount of provisions, and the performance was more accurate: the average precision, recall, and F1 score of 0.755, 0.705, and 0.728, respectively. Meanwhile, as a result of the second experiment (i.e., ‘Qatar_Qatar_2014_06_05’ and ‘United States_Connecticut_2018_19_04’), the control group spent 124.33 minutes and returned disappointing performances with the average precision, recall, and F1 score of 0.521, 0.563, and 0.541, respectively. The experiment group only 31.6% of the time (i.e., 39.33 minutes), while the performance was not improved consciously: the average precision, recall, and F1 score of 0.531, 0.540, and 0.535, respectively. Consequently, this dissertation discussed the experimental results in the aspect of time efficiency, accuracy of detecting erroneous provisions, and robustness to subjectivity, which are crucial for site risk management. In addition, the author suggested several industrial applications of the proposed method.

Chapter 7. Conclusions

This chapter summarizes and discusses the research findings and contributions. Opportunities for further improvement and future research works are also discussed.

7.1. Achievements to Research Objectives

The necessity of automation to review the construction specification has resonated with many researchers. However, the previous approaches to automate the review process had limitations in terms of applicability, not fully considering the semantic textual conflicts (i.e., different vocabulary, different sentence structures, and differently organized clauses) among the documents. Since every construction project provides a new construction specification and the specifications have different textual properties, semantic textual analysis is a critical factor in automating the review process of construction specifications with a sufficient level of applicability.

This dissertation developed an automated construction specification review method via semantic textual analysis. The specific objectives were (1) to develop the semantic construction thesaurus to understand the different vocabulary of the specifications using Word2Vec embedding and PageRank algorithm, (2) to recognize construction keywords of qualitative requirements from natural language sentences based on the Named Entity Recognition (NER) model using Word2Vec embedding and the Bi-directional Long Short-Term Memory (Bi-LSTM) architecture with Conditional Random Field (CRF) layer, and (3) to identify the most relevant clause from the standard specification for every clause in the construction specification using Doc2Vec embedding and semantic similarity calculation. The research objectives were addressed and achieved with the following outcomes:

(1) First, the author developed a semantic construction thesaurus to utilize the text comparison methods regardless of different vocabulary among different documents. The research extracted the information of the words that were similarly distributed within the sentence using the Word2Vec model and determined the pivot term for each closed network of converting words. As a result, the construction thesaurus included 208 conversion records.

(2) Second, the author developed a construction keyword recognition model to enable computers to understand the provision contents automatically regardless of the sentence structure. The five information types that are crucial in the risk management process were determined via in-depth collaboration with experienced contractors. Then, the NER model was developed based on RNN architecture, including Bi-LSTM and CRF layers, of which the input was word vectors embedded by Word2Vec. The model showed satisfactory results with an F1 score of 0.917 in classifying the word categories within the sentences. The robustness of the model was verified with 30 different sets of randomly split training and validation data.

(3) Third, the author proposed a relevant clause pairing approach to identify the most relevant text data regardless of the clause hierarchy. The text data were embedded by Doc2Vec to utilize the semantic features in the pairing process. Then, clause relevance that is based on the cosine similarity between the text vectors was calculated to identify the corresponding text. With the threshold of clause relevance of 0.8, the averaged accuracy of the relevant clause pairing that is proposed in this dissertation would be 81.8%, which means that the user can receive the pairs of corresponding clauses that are paired based on semantic text features.

To validate the proposed approaches, the author conducted experiments, of which validating indicators included time efficiency, the accuracy of detecting erroneous provisions, and robustness to subjectivity. The model outperformed the manual review process by reducing working hours, improving performances, and providing more consistent results. In detail, the first experiment that reviewed 'Qatar_Qatar_2014_06_05' and 'Qatar_ Qatar_2010_06_05' reduced 22.3% of time (i.e., 62.75 minutes to 48.75 minutes) and increased 24.2% of performance (i.e., f1 score of 0.586 to 0.728). In addition, the second experiment that reviewed 'Qatar_Qatar_2014_06_05' and 'United States_Connecticut_2018_19_04' reduced 68.4% of the time (i.e., 124.33 minutes to 39.33 minutes) and acquired more consistent performances among participants, despite the slightly decreased f1 score (i.e., 0.541 to 0.535). The experimental results demonstrated that the proposed method is positively necessary and useful to works in practice.

7.2. Contributions

The main contributions of this research include the following: (1) identification of semantic textual conflicts (i.e., different vocabulary, different sentence structures, and differently organized clauses) that disturb the automation in construction document analysis, (2) development of machine learning-based NLP approaches to facilitate the automated construction specification review, (3) proposition of an expandable NLP approach that can be utilized in other types of construction documents; and (4) an in-depth understanding of the construction specification and review process of the document that can lead to the improvement of construction automation and risk management. This dissertation specifically contributed to the body of knowledge by conducting the following studies:

(1) The author identified the three types of semantic textual conflicts in the construction specifications that cause difficulties in the automation of the review process. The different vocabulary, different sentence structures, and differently organized clauses of different documents required additional tasks for the automated approaches that were previously proposed by numerous studies to be utilized in practice. Addressing these limitations would facilitate a fully automated review of the construction documents.

(2) The author developed an automated construction specification review method using widely applied NLP to address the limitations of existing approaches. The developed methods are not restricted to the analyzed data or language and can be utilized in reviewing other construction specifications. Every step of the proposed framework learns the textual features from the new data and automatically provides the user with the required information.

(3) The developed methods can be utilized to analyze other types of construction documents, including contract documents, non-conformance reports, accident reports, and inspection reports, after minor customization of the hyperparameters. As the developed methods are based on machine learning-based NLP that can address the semantic textual conflicts among text data, the approaches are competitive in terms of expandability, as they consist of data-driven methods. Therefore, the research can contribute to risk management and mitigation in the construction industry related to various documents.

(4) This dissertation facilitated an in-depth understanding of the structures and contents of various construction specifications and the review process of the documents. This knowledge can bring further opportunities for improvements in the areas of construction automation and risk management.

The results of this research can support the risk management of construction projects in which reviewing the construction specification is difficult because of the tight schedule of the bidding process, the insufficient number of available professionals, and the large volume of information (over several thousand pages). The contractors can check whether construction requirements on the specification meet the local conditions and are consistent over the entire specification at the bidding stages. They would be able to review the provisions from the specification more efficiently and find erroneous provisions with different standard criteria, which facilitates preparation for potential troubles in advance.

7.3. Opportunities for Improvement and Future Research

Nevertheless, there are potential opportunities for improvement. In order to improve the performance and applicability of the research findings, the following recommendations should be followed:

(1) As data quality directly impacts the analysis performance, the original PDF should be converted into TXT data as clean as possible. In this research, the author had to convert data manually because of the functional limitation of the open-source conversion software. A better PDF-to-TXT conversion technique customized to the construction specifications needs to be developed to replace the manual process and thus process a larger amount of data, accurately and at a low cost.

(2) The more labeled data the model trains, the better the performance the model can achieve. Active learning can be considered to reduce the sample size for training but maintain high accuracy while minimizing the researchers' manual labeling efforts.

(3) Although the proposed approaches were evaluated by comparing the performance with the manual review, it still required testing in practice to enhance and convince the applicability. Currently, the alpha-prototype (i.e., the initial attempt to meet the product requirements) is developed as a web-based user interface (UI) named DICCI, of which the detailed information is provided in Appendix A. Research prototype: DICCI. Besides, the beta-prototype (i.e., the design refinements, function supplementations, debugging, and user experience improvements, which are to be tested in a real case) is under contemplation with the construction practitioners and the UI specialists.

(4) The research findings should be connected to information coming from other types of construction documents (e.g., contract documents, non-conformance reports, accident reports, and inspection reports) or digital data (e.g., Building Information Modeling; BIM) to provide users with more useful information for risk management. For example, future research could analyze the relationship between provisions from contract documents and provisions from BIM, and derive the risky provisions in terms of appropriateness in construction stages. For another example, future research could utilize information coming from the non-conformance reports, accident

reports, and inspection reports as the output of the erroneous provisions of the construction specification and predict the expected results of the provisions.

(5) Furthermore, the developed method can be equipped with information retrieval functions to search keywords, key-phrases, or quantitative values. Eventually, the author suggests developing an automatic specification generation model.

