Journal of the Korean Society of Civil Engineers
Vol. 38, No. 4: 621-625/ August, 2018
DOI: https://doi.org/10.12652/Ksce.2018.38.4.0621

ISSN 1015-6348 (Print)
ISSN 2287-934X (Online)
www.kscejournal.or.kr

Construction Management **시공관리**

딥러닝 기반 교량 점검보고서의 손상 인자 인식

정세환* · 문성현** · 지석호***
Chung, Sehwan*, Moon, Seonghyeon**, Chi, Seokho***

Bridge Damage Factor Recognition from Inspection Reports Using Deep Learning

ABSTRACT

This paper proposes a method for bridge damage factor recognition from inspection reports using deep learning. Bridge inspection reports contains inspection results including identified damages and causal analysis results. However, collecting such information from inspection reports manually is limited due to their considerable amount. Therefore, this paper proposes a model for recognizing bridge damage factor from inspection reports applying Named Entity Recognition (NER) using deep learning. Named Entity Recognition, Word Embedding, Recurrent Neural Network, one of deep learning methods, were applied to construct the proposed model. Experimental results showed that the proposed model has abilities to 1) recognize damage and damage factor included in a training data, 2) distinguish a specific word as a damage or a damage factor, depending on its context, and 3) recognize new damage words not included in a training data.

Key words : Bridge inspection reports, Damage factor recognition, Word embedding, Recurrent neural network

초 록

본 연구는 딥러닝을 활용하여 교량 점검보고서에서 손상 및 손상 인자를 자동으로 식별하는 방법을 제안한다. 교량 점검보고서에는 점검 결과 발견된 손상 및 원인 분석 결과가 기록되어 있다. 그러나 점검보고서의 양이 방대하여 인력으로 보고서로부터 정보를 수집하는 데 한계가 있다. 따라서 본 연구에서는 딥러닝 기반 개체명 인식 방법을 활용하여 교량 점검보고서 텍스트로부터 손상 및 손상 인자에 해당하는 단어들을 식별할 수 있는 모델을 제안한다. 모델 구현의 주요 방법론으로는 개체명 인식(Named Entity Recognition), 워드 임베딩(Word Embedding), 딥러닝의 일종인 순환신경망(Recurrent Neural Network)을 활용하였다. 실험 결과 제안된 모델은 1)훈련 데이터에 포함된 손상 및 손상 인자 단어들을 잘 식별할 수 있고, 2)단어 주변 맥락에 따라 특정 단어가 손상에 해당하는지 손상 인자에 해당하는지 잘 판별할 수 있을 뿐만 아니라, 3)훈련 데이터에 포함되지 않은 새로운 종류의 손상 단어도 잘 인식할 수 있는 것으로 확인되었다.

검색어 : 교량 점검보고서, 손상 인자 식별, 워드 임베딩, 순환신경망

1. 서 론

교량시설물에 대한 "안전점검 및 정밀안전진단 보고서"(이하 "점검보고서")는 점검을 통해 발견된 교량의 손상 및 해당 손상의 원인 분석 결과를 담고 있다. 교량 점검보고서에 기록된 손상 및 영향 인자 분석 정보는 어떤 교량에 어떤 손상이 어떤 원인에 의해

* 정회원 · 서울대학교 건설환경공학부 석사과정 (Seoul National University · hwani751@snu.ac.kr)
** 정회원 · 서울대학교 건설환경공학부 석박통합과정 (Seoul National University · blank54@snu.ac.kr)
*** 종신회원 · 교신저자·서울대학교 건설환경공학부 교수, 서울대학교 건설환경종합연구소 겸임교수
(**Corresponding Author** · Seoul National University, The Institute of Construction and Environmental Engineering (ICEE) · shchi@snu.ac.kr)

Received April 18, 2018/ **revised** May 3, 2018/ **accepted** May 8, 2018

발생했는지 설명해 주며, 이러한 정보는 교량시설물의 건설 및 유지관리 시 발생할 수 있는 손상을 사전에 예측하고 이를 미리 방지하기 위한 대책을 세우는 데 활용될 수 있다. 따라서 교량 점검보고서에 포함된 손상 및 손상 인자 정보를 수집, 분석, 활용하는 것은 국내외적으로 많은 연구적인 관심을 끌고 있다(Lee et al., 2014; Liu and El-Gohary, 2017; Lokuge et al., 2016; Peris-Sayol et al., 2017; Ryu and Shin, 2014).

