Journal of the Korean Society of Safety
Vol. 36, No. 1, pp. 56-63, February 2021
https://doi.org/10.14346/JKOSOS.2021.36.1.56

ISSN 1738-3803 (Print)
ISSN 2383-9953 (Online)
http://www.kosos.or.kr/jkosos

자연어 처리 기법을 활용한 산업재해 위험요인 구조화

강성식* · 장성록** · 이종빈*** · 서용윤****†

Structuring Risk Factors of Industrial Incidents Using Natural Language Process

Sungsik Kang* · Seong Rok Chang** · Jongbin Lee*** · Yongyoon Suh****†

†Corresponding Author
Yongyoon Suh
Tel : +82-51-629-6467
E-mail : ysuh@pknu.ac.kr

Received : February 3, 2021
Revised : February 17, 2021
Accepted : February 24, 2021

Abstract : The narrative texts of industrial accident reports help to identify accident risk factors. They relate the accident triggers to the sequence of events and the outcomes of an accident. Particularly, a set of related keywords in the context of the narrative can represent how the accident proceeded. Previous studies on text analytics for structuring accident reports have been limited to extracting individual keywords without context. We proposed a context-based analysis using a Natural Language Processing (NLP) algorithm to remedy this shortcoming. This study aims to apply *Word2Vec* of the NLP algorithm to extract adjacent keywords, known as word embedding, conducted by the neural network algorithm based on supervised learning. During processing, *Word2Vec* is conducted by adjacent keywords in narrative texts as inputs to achieve its supervised learning; keyword weights emerge as the vectors representing the degree of neighboring among keywords. Similar keyword weights mean that the keywords are closely arranged within sentences in the narrative text. Consequently, a set of keywords that have similar weights presents similar accidents. We extracted ten accident processes containing related keywords and used them to understand the risk factors determining how an accident proceeds. This information helps identify how a checklist for an accident report should be structured.

Key Words : natural language process, risk factor, accident process, accident report, narrative text, Word2Vec

1. 서 론

산업재해조사표와 같은 재해보고문서는 사업장 정보와 재해자 정보, 재해 발생 정보 등 사고의 유용한 정보를 포함하고 있다. 재해보고문서는 사고 발생 시 사고개요에 대하여 현장의 작업자 및 감독자에 의해 육하원칙에 따라 서술형으로 작성되며, 사고 예방을 위한 기초 자료로 수집 및 활용되고 있다[1]. 재해보고문서는 지속적으로 축적된다는 점에서, 방대한 재해 서술 데이터의 분석가치가 높아지고 있으며, 이를 분석할 수 있는 도구의 필요성 역시 증가하고 있다[2]. 사고개요와 같이 서술형 텍스트로 이루어진 데이터를 체계적으로 분석하기 위하여 자연어 처리기법(natural language processing)이 활용되고 있다. 자연어 처리기법은 비정형 데이터인 텍스트 정보를 분석하기 위하여 문서 내의 단어를 매트릭스 형태로 정형화하여 수학적 분석을 활용할 수 있도록 하는 처리기법이다[3]. 이는 음성 인식과 내용 요약, 번역, 감성 분석, 텍스트 분류작업 등 다양한 분야에서 사용되고 있다.

안전 분야의 경우 역시, 자연어 처리기법 중 텍스트

*부경대학교 안전공학과 박사과정 (Department of Safety Engineering, Pukyong National University)
**부경대학교 안전공학과 교수 (Department of Safety Engineering, Pukyong National University)
***부경대학교 방재연구소 전임연구원 (Laboratory of Disaster Management, Pukyong National University)
****부경대학교 안전공학과 부교수 (Department of Safety Engineering, Pukyong National University)

마이닝을 활용하여 사고개요를 분석한 연구들이 진행되고 있다. 이와 같은 연구들은 문서 내에 도출된 키워드의 빈도를 활용하여 분석하거나 문서에 포함되어있는 키워드의 노출 정도에 따라 가중치를 부여하는 등 키워드의 빈도 위주로 대부분 연구되고 있다. 예를 들면, 고위험 업종의 재해보고문서로부터 사고의 개요를 키워드분석을 통해 위험요인을 도출하여 연관성을 분석하거나 사고의 발생 유형과 관계를 분석하는 등 다양한 연구들이 진행되었다[4-7].