Bibliography

Aitchison, J., Gilchrist, A., and Bawden, D. (2003). *Thesaurus Construction and Use: a Practical Manual*. Routledge.

Caldas, C. H., and Soibelman, L. (2003). "Automating hierarchical document classification for construction management information systems." *Automation in Construction*, 12(4), 395–406.

Chung, S., Lim, S., Chi, S., and Hwang, B.-G. (2017). "Developing a Framework for Identifying Bridge Damage Patterns Based on Text Mining." *2017 MAIREINFRA*, Seoul, South Korea.

Craig, N., and Sommerville, J. (2006). "Information management systems on construction projects: case reviews." *Records Management Journal*, 16(3), 131–148.

Croft, W. B., Metzler, D., and Strohman, T. (2010). *Search Engines: Information Retrieval in Practice*. Pearson Education.

Cucerzan, S., and Yarowsky, D. (1999). "Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence." *Proceedings of 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 90–99.

Cui, Z., Ke, R., Pu, Z., and Wang, Y. (2018). “Deep Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction.” *arXiv preprint arXiv:1801.02143*.

Curran, J. R., and Moens, M. (2002). “Improvements in automatic thesaurus extraction.” *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition*, 59–66.

Fan, H., and Li, H. (2013). “Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques.” *Automation in Construction*, 34, 85–91.

Google Code Archive. (2013). “word2vec.” <https://code.google.com/archive/p/word2vec/> (Nov. 26, 2019).

Hochreiter, S., and Schmidhuber, J. (1997). “Long Short-Term Memory.” *Neural Computation*, 9(8), 1735–1780.

Huang, Z., Xu, W., and Yu, K. (2015). “Bidirectional LSTM-CRF Models for Sequence Tagging.” *arXiv preprint arXiv:1508.01991.*

Jing, Y., and Croft, W. B. (1994). “An Association Thesaurus for Information Retrieval.” *Proceedings of Intelligent Multimedia Information Retrieval Systems and Management*, 146–160.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). “Bag of Tricks for Efficient Text Classification.” *Proceedings of the 15th Conference of* *the European Chapter of the Association for Computational Linguistics (EACL)*, 1–5.

Kerrigan, S. L., and Law, K. H. (2005). “Regulation-Centric, Logic-Based Compliance Assistance Framework.” *Journal of Computing in Civil Engineering*, 19(1), 1–15.

Kim, T., and Chi, S. (2019). “Accident Case Retrieval and Analyses: Using Natural Language Processing in the Construction Industry.” *Journal of Construction Engineering and Management*, 145(3), 04019004.

Kim, Y., Lee, J., Lee, E.-B., and Lee, J.-H. (2020). “Application of Natural Language Processing (NLP) and Text-Mining of Big-Data to Engineering-Procurement-Construction (EPC) Bid and Contract Documents.” *Proceedings of 2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*, 123–128.

Kleinberg, J. M. (1999). “Authoritative sources in a hyperlinked environment.” *Journal of the ACM*, 46(5), 604–632.

Lafferty, J., Mccallum, A., and Pereira, F. C. N. (2001). “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.” *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, 282–289.

Lam, P. T. I., Kumaraswamy, M. M., and Ng, T. S. T. (2007). “International Treatise on Construction Specification Problems from a Legal Perspective.” *Journal of Professional Issues in Engineering Education and Practice*, 133(3), 229–237.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). “Neural Architectures for Named Entity Recognition.” *Proceedings of NAACL 2016*, 260–270.

Lau, J. H., and Baldwin, T. (2016). “An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation.” *Proceedings of the 1st Workshop on Representation Learning for NLP*, 78–86.

Le, Q., and Mikolov, T. (2014). “Distributed Representations of Sentences and Documents.” *Proceedings of the 31st International Conference on Machine Learning*, 1188–1196.

Le, T., Le, C., David Jeong, H., Gilbert, S. B., and Chukharev-Hudilainen, E. (2019). “Requirement Text Detection from Contract Packages to Support Project Definition Determination.” *Proceedings of Advances in Informatics and Computing in Civil and Construction Engineering*, 569–576.

Lee, H., Lee, J. K., Park, S., and Kim, I. (2016a). "Translating building legislation into a computer-executable format for evaluating building permit requirements." *Automation in Construction*, 71(1), 49–61.

Lee, J., and Yi, J.-S. (2017). "Predicting Project's Uncertainty Risk in the Bidding Process by Integrating Unstructured Text Data and Structured Numerical Data Using Text Mining." *Applied Sciences*, 7(11), 1141.

Lee, J., Yi, J.-S., and Son, J. (2016b). "Unstructured Construction Data Analytics Using R Programming - Focused on Overseas Construction Adjudication Cases -." *Journal of the Architectural Institute of Korea Structure & Construction*, 32(5), 37–44.

Lee, J., Yi, J.-S., and Son, J. (2019). "Development of Automatic-Extraction Model of Poisonous Clauses in International Construction Contracts Using Rule-Based NLP." *Journal of Computing in Civil Engineering*, 33(3), 04019003.

Liu, K., and El-Gohary, N. (2017). "Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports." *Automation in Construction*, 81, 313–327.

Manning, C. D., Raghaven, P., and Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McCallum, A., and Li, W. (2003). "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons." *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, 188–191.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space." 1, 1–12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2000). "Distributed Representations of Words and Phrases and their Compositionality." 1, 3111–3119.

Moon, S., Lee, G., Chi, S., and Oh, H. (2019). "Automatic Review of Construction Specifications Using Natural Language Processing." *Proceedings of ASCE International Conference on Computing in Civil Engineering 2019*, 401–407.

Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C., and Xiang, B. (2016). "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond." *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 280–290.

Newman, D., Chemudugunta, C., Smyth, P., and Steyvers, M. (2006). "Analyzing Entities and Topics in News Articles Using Statistical Topic Models." *Lecture Notes in Computer Science*, 93–104.

Noonburg, D. (2017). “pdftotext.”

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). “The PageRank Citation Ranking: Bringing Order to the Web.” *World Wide Web Internet And Web Information Systems*, 66, 1–17.

Al Qady, M., and Kandil, A. (2013a). “Document Discourse for Managing Construction Project Documents.” *Journal of Computing in Civil Engineering*, 27(5), 466–475.

Al Qady, M., and Kandil, A. (2013b). “Document Management in Construction: Practices and Opinions.” *Journal of Construction Engineering and Management*, 139(10), 06013002.

Al Qady, M., and Kandil, A. (2014). “Automatic clustering of construction project documents based on textual similarity.” *Automation in Construction*, Elsevier B.V., 42, 36–49.

Al Qady, M., and Kandil, A. (2015). “Automatic Classification of Project Documents on the Basis of Text Content.” *Journal of Computing in Civil Engineering*, 29(3), 04014043.

Ratinov, L., and Roth, D. (2009). “Design challenges and misconceptions in named entity recognition.” *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL)*, 147–155.

Ryoo, B. Y., Skibniewski, M. J., and Kwak, Y. H. (2010). "Web-Based Construction Project Specification System." *Journal of Computing in Civil Engineering*, 24(2), 212–221.

Salama, D. A., and El-Gohary, N. M. (2013). "Automated Compliance Checking of Construction Operation Plans Using a Deontology for the Construction Domain." *Journal of Computing in Civil Engineering*, 27(6), 681–698.

Sang, E. F. T. K., and De Meulder, F. (2003). "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition." *Proceedings of CoNLL-2003*, 142–147.

Solihin, W., and Eastman, C. (2015). "Classification of rules for automated BIM rule checking development." *Automation in Construction*, 53, 69–82.

Wang, S., and Manning, C. D. (2012). "Baselines and bigrams: Simple, good sentiment and topic classification." *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, 2(July), 90–94.

Wielinga, B. J., Schreiber, A. T., Wielemaker, J., and Sandberg, J. A. C. (2001). "From thesaurus to ontology." *Proceedings of the international conference on Knowledge capture*, 194–201.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *arXiv preprint arXiv:1609.08144*.

Xiao, J., Li, X., Zhang, Z., and Zhang, J. (2018). "Ontology-Based Knowledge Model to Support Construction Noise Control in China." *Journal of Construction Engineering and Management*, 144(2), 04017103.