그러나 교량 점검보고서의 양이 방대하기 때문에, 인력으로 보고서에서 손상 및 손상 인자 분석 정보를 수집하는 데에는 한계가 있다. 따라서 자연어로 작성된 점검보고서에 기계학습 기반 텍스트 마이닝 기술을 적용함으로써 사용자가 원하는 정보를 자동으로 수집하는 데 도움을 줄 수 있다. 특히, 텍스트에서 인명, 지명, 기관명 등 개체명을 식별하기 위한 개체명 인식(Named Entity Recognition) 방법이 다양한 분야의 텍스트 데이터에 적용되고 있는데, 개체명 인식 기술은 일반적인 개체명을 인식 및 분류하는 외에 특정 전문 분야의 도메인 지식과 관련된 단어들을 자동으로 식별하는 데에도 활용되고 있다(Tanabe et al., 2005; Zhu et al., 2013).

건설분야에서도 산재한 문서 데이터로부터 유용한 정보를 추출하기 위해 텍스트마이닝 기술을 적용한 선행연구가 진행되었으나, 대부분의 선행연구는 단어의 빈도수를 기반으로 자주 등장하는 주요 단어를 도출하고 주요 단어들의 시간적 추세 또는 단어 간 연관관계 분석 등에 초점을 맞추고 있다(Jeong and Kim, 2012; Lee et al., 2016). 이와 같은 단어 빈도수 기반 분석 방법은 문서에 기록된 지식정보의 전체적인 경향성을 도출하는 데는 적합하나, 교량 점검보고서에 기록된 손상 인자와 같은 매우 다양하고 이질적인 종류의 단어들을 자동으로 인식 및 분류하는 작업에는 적합하지 않다.

본 연구는 개체명 인식(Named Entity Recognition) 방법을 적용한 교량 점검보고서의 손상 인자 인식 모델을 제안한다. 본 연구에서 제안하는 모델은 교량 점검보고서의 각 문장을 입력으로 받아 각 문장에 포함된 손상 및 손상 인자에 해당하는 단어를 자동으로 인식하는 것을 목적으로 한다. 손상 인자 인식 모델 구축에는 딥러닝의 일종인 순환신경망(Recurrent Neural Network) 방법을 활용하였으며, 순환신경망 모델이 점검보고서 텍스트를 입력값으로 받을 수 있도록 단어를 벡터로 변환하는 기법 중 하나인 워드 임베딩(Word Embedding)을 점검보고서 텍스트에 적용하였다.

2. 연구 방법론

2.1 개체명 인식(Named Entity Recognition)

개체명 인식은 텍스트에서 인명, 지명, 기관명 등 사용자가 설정한 종류에 속하는 단어들을 자동으로 식별하는 것을 목적으로 하는 방법이다. 구체적으로, 개체명 인식 모델은 한 문장을 입력으로 받고, 그 문장에 포함된 단어 중 개체명에 해당하는 단어들을 식별한 뒤 각 단어가 어떤 개체명에 해당하는 지를 분류한다. 본 연구에서 제안하는 개체명 인식 모델은 교량 점검보고서의 각 문장을 입력으로 받고, 문장에 포함된 각각의 단어가 손상 종류에 해당하는지, 손상 인자에 해당하는지, 또는 둘 중 아무 것에도 해당하지 않는지를 판별한다(Fig. 1).

본 연구에서 제안하는 손상 인자 인식 작업은 기계학습의 관점에서 입력 데이터의 레이블이 '손상' 또는 '원인'에 해당하는지 판별하는 분류분석(Classification) 작업에 속하며, 또한 손상 인자 인식 모델은 입력 및 출력 값이 개별 단어나 레이블이 아니라 단어 또는 레이블의 배열인 Sequence Model에 해당한다. Sequence model의 필요조건은, 첫째, 다양한 길이의 입력값을 모두 처리할 수 있어야 하며, 둘째, 입력값으로 들어온 데이터 배열 내의 순차적 패턴을 학습할 수 있어야 한다.