그러나 사고개요의 키워드 빈도분석은 전체 문서에 포함되어있는 키워드를 추출하여 많이 도출된 키워드를 위험요인으로 선정하여 분석하거나, 각 문서의 키워드 빈도를 분석하여 문서 간의 관계를 분석하는데 사용된다. 기존의 텍스트마이닝을 이용한 키워드 분석은 서술형으로 작성되는 사고개요에 대하여 단어 간 연관관계인 문맥을 고려하지 않고 키워드의 빈도만을 분석하기 때문에 인접 키워드의 관계를 고려한 문맥분석에는 어려움이 있다.

이와 같은 문제를 해결하기 위해 자연어 처리 분야에서는 임베딩(embedding) 기법이 활용되고 있다. 임베딩기법을 활용하여 문장에서 단어의 위치에 따라 가중치를 다르게 부여하여 비슷한 의미로 사용되는 단어와 인접하게 나타나는 단어들을 유추하여, 비슷한 문맥에서 유사하게 사용되는 단어들을 도출할 수 있다[8]. 이를 활용하여 문맥에서 나타나는 키워드의 특성을 분석할 수 있으며, 사고개요에서 인접하게 나타나는 위험요인들의 관계를 파악할 수 있다. 임베딩을 활용한 기존연구에서는 문맥을 고려하여 단어의 관계를 파악하여, 효과적으로 검색할 수 있는 도구를 만들거나 사전 내의 단어를 군집화하는 등 문서를 활용하는 분야에서 사용되고 있다[9,10].

본 연구에서는 재해보고문서에 포함되어 있는 위험요인을 키워드 수준으로 도출하고, 문장 구조를 통해 인접단어들을 분석하기 위하여 자연어처리기법 중 지도 학습 모형의 워드 임베딩(word embedding) 기법인 Word2Vec을 활용한다. 먼저, Word2Vec이란 서술형 문장으로 이루어진 사고개요로부터 단어 수준의 키워드를 분석하기 위한 기법으로, 문장에서 인접하게 나타나는 위험요인을 도출한다. 인접한 위험요인들을 도출한 후 키워드의 인접도에 따라 요인들을 이차원의 지도 형태로 시각화하여 키워드 간의 인접 관계를 효과적으로 파악할 수 있다. 즉, 서술형으로 작성된 사고개요의 문맥에서 인접한 단어들을 도출할 수 있다. 결과적으로, 도출된 인접단어를 활용하여 사고개요 작성에 사용되는 용어들을 체크리스트 형태로 제시하고, 대표 사고 프로세스와 그에 포함된 위험요인을 확인하는데 도움이 되리라 기대된다.

2. 배경이론

2.1 재해보고문서 활용 연구

사업장의 안전관리를 위하여 산업재해조사는 필수적으로 요구되고 있다. 사업장에서 발생하는 사고원인을 조사하여 동종업종에서 발생할 수 있는 유사재해에 대하여 예방대책을 수립할 수 있으며, 재해보고문서의 분석을 통하여 다양한 연구가 진행되고 있다. 조재환의 연구에서는 건설업의 재해사례 분석을 통하여 재해유형별 원인 분석과 예방대책을 연구하였다[11]. 이 연구는 산업재해 특성을 분석하기 위하여 「산업재해 기록·분류에 관한 지침」의 분류체계를 활용하여 재해보고문서에서 사업장 특성과 재해자 특성, 재해발생 특성에 따라 재해 유형별 발생 원인을 분석하여 예방대책을 제시하고 있다.

2.2 텍스트마이닝 활용 연구

텍스트마이닝을 활용한 사고 분석은 서술형으로 작성된 사고개요를 분석하기 위하여 다양하게 사용되고 있다. 이영재와 강성경, 유환의 연구는 사고가 가장 많이 발생하는 건설업에 대하여, 텍스트마이닝을 활용하여 재해사례 분석을 통해 전문건설업종별 위험요인을 탐색하였다[12]. 2017년에 발생한 사고사례를 텍스트마이닝을 활용하여 키워드의 빈도수를 도출하였다. 또한 장우현과 서용윤는 화학제조업의 재해보고문서에서 local outlier factor 알고리즘과 의사결정나무 알고리즘을 활용하여 이상사고를 도출하고 사고의 원인을 분석하였다[13]. 이 연구는 빈도가 높은 사고보다 상대적으로 관리가 미흡할 수 있는 이상사고의 주요원인들을 탐색하고 분석하였다. 텍스트마이닝을 활용하여 사고개요에서 키워드의 빈도를 추출하여 이상치 탐색 알고리즘을 활용하였다.