Zhang, J., and El-Gohary, N. M. (2014). "Automated Reasoning for Regulatory Compliance Checking in the Construction Domain." *Proceedings of Construction Research Congress 2014*, 907–916.

Zhang, J., and El-Gohary, N. M. (2015). "Automated Information Transformation for Automated Regulatory Compliance Checking in Construction." *Journal of Computing in Civil Engineering*, 29(4), 1–16.

Zhang, J., and El-Gohary, N. M. (2016). "Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated

Compliance Checking." *Journal of Computing in Civil Engineering*, 30(2), 04015014.

Zhang, J., and El-Gohary, N. M. (2017). "Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking." *Automation in Construction*, 73, 45–57.

Zhang, R., and El-Gohary, N. M. (2018). "A Clustering Approach for Analyzing the Computability of Building Code Requirements." *Proceeding of Construction Research Congress 2018*, New Orleans, Louisiana, 86–95.

Zhong, B., Li, H., Luo, H., Zhou, J., Fang, W., and Xing, X. (2020a). "Ontology-Based Semantic Modeling of Knowledge in Construction: Classification and Identification of Hazards Implied in Images." *Journal of Construction Engineering and Management*, 146(4), 04020013.

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." *Automation in Construction*, 28, 58–70.

Zhong, B., Xing, X., Luo, H., Zhou, Q., Li, H., Rose, T., and Fang, W. (2020b). "Deep learning-based extraction of construction procedural constraints from construction regulations." *Advanced Engineering Informatics*, 43, 101003.

Zhou, P., and El-Gohary, N. (2016). "Ontology-Based Multilabel Text Classification of Construction Regulatory Documents." *Journal of Computing in Civil Engineering*, 30(4), 04015058.

Zou, Y., Kiviniemi, A., and Jones, S. W. (2017). "Retrieving similar cases for construction project risk management using Natural Language Processing techniques." *Automation in Construction*, 80, 66–76.

국문 초록

의미기반 텍스트 분석을 통한 건설공사 시방서 자동 검토

문성현

서울대학교 대학원
건설환경공학부

건설 프로젝트의 리스크 관리를 위해서는 건설공사 시방서의 시공기준이 현장 상황에 적합한지 사전에 검토하는 것이 중요하다. 하지만, 계약 단계의 촉박한 일정, 활용 가능한 전문인력의 부족, 검토해야 하는 다량의 정보 등으로 인해 시방서 검토 과정에 어려움이 존재한다. 또한, 시방서 검토 작업은 수작업으로 진행되기 때문에 시간이 오래 걸리고, 주관적인 해석, 착오, 누락 등의 오류에 취약하다. 건설 문서를 분석하고 사용자가 필요로 하는 정보를 제공하는 다수의 연구 결과가 만족스러운 성능을 보였지만, 서로 다른 문서에 존재하는 텍스트의 의미 모호성을 고려하지 않았다는 점에서 기술적인 개선이 요구된다. 건설공사 시방서는 매 건설 프로젝트마다 작성되며 주기적으로 갱신되기 때문에, 실무자는 서로 다른 어휘, 문장 구조, 조항 구성 등을 가지는 새로운 문서를 매번 새로 분석해야 한다. 건설공사 시방서 검토 작업을 자동화하고 프로젝트 리스크 관리를 지원하기 위해 이러한 텍스트의 특성을 분석하는 연구가 필요하다.

본 연구는 의미기반 텍스트 비교분석을 통한 건설공사 시방서 자동 검토 방법론을 제안한다. 첫 째로, 같은 대상이 시방서 마다 다른 단어로 표현되는 문제를 해결하기 위해, Word2Vec 임베딩 기법과 PageRank 알고리즘을 활용하여 건설어 시소러스를 구축한다. 둘 째로, 서로 다른 형식으로 작성된 문장으로부터 시공기준 정보를 추출하기 위해, Word2Vec 임베딩 기법과 Bi-LSTM 및 CRF 아키텍처를 활용하여 NER 모델을 개발한다. 셋 째로, 서로 다른 시방서로부터 관련성이 높은 조항을 대응하기 위해 Doc2Vec 임베딩 기법과 의미기반 유사도 분석 방법론을 활용하여 조항 대응 모델을 개발한다. 본 연구의 결과는 건설공사 시방서의 모든 조항에 대해 각 조항에 가장 관련성 높은 조항과 해당 조항의 시공기준 정보를 표의 형태로 사용자에게 제공한다.

우선, 첫 번째 연구 목표를 달성하기 위해 Word2Vec 임베딩 기법을 적용하여 유사하게 사용되는 단어들을 분석했고, 각 단어들을 변환하는 중심 단어(pivot term)를 선정했다. 연구에서 수집한 56 개 시방서의 346,950 개 단어(19,346 개 문장)를 분석한 결과, 총 208 개의 단어 변환 규칙을 가지는 시소러스를 구축했다. 다음으로, 두 번째 연구 목표를 달성하기 위해 건설산업 실무자들과의 협업을 통해 리스크 관리 관점에서 중요하다고 여겨지는 5 개의 정보 타입(책임 주체, 작업 내용, 건설공사 객체, 시공기준, 참고문헌)을 선정했다. 4,659 개 문장의 실험 데이터를 사용해 Word2Vec 벡터를 인풋으로 받아 각 단어를 5 개 정보 타입으로 분류하는 NER 모델을 개발했으며, 모델은 클래스 평균 0.917 의 F1 스코어를 보이는 등 우수한 성능을 확보했다. 또한, 30 개의 무작위로 구분된 학습/검증 데이터셋을 통해 NER 모델이 특정한 학습 데이터에 과적합되지 않았다는 것을 증명했다. 마지막으로, 세 번째 연구 목표를 달성하기 위해 수작업으로

구축된 2,527 개의 조항들로부터 Doc2Vec 임베딩 기법으로 의미적 특징을 추출했다. 각 조항에 대응되는 조항을 찾기 위해 코사인 유사도에 기반하여 조항 연관성을 계산했고, 최종 결과는 시방서 검토 작업의 시간을 단축하고, 검토 결과의 품질을 향상시켰으며, 작업자의 주관성을 저감하는 효과를 보였다.

제안된 방법론을 검증하기 위해 본 연구는 자동 검토 모델과 건설 분야 실무자의 시방서 검토 과정 및 결과를 비교 분석했다. 모델의 자동 검토 능력을 평가하기 위해 시방서를 검토하는 데 소요되는 시간, 잘못된 조항을 검출하는 정확성, 검토 결과의 객관성 등 다양한 지표를 활용했다. 검증 결과, 의미기반 텍스트 비교분석 방법론을 활용하여 서로 다른 시방서의 모호한 특성에 따른 검토의 어려움을 해소할 수 있다는 것을 확인했다.

결론적으로, 본 논문은 건설공사 시방서 검토 과정을 자동화하기 위해 텍스트의 의미적 모호성을 분석했다. 건설공사 시방서의 자동화를 저해하는 요소인 텍스트의 의미적 모호성을 정의했고, 머신러닝 기반 자연어 처리 기법을 적용하여 각 문제에 대응했다. 이는 건설 문서를 자동으로 분석하는 연구 분야에서 서로 다른 문서의 의미적 특성을 고려한 첫 번째 시도이다. 제안된 방법은 건설 프로젝트의 초기 단계에 시방서를 검토하려는 실무자, 시공 단계에 각 조항의 내용을 분석하려는 시공자, 새로운 프로젝트 발주를 위해 시방서를 제작하려는 발주처 등 다양한 관점에서 사용된다. 연구 결과는 간단한 처리를 거쳐 계약 문서, 부적합 보고서, 안전사고 보고서, 정밀점검 보고서 등 건설 분야의 다양한 텍스트 데이터에 적용될 수 있다. 또한, 건설공사 시방서의

구조와 검토 과정을 심층적으로 분석함으로써 건설 자동화에 기여하고, 이를 통해 건설 프로젝트의 리스크 대응을 효과적으로 지원할 수 있다.