2.2 워드 임베딩(Word Embedding)

텍스트 데이터를 기계학습 모델에 입력값으로 넣기 위해서는 각 단어를 숫자, 즉 벡터로 변환하는 과정이 필요하다. 일반적으로 명목형 변수를 벡터화할 때 사용하는 One-hot Encoding 방식을 적용하여 텍스트 데이터에 존재하는 각각의 단어를 벡터로 변환한다면, 각 단어 벡터는 그 크기가 어휘의 개수와 같고 그 중 한 개의 원소만 1이고 나머지는 0인 벡터로 표현된다(Fig. 2). One-hot

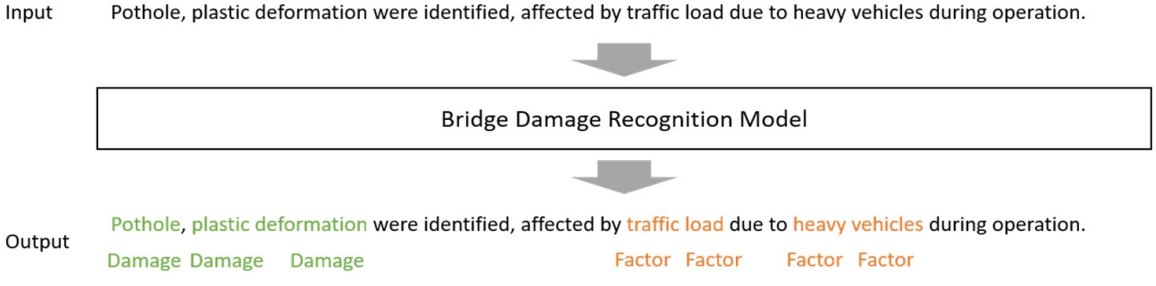

Fig. 1. Damage Factor Recognition Model

Encoding 방식은 표현 방식이 직관적인 반면, 단어와 단어 사이에 존재하는 속성 간 유사성을 표현할 수 없다는 한계점이 있다.

워드 임베딩(Word Embedding)은 One-hot으로 표현된 단어 벡터를 보다 작은 차원의 벡터로 변환(embedding)하는 방법이다. 본 연구에서는 Mikolov et al. (2013)이 제안한 "Word2vec" 알고리즘을 활용하여 워드 임베딩을 수행하였다. Word2vec 알고리즘은 함께 등장하는 단어들은 의미가 비슷할 것이라는 가정 하에, 한 단어를 변환하기 위해 그 단어 주변의 다른 단어들을 이용하는 신경망 모델을 활용한다(Fig. 3). Word2Vec에서 워드 임베딩을 수행하는 과정은 다음과 같다. 우선, 변환하고자 하는 단어의 주변 단어들의 one-hot 벡터를 신경망 모델에 입력한다. 입력된 one-hot 벡터들은 신경망 모델의 은닉층으로 전달되며, 은닉층의 노드 개수는 사용자가 임의로 지정한다. 은닉층의 값은 출력층으로 전달되어 변환하고자 하는 단어의 one-hot 벡터를 예측하도록 가중치가 조절된다. 이 과정에서 은닉층의 값이 최초에 변환하고자 했던 단어의 임베딩 벡터가 된다. 텍스트 데이터에 존재하는 모든 단어들을 대상으로 위의 변환 과정을 적용함으로써 최종적으로 맥락이 비슷한 단어들끼리 비슷한 벡터 값을 갖는다.

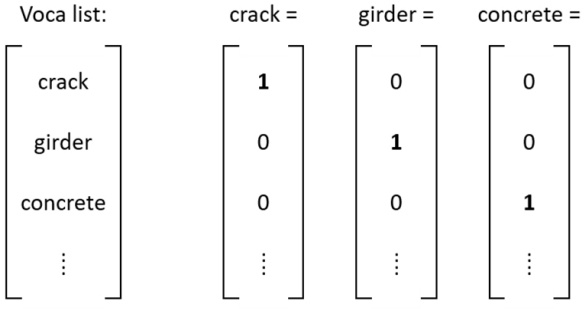

Fig. 2. Examples of One-hot Encoded Word Vectors

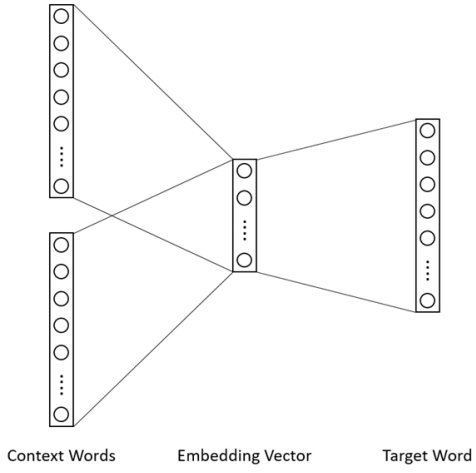

Fig. 3. Neural Network of Word2vec Algorithm for Word Embedding

2.3 순환신경망(Recurrent Neural Network)