2.3 워드 임베딩 기법

앞서 주요 기존연구들은 사고개요에서 키워드의 빈도를 도출하여 분석하였으며, 빈도를 도출하는 텍스트마이닝 기법은 문서-키워드 매트릭스를 통해 데이터를 분석하며 문서 내에 포함된 키워드를 중심으로 분석한다. 이는 데이터분석에 많은 시간이 걸리므로 많은 양의 데이터를 분석하기에 적절하지 않다. 또한 사고개요는 작성자에 따라 주관적인 단어 선택으로 인해 정확한 키워드 도출이 어렵다는 단점이 있다. 이와 같은

문제를 해결하기 위하여 전체 문서에 존재하는 키워드에 대하여 키워드-벡터 매트릭스 형태로 분석하는 워드 임베딩을 활용한 연구가 진행되고 있다.

워드 임베딩은 단어를 벡터화시키는 학습 과정을 의미한다. 이는 '비슷한 분포를 가진 단어들은 비슷한 의미를 갖는다'라는 가정으로 개발되었다. 비슷한 문맥에서 함께 등장하는 단어들은 유사한 의미와 관계를 가진다는 것을 유추할 수 있다는 것이 이론의 가정이며, 가장 기본이 되는 모델은 신경망을 활용한 Feed-Forward Neural Net Language Model(NNLM) 이다.

또한, 임베딩 기법은 텍스트 데이터가 추가됨에 따라 계속해서 증가할 수 있는 매트릭스의 차원을 단어 집합의 크기로 상정하지 않고, 단어 집합을 축약하는 새로운 차원으로 표현하는 밀집 표현(dense representation)을 사용한다. 밀집표현은 차원의 감소를 통해 더 적은 차원으로 단어의 의미를 표현하며, 빈도 위주의 키워드 표현처럼 데이터가 0으로 나타났던 문제를 실수값으로 표현되도록 변환한다. 밀집표현으로 작성된 매트릭스는 각각의 차원이 모두 단어의 의미를 가지고 있으며, 많은 양의 데이터를 축소하여 분석할 수 있는 장점이 있다[14].

Word2Vec은 2013년 구글에서 발표된 방법론으로, 기존에 사용되던 NNLM 학습 방법에 비해 빠른 시간에 학습을 하고, 더 좋은 성능을 내도록 고안된 방법이다. 이 방법론은 문장의 형태소를 중심으로 고차원 공간에 각 형태소의 좌표값을 부여한다. 좌표값은 단일한 숫자로 나타나며, 비슷한 좌표를 가지는 형태소들은 유사하다고 판단할 수 있다. 또한 Word2Vec은 CBOW (Continuous Bag of Words)와 Skip-gram 두 가지 방식이 있다. CBOW는 주변에 있는 단어들을 가지고, 중간에 있는 단어를 예측한다. 반대로, Skip-gram은 중간에 있는 단어로 주변 단어들을 예측하는 방법이다. 이처럼 Word2Vec은 저차원 벡터공간에 임베딩된 단어벡터 사이의 유사도를 측정하여 인접단어 분석에 사용되는 알고리즘이며, 단어의 의미를 파악하고 사고개요 문맥에서 나타나는 인접단어의 분석에 활용한다[15].