주요어: 건설관리; 리스크 관리; 건설공사 시방서; 자동 문서 검토; 텍스트 의미적 모호성; 자연어 처리; 기계 학습

학 번: 2015-21288

Appendix A. Research Prototype: DICCI

The author developed a prototype (named DICCI) to visualize the research results, verify the applicability, and discover the requirements for refinement. In order to utilize the automated construction specification review program effectively in the construction site, the data server in the center should collect, manage, and analyze the data, the on-site practitioners should be able to access the analyzed results, and the data and results should be linked intimately. Therefore, the author developed the prototype as a web-based UI to maximize the applicability of the program by addressing the restriction to physical spaces. The UI was developed based on Django that is an open-sourced web application framework based on Python and connected to the database and the Python modules of the research by JSON (JavaScript Object Notation).

A.1. UI Functions

The UI provides three types of analysis, including ‘full analysis’, ‘section analysis’, and ‘sentence analysis’, which indicates analyzing the whole document at a time, analyzing a specific pair of clauses, and analyzing a specific pair of paragraphs (Figure A.1). Note that the notations are different because the UI template was developed at the beginning of the research.

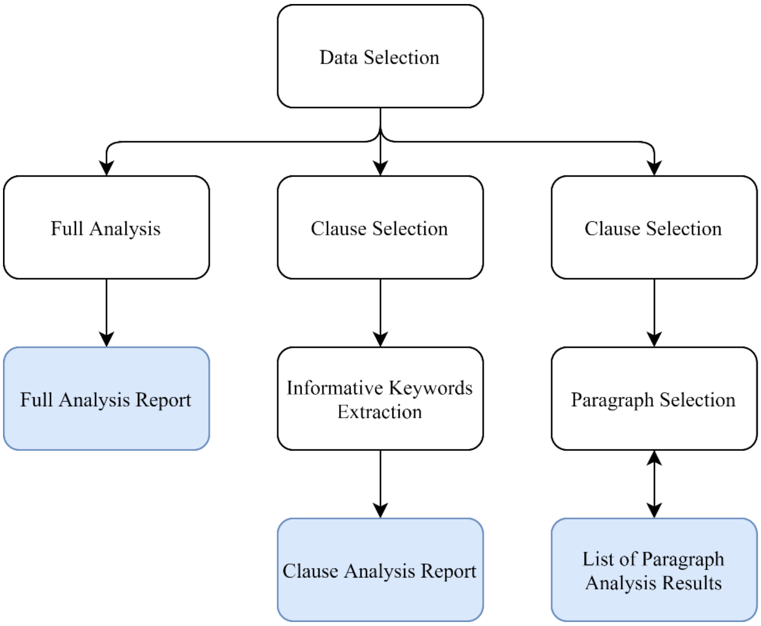
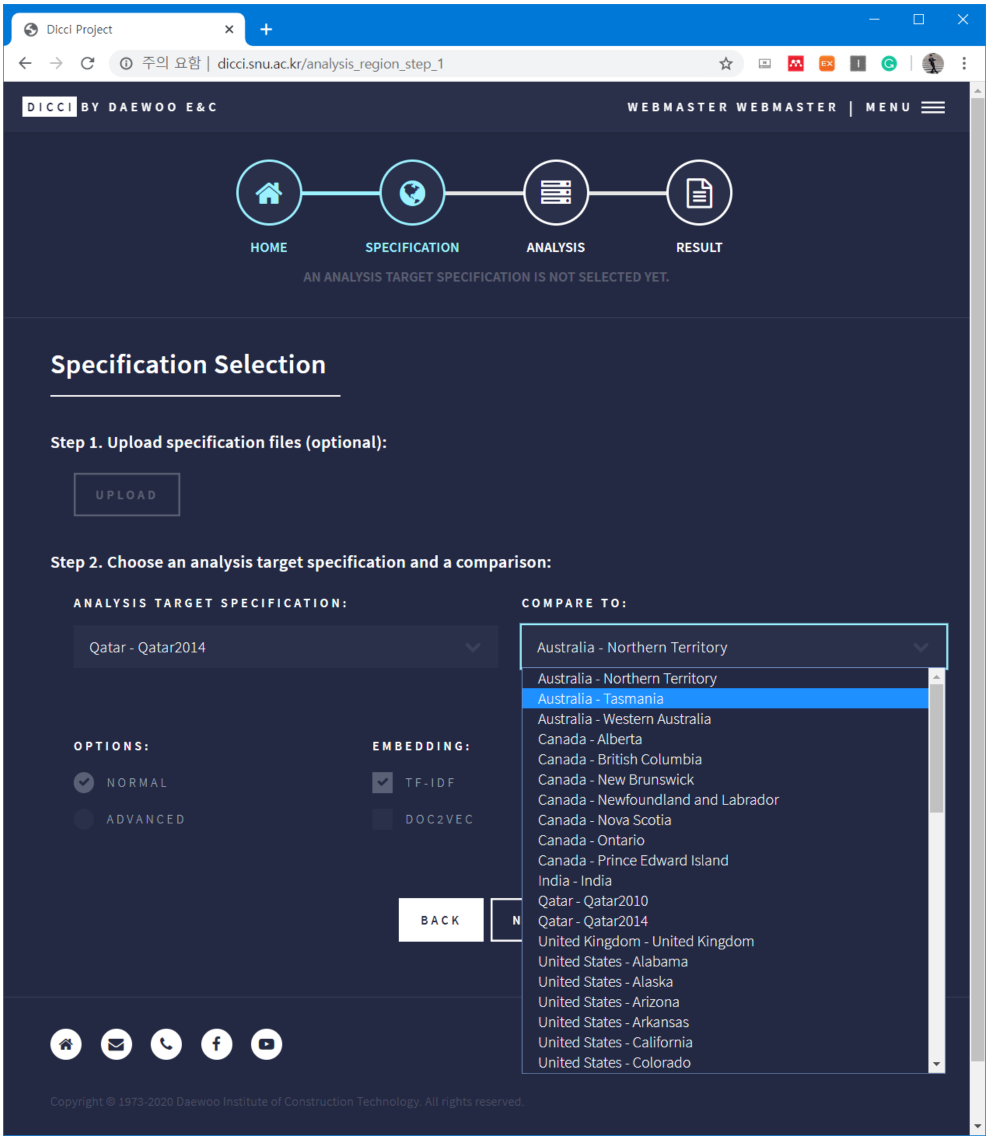


Figure A.1 DICCI functions

A.2. Data Selection for Analysis

The first step of the analysis is to select the data. The user can select the target specification (i.e., the analysis target that the user wants to review) and the comparative specification (Figure A.2). It is interconnected to the database that included 58 construction specifications and 7,820 clauses. Note that only 889 clauses from 5 specifications (i.e., Australia_Tasmania, Qatar_Qatar2010, Qatar_Qatar2014, United Kingdom_United Kingdom, United States_ Connecticut) can be accessed at the moment.



AN ANALYSIS TARGET SPECIFICATION IS NOT SELECTED YET.

Specification Selection

Step 1. Upload specification files (optional):

Step 2. Choose an analysis target specification and a comparison:

ANALYSIS TARGET SPECIFICATION: Qatar - Qatar2014

COMPARE TO: Australia - Northern Territory

OPTIONS:

● NORMAL

○ ADVANCED

EMBEDDING:

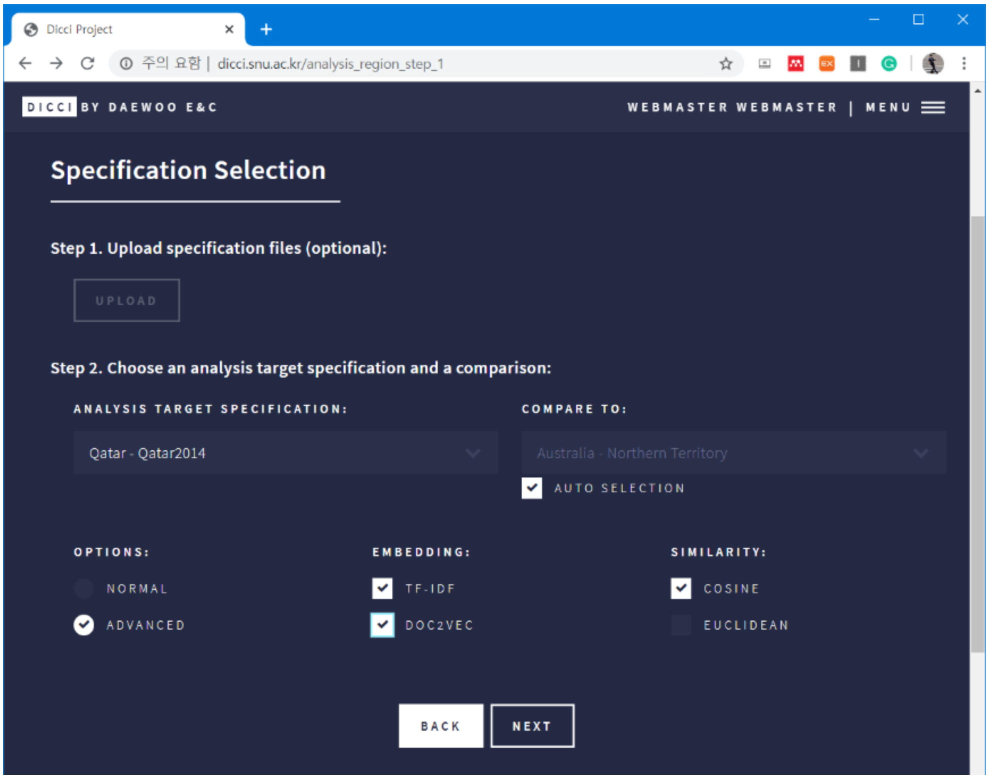
☑ TF-IDF

□ DOC2VEC

Copyright © 1973-2020 Daewoo Institute of Construction Technology. All rights reserved.

Figure A.2 Manual selection of data for analysis

The UI also provides an automatic recommendation of the comparative specification based on the text similarity between the specifications (Figure A.3), of which result asks the user to select one of the five most similar specifications (Figure A.4).


Dicci Project
주의 요함 | dicci.snu.ac.kr/analysis_region_step_1
DICCI BY DAEWOO E&C
WEBMASTER WEBMASTER | MENU
UPLOAD
BACK
NEXT

Specification Selection

Step 1. Upload specification files (optional):

Step 2. Choose an analysis target specification and a comparison:

ANALYSIS TARGET SPECIFICATION: Qatar - Qatar2014

COMPARE TO: Australia - Northern Territory

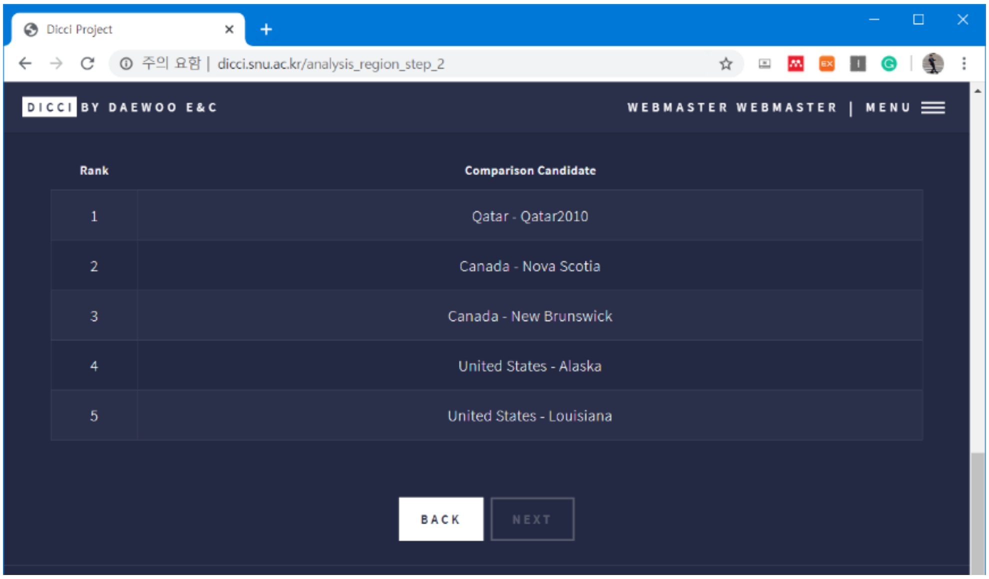
☑ AUTO SELECTION

OPTIONS: ○ NORMAL ● ADVANCED

EMBEDDING: ☑ TF-IDF ☑ DOC2VEC

SIMILARITY: ☑ COSINE ☐ EUCLIDEAN

Figure A.3 Automatic selection of data for analysis

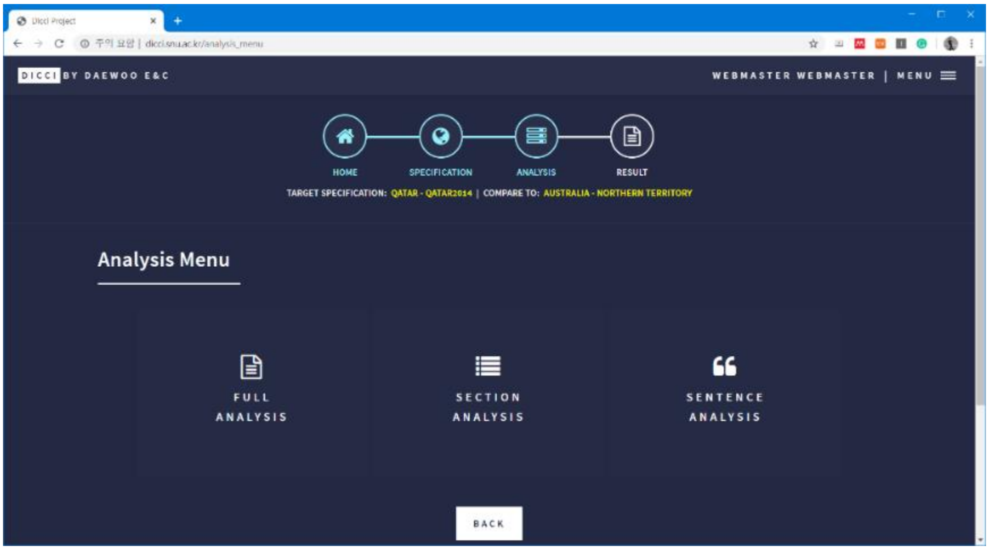

Dicci Project
주의 요함 | dicci.snu.ac.kr/analysis_region_step_2
DICCI BY DAEWOO E&C
WEBMASTER WEBMASTER | MENU
BACK
NEXT

Rank	Comparison Candidate
1	Qatar - Qatar2010
2	Canada - Nova Scotia
3	Canada - New Brunswick
4	United States - Alaska
5	United States - Louisiana

Figure A.4 Results of automatic data recommendation

A.3. Selection of Analysis Type

The second step of the analysis to select the analysis type. The UI provides three types of analysis, as described in the 'A.1. UI Functions' (Figure A.5). Note that the 'full analysis' is the combination of the result of every clause pair from the selected specifications.

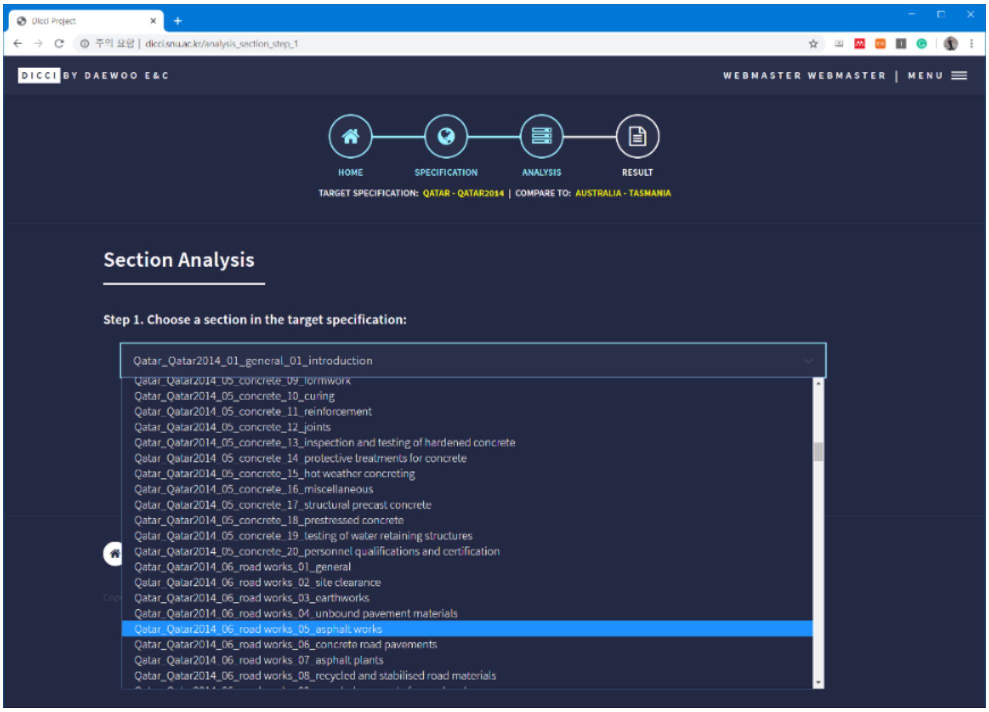


TARGET SPECIFICATION: QATAR - QATAR2014 | COMPARE TO: AUSTRALIA - NORTHERN TERRITORY

Figure A.5 Analysis type selection page

A.4. Clause Pairing

In the 'section analysis', the user selects the target clause (i.e., the clause that the user wants to review) from the selected target specification (Figure A.6). Then, the DICCI recommends the comparative clauses that are the five most similar clauses from the comparative specification to the selected clause.