본 연구에서는 딥러닝의 일종인 순환신경망(Recurrent Neural Network)을 활용하여 손상 인자 인식 모델을 구현하였다. 순환신경망(Recurrent Neural Network)은 Sequence model에 적용될 수 있는 신경망 구조로, 개별 단어를 입력값으로 받는 일반적인 인공신경망과 달리 단어의 배열 전체를 입력값으로 받고 각각의 단어에 해당하는 레이블의 배열을 출력한다(Fig. 4). 한 단어의 레이블을 예측하는 과정에서, 해당 단어의 임베딩 벡터를 입력값으로 받을 뿐만 아니라 이전 단어 은닉층의 값 또한 이번 단어의 은닉층에 입력값으로 전달한다. 마찬가지로 이번 단어의 은닉층 값은 해당 단어의 레이블을 판별하기 위해 출력층에 전달될 뿐만 아니라 다음 단어의 은닉층에 입력값으로서 전달된다. 이와 같은 구조로 인해 순환신경망 모델은 입력 데이터에 존재하는 순차적 패턴을 학습할 수 있다.

그러나 실제 텍스트 데이터에 존재하는 순차적 패턴은 이와 같이 한 방향으로만 존재하지 않는다. 예를 들어, "중차량 통행에 의한 윤하중의 영향으로 포장마모 등의 손상이 조사되었다." 와 같은 문장에서, "윤하중"이란 단어가 손상인지 영향인자인지 식별하기 위해서는 그 단어 이후에 등장하는 "영향으로", "포장마모" 등과 같은 단어가 중요한 역할을 한다. 따라서 본 연구에서는 텍스트 데이터의 양방향 패턴을 모델이 학습할 수 있도록 양방향 순환신경망(Bidirectional Recurrent Neural Network)을 활용하여 최종적

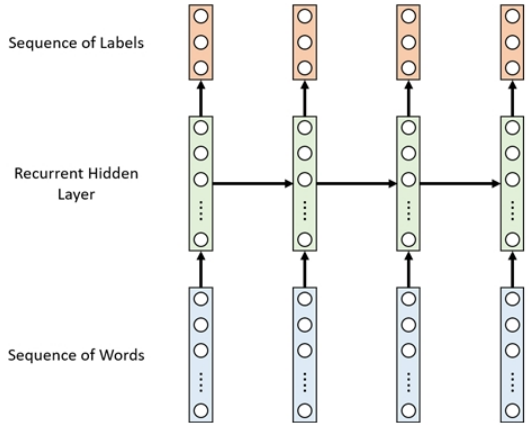

Fig. 4. Recurrent Neural Network

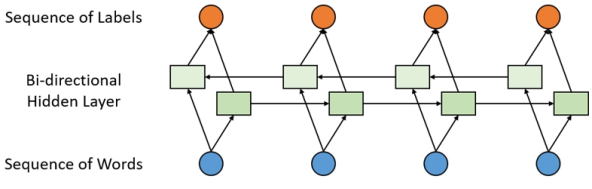

Fig. 5. Bidirectional Recurrent Neural Network

으로 개체명 인식 모델을 구축하였다(Schuster and Paliwal, 1997) (Fig. 5).

3. 연구 결과

본 연구에서는 개체명 인식, 워드 임베딩 및 순환신경망을 활용한 손상 인자 인식 모델을 제안하였다. 모델의 손상 및 손상 인자 인식 성능 검증을 위해 105건의 국도교량 점검보고서에 포함된 총 3,583개의 문장을 활용하였다. 보고서에 존재하는 각각의 단어들에 대해 Word2vec 알고리즘을 적용하여 단어 벡터로 변환하였다. 전체 문장 중 손상 인과관계를 표현할 수 있는 "에 의한" 또는 "의 영향으로"와 같은 문구를 포함하는 10개의 문장을 모델 학습을 위한 훈련 데이터로 선별하였다. 훈련 데이터의 각각의 단어에 대해 손상, 원인, 무속성 중 어디에 해당하는지 레이블을 매겼으며, 이를 바탕으로 손상 인자 인식 모델을 훈련시킨 뒤 레이블을 매기지 않은 나머지 문장에 대한 모델의 예측 결과를 확인하였다 (Table 1).