3. 연구방법 및 결과

3.1 연구절차

본 연구는 다음 Fig. 1과 같이 진행되었다. 먼저, 재해보고문서에서 서술형 문장으로 되어있는 사고개요를 수집하고, 텍스트마이닝을 활용하여 전처리과정을 실시하였다. 다음으로, 전처리된 데이터를 문장 속 단어를 벡터화하는 워드 임베딩의 한 방법인 Word2Vec

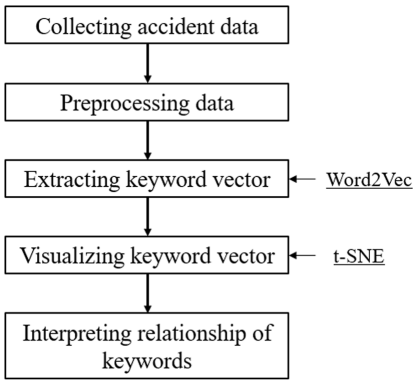

Fig. 1. Flowchart for the methodology.

모델을 사용하여 문맥에서의 인접단어를 분석하였고, 이를 고차원 벡터 모델의 시각화 방법인 t-SNE를 통해 2차원으로 시각화하였다.

3.2 데이터 수집

본 연구의 분석 데이터는 2012년부터 2017년까지 6년간 건설업에서 발생한 사고 142,260건의 사고개요를 수집하여 분석하였다. 수집한 사고개요는 근로자 연령과 사업장 규모를 포함한 기본 정보와 재해 일자와 사고 발생형태 등을 나타내는 사고의 전반적인 내용이 서술형으로 기록되어 있다. 사고 데이터는 엑셀을 활용하여 문서번호와 사고개요 내용으로 정리하였다.

3.3 데이터 전처리

수집된 데이터 전처리는 프로그램 *R*의 *tm* package와 *KoNLP* package를 이용하여 수집된 사고개요에서 명사를 추출하고 문서에 포함된 숫자와 기호, 공백을 제거하였다. 또한 불완전한 단어와 분석에 부적절한 단어는 제거하였다. 전처리 된 전체 데이터는 13,167개의 키워드로 이루어졌으며, 중복된 키워드를 포함하여 7,338,819개의 단어로 구성되어있다.

3.4 유사도 벡터 도출 : Word2Vec(skip-gram)

Word2Vec은 비슷한 맥락에서 나타나는 단어에 대하여 유사한 가중치 벡터를 부여하는 방법론으로 인공신경망을 활용한 기계학습 기반의 워드 임베딩 알고리즘이다. 주변 단어 혹은 중심 단어(target word)를 예측하는 문제에 사용되며, Fig. 2와 같이 sliding window 방식으로 학습한다. Window는 중심 단어 주위에 있는 단어들을 의미하며, 문맥 내에서 중심 단어를 중심으로 window 내에 등장하지 않으면 결과값을 감소시키고, 등장할 경우 결과값을 증가시키며, 계산이 완료되면 다음 인접단어의 결과값을 계산하는 방법으로 키워드

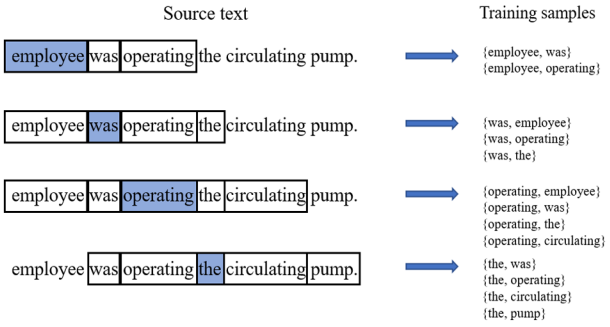

Fig. 2. Sliding window process (window=2).

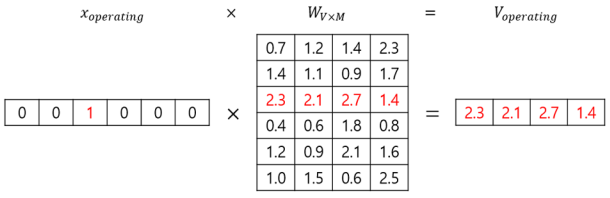

Fig. 3. Example of weight calculation.

의 유사도를 계산하여 벡터로 나타낸다.