Dicci Project
주의 요함 | dicci.snu.ac.kr/analysis_section_step_1
DICCI BY DAEWOO E&C
WEBMASTER WEBMASTER | MENU
HOME
SPECIFICATION
ANALYSIS
RESULT

TARGET SPECIFICATION: QATAR - QATAR2014 | COMPARE TO: AUSTRALIA - TASMANIA

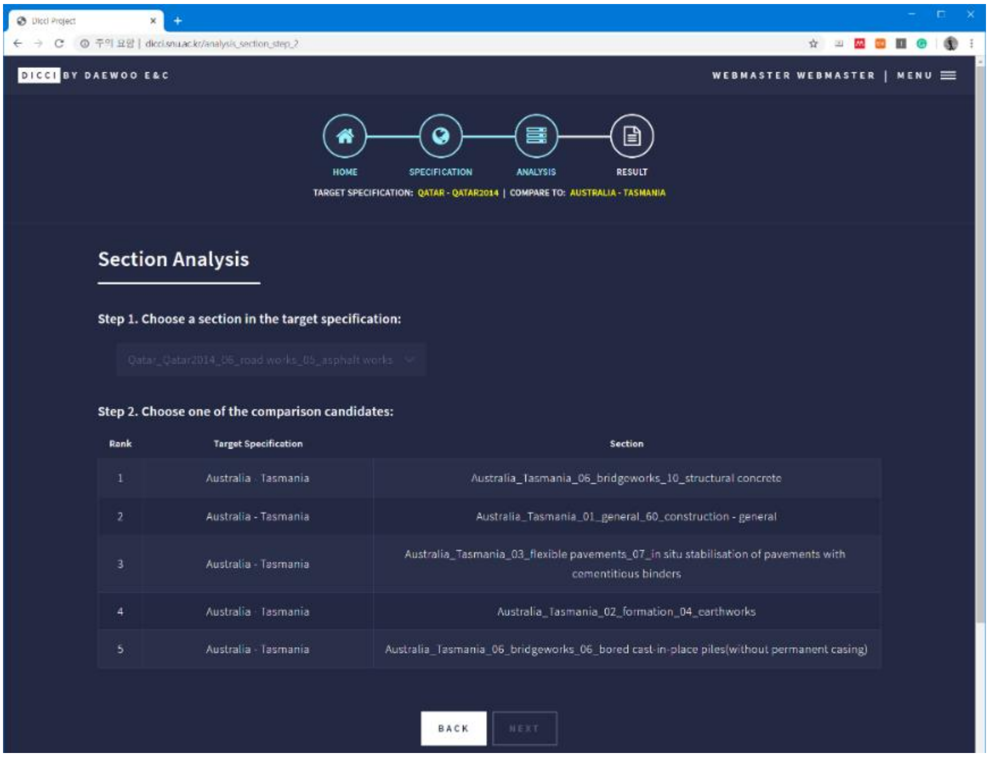
Section Analysis

Step 1. Choose a section in the target specification:

Qatar_Qatar2014_01_general_01_introduction

Qatar_Qatar2014_05_concrete_09_formwork
Qatar_Qatar2014_05_concrete_10_curing
Qatar_Qatar2014_05_concrete_11_reinforcement
Qatar_Qatar2014_05_concrete_12_joints
Qatar_Qatar2014_05_concrete_13_inspection and testing of hardened concrete
Qatar_Qatar2014_05_concrete_14_protective treatments for concrete
Qatar_Qatar2014_05_concrete_15_hot weather concreting
Qatar_Qatar2014_05_concrete_16_miscellaneous
Qatar_Qatar2014_05_concrete_17_structural precast concrete
Qatar_Qatar2014_05_concrete_18_prestressed concrete
Qatar_Qatar2014_05_concrete_19_testing of water retaining structures
Qatar_Qatar2014_05_concrete_20_personnel qualifications and certification
Qatar_Qatar2014_06_road works_01_general
Qatar_Qatar2014_06_road works_02_site clearance
Qatar_Qatar2014_06_road works_03_earthworks
Qatar_Qatar2014_06_road works_04_unbound pavement materials
Qatar_Qatar2014_06_road works_05_asphalt works
Qatar_Qatar2014_06_road works_06_concrete road pavements
Qatar_Qatar2014_06_road works_07_asphalt plants
Qatar_Qatar2014_06_road works_08_recycled and stabilised road materials

Figure A.6 Selection of target clause


Dicci Project
주의 요함 | dicci.snu.ac.kr/analysis_section_step_2
DICCI BY DAEWOO E&C
WEBMASTER WEBMASTER | MENU
HOME
SPECIFICATION
ANALYSIS
RESULT

TARGET SPECIFICATION: QATAR - QATAR2014 | COMPARE TO: AUSTRALIA - TASMANIA

Section Analysis

Step 1. Choose a section in the target specification:

Qatar_Qatar2014_06_road works_05_asphalt works

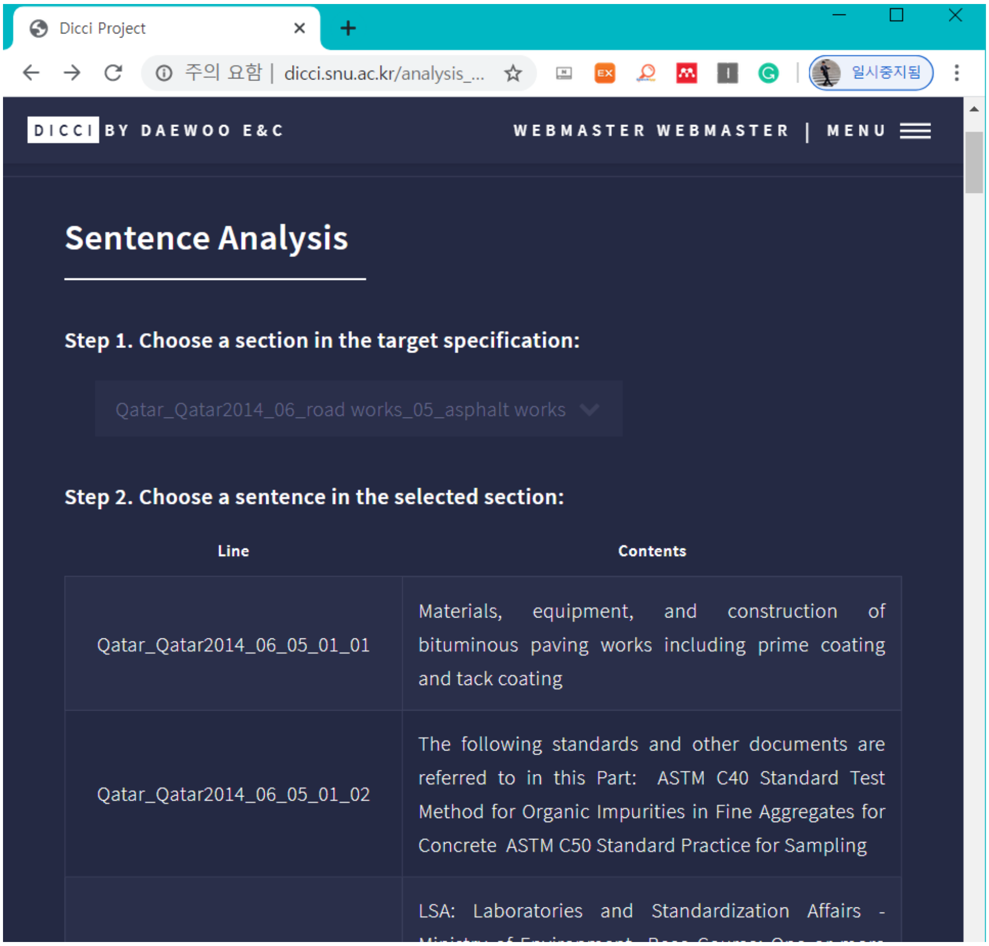
Step 2. Choose one of the comparison candidates:

Rank	Target Specification	Section
1	Australia - Tasmania	Australia_Tasmania_06_bridgeworks_10_structural concrete
2	Australia - Tasmania	Australia_Tasmania_01_general_60_construction - general
3	Australia - Tasmania	Australia_Tasmania_03_flexible pavements_07_in situ stabilisation of pavements with cementitious binders
4	Australia - Tasmania	Australia_Tasmania_02_formation_04_earthworks
5	Australia - Tasmania	Australia_Tasmania_06_bridgeworks_06_bored cast-in-place piles(without permanent casing)

Figure A.7 Recommendation results of relevant clauses

A.5. Paragraph Pairing

If the user selected ‘sentence analysis’ at the analysis type selection step, DICCI requires to select the target paragraph that is to be analyzed (Figure A.8). After the user selected a paragraph, the DICCI returns the most comparative paragraph from the selected clause (Figure A.9).


Dicci Project
주의 요함 | dicci.snu.ac.kr/analysis_...
일시중지됨
DICCI BY DAEWOO E&C
WEBMASTER WEBMASTER | MENU

Sentence Analysis

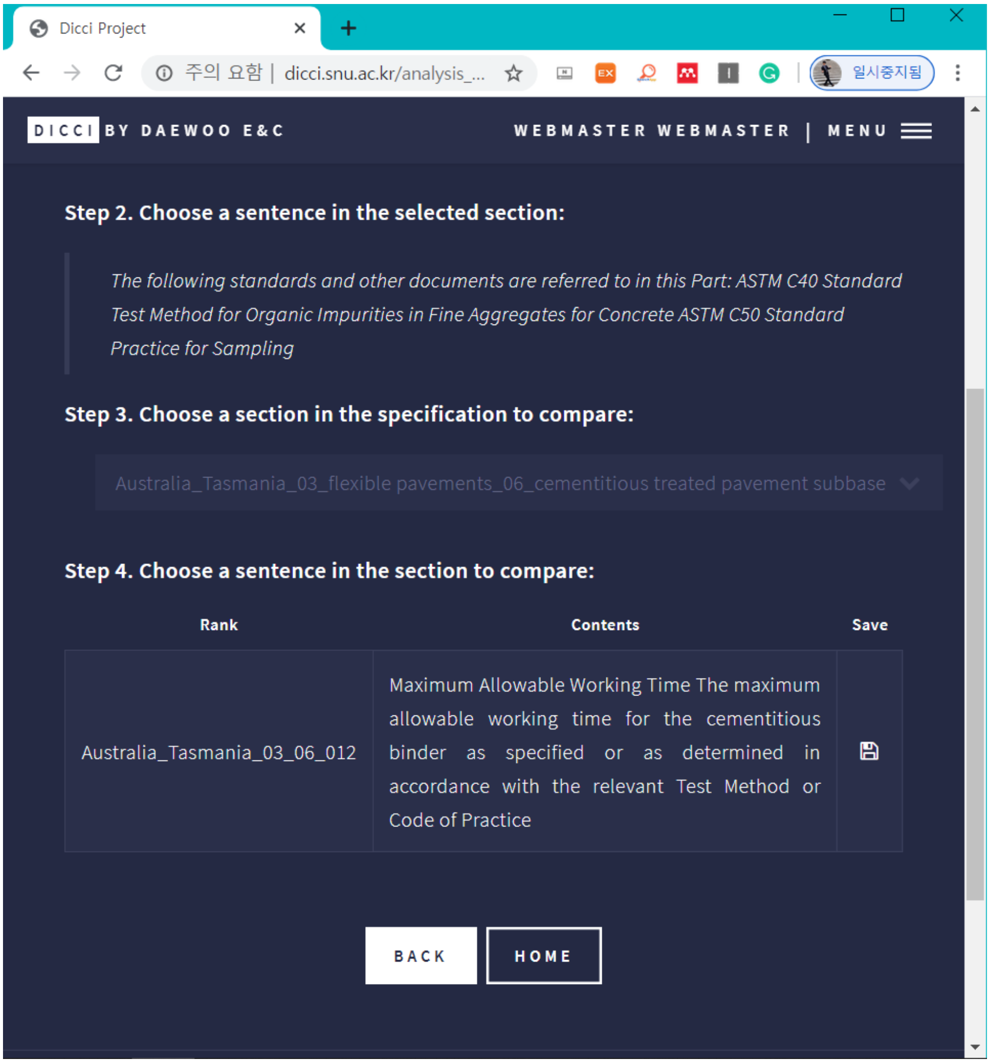
Step 1. Choose a section in the target specification:

Qatar_Qatar2014_06_road works_05_asphalt works

Step 2. Choose a sentence in the selected section:

Line	Contents
Qatar_Qatar2014_06_05_01_01	Materials, equipment, and construction of bituminous paving works including prime coating and tack coating
Qatar_Qatar2014_06_05_01_02	The following standards and other documents are referred to in this Part: ASTM C40 Standard Test Method for Organic Impurities in Fine Aggregates for Concrete ASTM C50 Standard Practice for Sampling
	LSA: Laboratories and Standardization Affairs -

Figure A.8 Selection of target paragraph



Dicci Project
주의 요함 | dicci.snu.ac.kr/analysis_...
일시중지됨
DICCI BY DAEWOO E&C
WEBMASTER WEBMASTER | MENU

Step 2. Choose a sentence in the selected section:

> *The following standards and other documents are referred to in this Part: ASTM C40 Standard Test Method for Organic Impurities in Fine Aggregates for Concrete ASTM C50 Standard Practice for Sampling*

Step 3. Choose a section in the specification to compare:

Australia_Tasmania_03_flexible pavements_06_cementitious treated pavement subbase

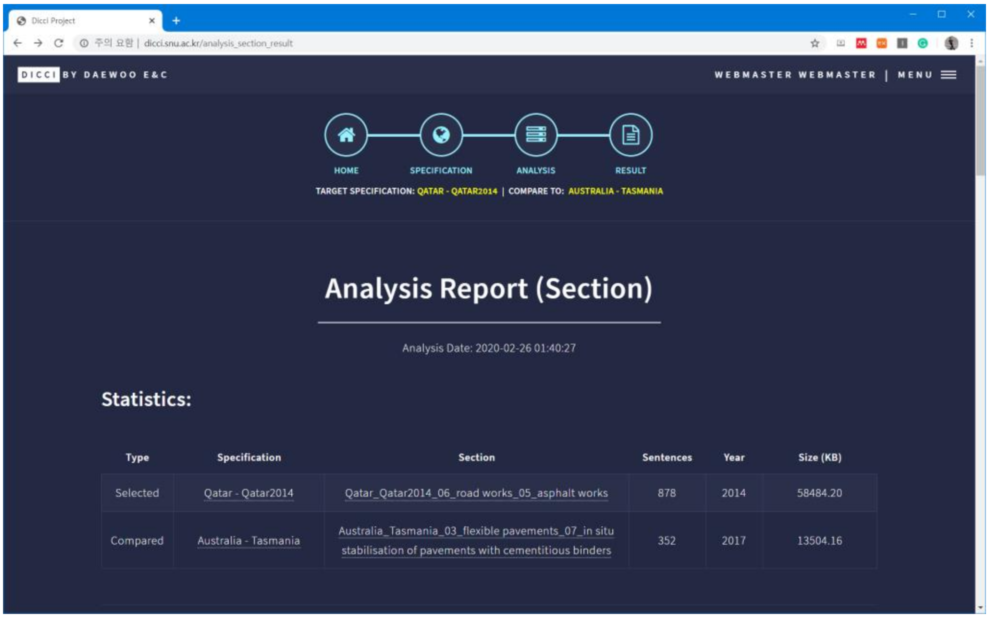
Step 4. Choose a sentence in the section to compare:

Rank	Contents	Save
Australia_Tasmania_03_06_012	Maximum Allowable Working Time The maximum allowable working time for the cementitious binder as specified or as determined in accordance with the relevant Test Method or Code of Practice	💾

Figure A.9 Recommendation results of corresponding paragraphs

A.6. Informative Keywords Extraction

The final step of the analysis is the informative keywords extraction. If the user selected the full analysis or the clause analysis at the 'selection analysis type' step, the UI returns a result page with the statistical information of the two selected text data on the top (Figure A.10). Then, the keywords extraction results are following on at a time. The user can read the original text with a mouseover (Figure A.11). Besides, the user can access the original PDF file by clicking the name of the specification (Figure A.12).