실험 결과 제안된 모델은 문장에 포함된 손상 및 손상 인자에 해당하는 단어들을 효과적으로 인식하고 분류할 수 있는 것으로 확인되었다. 첫째로, 제안된 모델은 초기 훈련 데이터에 포함된 손상 종류 단어들을 새로운 문장에서도 잘 인식하는 것으로 확인되었다. 예를 들어, 훈련 데이터에 포함된 "부식", "박락", "파손"과 같은 손상 종류 단어들이 포함된 문장을 입력으로 받았을 때, 모델은 해당 단어가 손상 종류에 해당한다는 것을 정확히 인식하였다.

둘째로, 맥락에 따라 특정 단어가 손상으로 분류될 수도, 손상 인자로 분류될 수도 있는 경우를 잘 구별하는 것으로 확인되었다. 예를 들어, "부재 손상은 접속부 누수에 의한 손상으로 검토되었다"와 "균열, 유입된 우수에 의한 누수 흔적이 조사되었다"의 두 문장에서 "누수"라는 단어는 주변 맥락에 따라 손상 인자가 될 수도 있으며 손상 자체일 수 있다. 두 경우 모두 제안된 모델은 "누수"라는 단어가 각각 손상 인자 및 손상에 해당하는 단어라는 것을 잘 인식하는 것으로 확인되었다.

셋째로, 모델은 훈련 데이터에 포함되어 있지 않은 새로운 종류의 손상 단어도 효과적으로 손상으로서 인식하는 것으로 확인되었다.

Table 1. Manually Labeled Training Data (Translated in English)

No.	Sentences	Factor	Damage
1	Deformation in the railing, which had been investigated due to vehicle collision, causes no additional damage, but repair is required to ensure the safety of the vehicle.	vehicle, collision	deformation
2	In the visual inspection on the expansion joint, it was investigated that sedimentation in the expansion joint occurred due to dust of passing vehicle, and crack on the concrete partially occurred.	vehicle, dust	sedimentation, crack
3	In addition, some cracks were investigated due to drying shrinkage and repeated wheel load	repeated, wheel load, shrinkage	crack
4	The deck was investigated that damages such as crack(CW>0.3mm) due to shrinkage, efflorescence, deterioration, and spalling occurred, and requires short-term repair.	shrinkage	crack, efflorescence, deterioration, spalling
5	In the visual inspection, damages such as vertical crack(CW<0.3mm), traces of leakage due to rainwater inflowing, material separation, and breakage were investigated	rainwater	vertical crack, leakage, material separation, breakage
6	In addition, re-bar exposure due to lack of coating was investigated in some sections of the barrier, and repairs for the expected corrosion and spalling on surrounding concrete, caused by exposure to outside air and rainwater, are required	lack of coating, exposure to outside air, rainwater	re-bar, exposure, corrosion, spalling
7	In the visual inspection on the drainage facilities, some clogging of the drainage holes, caused by sedimentation due to environmental factors, was investigated, and one poor installation of the drainage pipe was investigated.	sedimentation	drainage hole, clogging, installation, poor
8	In the visual inspection on the bridge supports, one corrosion of the main body caused by inflow of rainwater and some cracks of supporting concrete caused by vibration due to repetitive vehicle traffic were investigated.	rainwater inflow, vehicle, traffic, vibration	main body, corrosion, crack
9	In the visual inspection, clogging of the drainage holes caused by sedimentation due to environmental factors occurred, grating was missing due to external impact, and missing of the drainage pipe support was identified due to inadequate construction.	sedimentation, external, impact, construction, inadequate	drainage hole, clogging, grating, missing, support, missing
10	Spalling and breakage of the curb are considered to be caused by salt and freeze-thaw during operation, and suface corrosion and deterioration were investigated as well as the spalling and the corrosion.	salt, freeze-thaw	spalling, breakage, surface, corrosion, deterioration

예를 들어, “박리”, “표면변색”, “망상균열”과 같은 단어는 훈련 데이터로 활용된 10개의 문장에 포함되어 있지 않은 단어이지만, 모델은 위와 같은 단어들이 포함된 새로운 문장을 입력으로 받았을 때 해당 단어들이 손상 종류에 해당한다는 것을 잘 인식하는 것으로 확인되었다.