Word2Vec의 인공신경망을 도식화하면 Fig. 3과 같다. 먼저, 문맥 내의 각 단어에 대해서, 단어가 존재하는 경우에 1로 입력된 one-hot vector를 작성한다. 다음으로 입력층(input layer)은 각각의 중심단어를 입력하며, 출력층(output layer)은 예측하고자 하는 단어인 분석자가 정한 window의 크기에 따라 주변단어들의 one-hot vector를 찾는다. 단, Fig. 3의 예시는 영어의 사례로 was나 the는 제거하고 벡터화할 필요가 있으며, 한글을 대상으로는 조사를 제외한 용어로 벡터화해야 한다. 은닉층(hidden layer)은 입력층과 가중치에 따라서 작성되며, 분석자가 정한 크기(M)에 따라 임베딩하여 벡터의 차원이 결정된다.

Fig. 4와 같이 Word2Vec은 입력층과 출력층을 학습하여 가중치(W)를 훈련하는 기계학습이며, 단어 집합의 크기(V)와 분석자가 정한 벡터의 크기로 가중치 벡터를 구성한다. 또한 훈련 전의 가중치 벡터는 무작위 값을 갖은 상태로 중심단어에서 주변 단어를 정확히 예측하기 위하여 학습한다. Word2Vec의 학습 과정은 출력된 벡터를 0과 1사이의 값으로 정규화하는 소프트맥스(softmax) 함수를 활용하여, 각 원소의 총합이 1인 스코어 벡터(score vector)를 작성한다. 스코어 벡터는 각 단어가 주변 단어일 확률을 나타내며, 벡터값의 오차를 줄이기 위하여 손실 함수(loss function)로 cross-entropy 함수를 사용한다. cross-entorpy 함수는 머신러닝을 통한 예측모형에서 훈련 데이터를 통해 실제값과 예측값의 차이를 계산하는데 사용된다. 스코어 벡터가 출력 벡터를 정확하게 예측한 경우, cross-entropy 값은

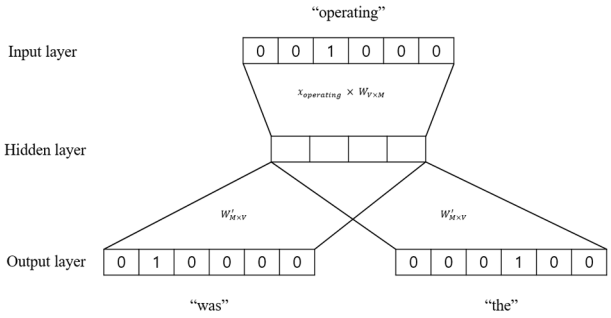

Fig. 4. Neural network diagram in Word2Vec.

0이 되며, 이 값을 최소화하는 방향으로 학습한다. 마지막으로 역전파(back propagation)를 수행하여 가중치 벡터를 임베딩 벡터로 사용한다.

본 연구에서는 인접성을 고려한 키워드 가중치 벡터의 추출은 프로그램 *R*의 *wordVectors* package를 사용하였으며, 분석의 설정값은 200개의 벡터와 window는 4, 최소 키워드 도출 빈도는 10으로 설정하였다. 이와 같은 설정값을 도출하기 위하여 관련 연구를 참고하여[16], 벡터의 크기는 50개와 100개, 200개에 대하여 분석하였고 window는 3과 4, 5로 9개의 조합에 대하여 사전분석을 진행하였다. 또한 키워드 도출 빈도는 전체 사고개요에서 추출한 키워드의 빈도를 의미하며, 9개 이하로 도출되는 단어들은 무의미한 단어라고 판단하고 분석하였다. Word2Vec의 키워드 분석을 통해 최종적으로 도출된 유사도 벡터를 다음 Fig. 5와 같이 12396×200의 데이터 매트릭스로 형태도 추출하였다.

3.5 데이터 시각화 : t-Stochastic Neighbor Embedding (t-SNE) 알고리즘

Stochastic Neighbor Embedding은 고차원의 벡터로 표현되는 데이터 사이의 거리를 최대한 보존하여 저차원의 확률적인 위치를 학습하는 방법론이며 Stochastic modeling은 시점마다 확률 분포를 따라 어떤 점이 다른 점으로 이동하는 모델을 의미한다. t-SNE는 현재 위치에서 다음 시점에 다른 위치로 이동할 확률을 정의하여 가까울수록 높은 확률로 점들을 이동하는 모델이다. 또한 Word2Vec으로 도출된 키워드 벡터를 시각화

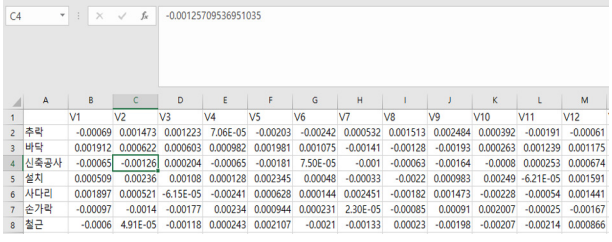

Fig. 5. Weight matrix using Word2Vec.