Dicci Project
주의 요함 | dicci.snu.ac.kr/analysis_section_result
DICCI BY DAEWOO E&C
WEBMASTER WEBMASTER | MENU
HOME
SPECIFICATION
ANALYSIS
RESULT

TARGET SPECIFICATION: QATAR - QATAR2014 | COMPARE TO: AUSTRALIA - TASMANIA

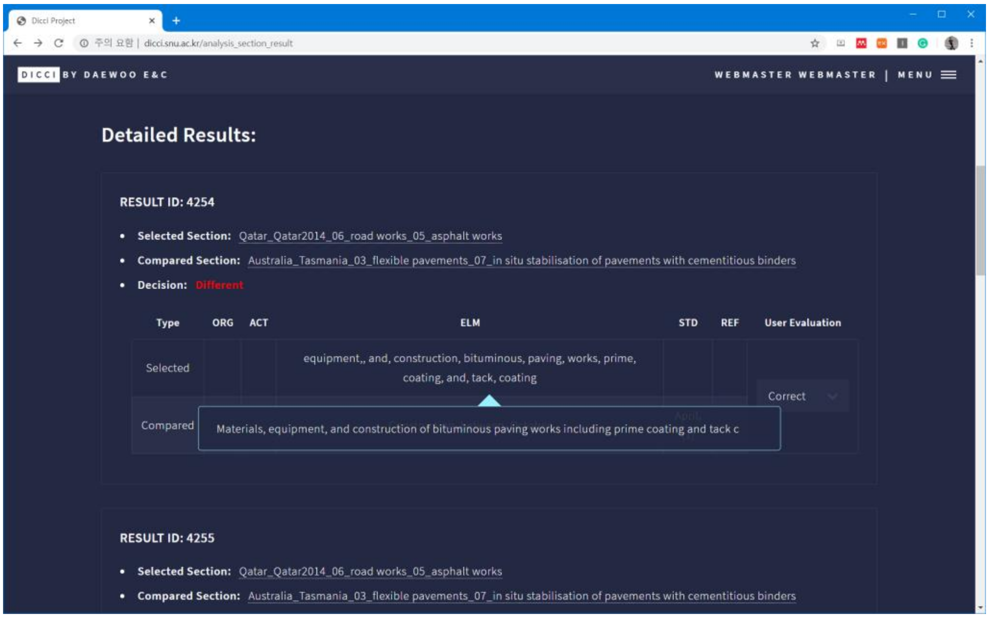
Analysis Report (Section)

Analysis Date: 2020-02-26 01:40:27

Statistics:

Type	Specification	Section	Sentences	Year	Size (KB)
Selected	Qatar - Qatar2014	Qatar_Qatar2014_06_road works_05_asphalt works	878	2014	58484.20
Compared	Australia - Tasmania	Australia_Tasmania_03_flexible pavements_07_in situ stabilisation of pavements with cementitious binders	352	2017	13504.16

Figure A.10 Statistics of selected data


Dicci Project
주의 요함 | dicci.snu.ac.kr/analysis_section_result
DICCI BY DAEWOO E&C
WEBMASTER WEBMASTER | MENU

Detailed Results:

RESULT ID: 4254

- **Selected Section:** Qatar_Qatar2014_06_road works_05_asphalt works
- **Compared Section:** Australia_Tasmania_03_flexible pavements_07_in situ stabilisation of pavements with cementitious binders
- **Decision:** Different

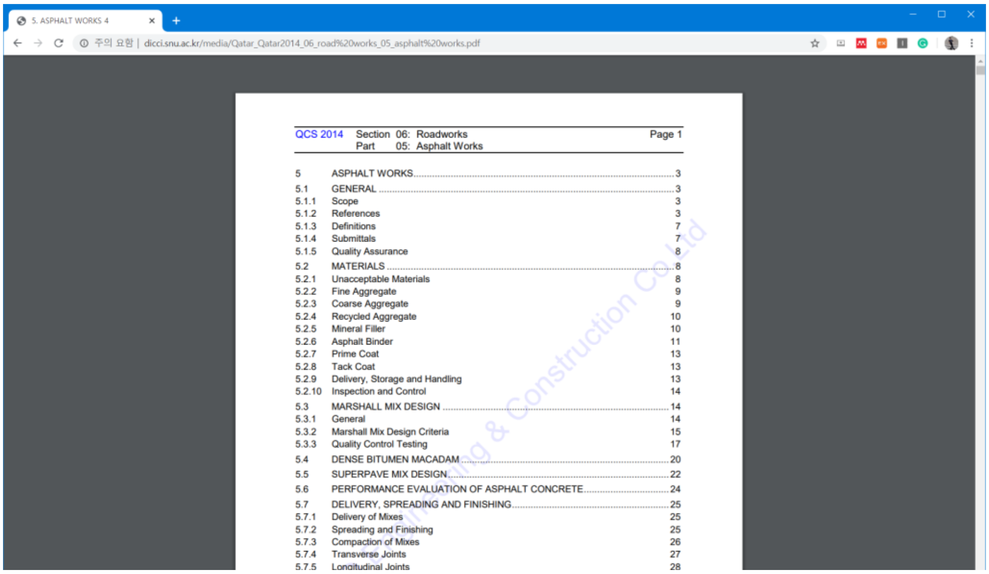
Type	ORG	ACT	ELM	STD	REF	User Evaluation
Selected			equipment,, and, construction, bituminous, paving, works, prime, coating, and, tack, coating			Correct
Compared						

Materials, equipment, and construction of bituminous paving works including prime coating and tack c

RESULT ID: 4255

- **Selected Section:** Qatar_Qatar2014_06_road works_05_asphalt works
- **Compared Section:** Australia_Tasmania_03_flexible pavements_07_in situ stabilisation of pavements with cementitious binders

Figure A.11 Results of informative keywords extraction


5. ASPHALT WORKS 4
주의 요함 | dicci.snu.ac.kr/media/Qatar_Qatar2014_06_road%20works_05_asphalt%20works.pdf

QCS 2014 Section 06: Roadworks Page 1
Part 05: Asphalt Works

5 ASPHALT WORKS....3
5.1 GENERAL....3
5.1.1 Scope 3
5.1.2 References 3
5.1.3 Definitions 7
5.1.4 Submittals 7
5.1.5 Quality Assurance 8
5.2 MATERIALS....8
5.2.1 Unacceptable Materials 8
5.2.2 Fine Aggregate 9
5.2.3 Coarse Aggregate 9
5.2.4 Recycled Aggregate 10
5.2.5 Mineral Filler 10
5.2.6 Asphalt Binder 11
5.2.7 Prime Coat 13
5.2.8 Tack Coat 13
5.2.9 Delivery, Storage and Handling 13
5.2.10 Inspection and Control 14
5.3 MARSHALL MIX DESIGN....14
5.3.1 General 14
5.3.2 Marshall Mix Design Criteria 15
5.3.3 Quality Control Testing 17
5.4 DENSE BITUMEN MACADAM....20
5.5 SUPERPAVE MIX DESIGN....22
5.6 PERFORMANCE EVALUATION OF ASPHALT CONCRETE....24
5.7 DELIVERY, SPREADING AND FINISHING....25
5.7.1 Delivery of Mixes 25
5.7.2 Spreading and Finishing 25
5.7.3 Compaction of Mixes 26
5.7.4 Transverse Joints 27
5.7.5 Longitudinal Joints 28

Figure A.12 Original PDF file of selected specification