반면, “바닥판 하면 녹오염은 신축이음 하부 부식 및 우수유입에 따른 오염으로 ...”와 같은 문장에서, “녹오염”은 손상 종류로 인식되지 않았으며 “부식”은 영향인자 대신 손상으로 분류되었다. “부식”을 손상으로 오분류한 것은 훈련 데이터의 부족으로 모델이 ‘<원인> 에 따른 <손상>’의 패턴을 학습하지 못했기 때문인 것으로 보이며, “녹오염”을 손상으로 인식하는 데 실패한 것은 “녹오염”과 비슷한 다른 종류의 손상 단어가 초기 훈련 데이터에 포함되어 있지 않았기 때문인 것으로 판단된다. 향후에는 더욱 많은 손상 및 영향인자 단어들과 다양한 언어적 패턴을 포함하는 훈련 데이터를 활용함으로써 모델의 손상 인자 인식 성능을 더욱 향상시킬 수 있을 것으로 기대된다.

4. 결 론

본 연구는 교량 점검보고서에서 자동으로 손상 인자를 인식하기 위한 딥러닝 기반 개체명 인식 모델을 제안하였다. 실험 결과, 양방향 순환신경망 모델은 적은 양의 훈련 데이터로부터 새로운 점검 보고서에 기록된 손상과 손상 인자를 식별할 수 있는 것으로 확인되었다. 특히 제안된 모델은 특정 단어가 서술된 맥락에 따라 손상인지 영향인자인지 구분할 수 있으며, 또한 사전에 입력되지 않은 새로운 종류의 손상 단어도 식별할 수 있는 것으로 나타났다.

본 연구는 방대한 양의 교량 점검보고서로부터 손상 인자를 자동으로 인식하는 방법론을 제안하였다는 데 학술적 기여점이 있을 뿐만 아니라, 점검보고서에 산재된 정보의 활용을 가능하게 함으로써 교량 건설 및 유지관리 시 손상 발생을 사전에 예측하고 이에 대한 대책을 수립할 수 있게 한다는 데 실무적 기여점이 있다. 향후 연구에서는 보다 많은 양의 훈련 및 검증 데이터를 구축하고, 이를 바탕으로 모델의 정량적 인식 성능을 검증 및 개선하여 실무적으로 활용 가능한 인식 성능을 갖춘 손상 인자 인식 모델을 개발할 수 있을 것이다.

감사의 글

본 연구는 국토교통부 국토교통과학기술진흥원 건설교통기술촉진연구사업의 연구비지원에 의해 수행되었습니다(17CTAP-C114956-02).

References

Jeong, C. W. and Kim, J. J. (2012). “Analysis of trend in construction using textmining method.” *Journal of the Korean Digital Architecture Interior Association*, Vol. 12, No. 2, pp. 53-60 (in Korean).

Lee, I. K., Moon, M. K, Park, H. S., Jeon, J. C. and Lee, H. H. (2014). “Statistical analysis of damages in expressway bridges.” *Magazine of the Korea Institute for Structural Maintenance and Inspection*, Vol. 18, No. 2, pp. 2-9 (in Korean).

Lee, J. H., Yi, J. S. and Son, J. (2016). “Unstructured construction data analytics using R programming - focused on overseas construction adjudication cases.” *Journal of the Architectural Institute of Korea Structure & Construction*, Vol. 32, No. 5, pp. 37-44 (in Korean).

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*, Vol. 81, pp. 313-327.

Lokuge, W., Gamage, N. and Setunge, S. (2016). “Fault tree analysis method for deterioration of timber bridges using an Australian case study.” *Built Environment Project and Asset Management*, Vol. 6, No. 3, pp. 332-344.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). “Efficient estimation of word representations in vector space.” Available at: http://arxiv.org/abs/1301.3781.

Peris-Sayol, G., Paya-Zaforteza, I., Balasch-Parisi, S. and Alós-Moya, J. (2017). “Detailed analysis of the causes of bridge fires and their associated damage levels.” *Journal of Performance of Constructed Facilities*, Vol. 31, No. 3, 04016108.

Ryu, J. M. and Shin, E. C. (2014). “Database construction plan of infrastructure safety inspection and in-depth inspection results.” *Journal of Korean Geosynthetics Society*, Vol. 13, No. 4, pp. 133-141 (in Korean).

Schuster, M. and Paliwal, K. K. (1997). “Bidirectional recurrent neural networks.” *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673-2681.

Tanabe, L., Xie, N., Thom, L. H., Matten, W. and Wilbur, W. J. (2005). “GENETAG: A tagged corpus for gene/protein named entity recognition.” *BMC Bioinformatics*, Vol. 6 (Suppl 1), pp. 1-7.

Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W. and Shen, B. (2013). “Biomedical text mining and its applications in cancer research.” *Journal of Biomedical Informatics*, Vol. 46, No. 2, pp. 200-211.