하기 위하여 주로 사용되는 방법론으로 데이터 간의 거리를 stochastic probability로 변환하여 임베딩에 이용하며, 차원의 축소과정에서 사라질 수 있는 데이터를 보존하여 지도의 형태로 시각화할 수 있다. 데이터 간의 거리는 perplexity를 통해 정의하며, 이는 언어 모델을 평가하기 위한 내부 평가 지표이다. 분석에 적절한 perplexity를 찾기 위하여 binary search를 활용하여 계산한다. 이를 통해 지나치게 크거나 작은 perplexity가 아니면 비슷한 임베딩을 도출할 수 있다[17].

본 연구에서는 Word2Vec으로 도출된 키워드 벡터를 시각화하기 위하여, *R*의 tsne package를 사용하여 t-SNE 분석을 진행하였다. 이를 통해 x와 y좌표를 갖는 2차원의 지도상에 키워드의 좌표를 도출하였다. 키워드 벡터의 perplexity는 binary search를 통해 50으로 설정하였다.

3.6 데이터 군집화 : k-means clustering

k-means clustering은 대표적인 비지도 학습 기법의 군집화 알고리즘이며, 각 데이터로부터 그 데이터가 속한 클러스터의 중심까지의 평균 거리를 최소화하는 방법론이다. 각 군집은 하나의 중심을 갖고, 각 개체는 가장 가까운 중심에 할당되어 같은 중심에 할당된 개체들이 모여 하나의 군집을 형성한다. 형성되는 군집의 수는 hyperparameter인 k를 분석자가 선정하여 군집수(K)를 정해 알고리즘을 실행한다. 군집 수는 데이터의 응집도로 선정하며, 각 데이터로부터 자신이 속한 군집의 중심까지의 거리를 의미하는 inertia 값으로 확인하며, 이 값이 낮을수록 군집화가 더 잘되었다고 볼 수 있다. k-means clustering은 먼저, k개의 임의의 중심점을 배치하고 각 데이터들을 가장 가까운 중심점으로 할당한다. 다음으로, 군집으로 지정된 데이터들을 기반으로 해당 군집의 중심점을 결과값이 수렴할 때까지 업데이트하여 군집화한다[18].

본 연구에서는 키워드 벡터의 인접성을 검토하기 위하여, 시각화한 키워드를 k-means clustering으로 군집화였다. 프로그램 *R*의 *caret* package를 사용하여 키워드를 구분하였으며, 전체 키워드에 대하여 Word2Vec에서 도출된 200개의 vector로 분석을 진행하였다. 클러스터는 Fig. 6과 같이 각 클러스터마다 비슷한 키워드의 수를 갖는 10개의 클러스터로 설정하여 구분하였으며, 각각의 클러스터와 클러스터에 해당되는 키워드는 Table 1과 같다. Table 1의 *표시는 영어단어로 변환과정 중 동의어에 대하여 표시하였고, 한글단어의 경우 다른 단어가 도출되었다. 자세한 한글-영어 변환은 부록에 표기한 표를 확인하기 바란다.

4. 토의 및 시사점

Word2Vec으로 도출된 키워드 벡터를 활용하기 위하여 클러스터에 따라 나누어진 키워드를 「산업재해 기록·분류에 관한 지침」의 분류체계를 활용하여 사고분석의 주요 요인들에 따라 분류하였다. 분류 카테고

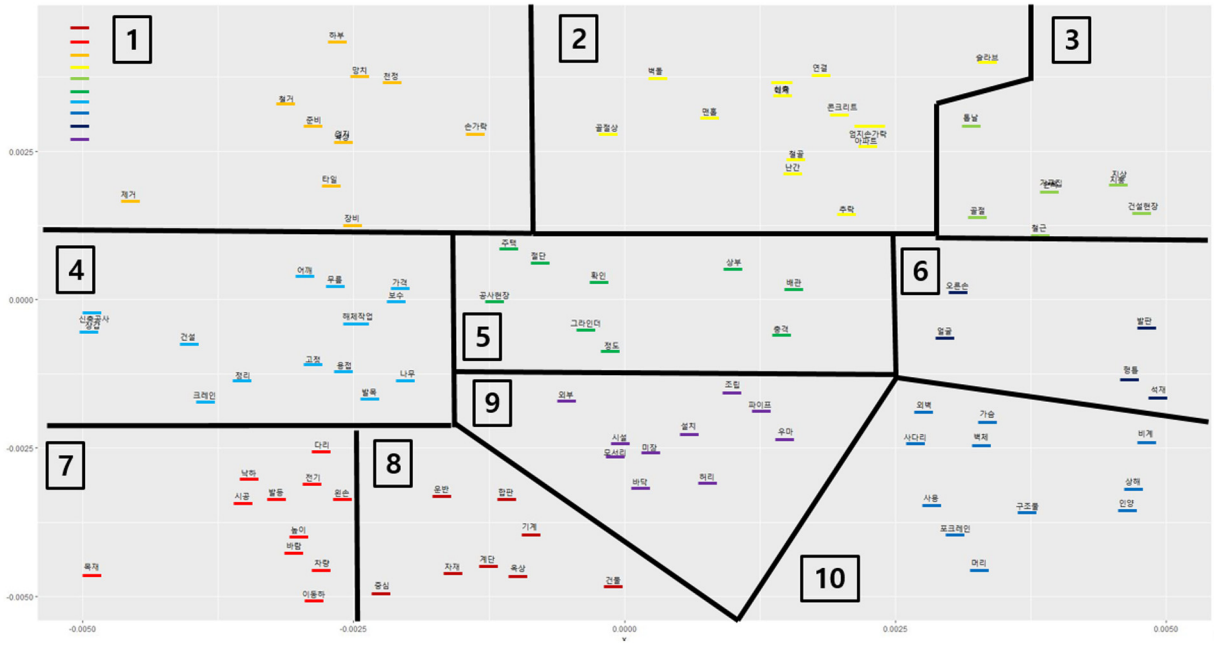

Fig. 6. Visualization of keyword vectors using t-SNE algorithm.

Table 1. Keywords driven by k-means clustering

Cluster number	Number of keywords	Keywords
1	11	Finger, Demolition, Ceiling, Falling, Remove, Bottom, Thumb*, Hammer, Equipment, Tile, Preparation
2	13	Fall, Dismantling, Concrete, Apartment, Connect, Brick, Thumb*, Iron, Fracture*, Slabs, Handrail, Manhole
3	8	Rebar, Formwork, Fracture*, Roof, Wrist, Saw, Ground, Construction site
4	14	New Construction*, Knee, Arrangement, Shoulder, Tree, Fixing, Crane, Dismantling, Welding, Repair, Erection, Price, Glove
5	9	Cut, Site, Housing, Grindstone, Top, Pipe*, Shock, Check, Degree
6	5	Scaffolding, Right hand, Mold, Face, Stone
7	11	Height, Leg, Left hand, Vehicle, Electricity, Move, Construction*, Wood, Fall*, Instep of a foot, Wind
8	8	Material, Stair, Plywood, Center, Carrying, Building, Machine, Rooftop
9	10	Floor, Install, Pipe*, Out, Waist, Facility, Assembly, Plasterer, Scaffold, Edge
10	11	Ladder, Scaffold, Head, Wall, Use, Salvage, Chest, Outer, Wound, Structure, Poclain

Table 2. Accident process and keywords by category

Cluster Number	Task type	Place	Original cause material	Injured area	Type of injury	Type of accident	Accident process
1	Demolition, Remove, Preparation	Ceiling	Hammer, Tile, Equipment	Finger, Thumb*	Fall*		Finger injuries and falls due to hammer during demolition work from the ceiling
2	Dismantling, Connect	Apartment, Slabs, Handrail	Concrete, Brick, Iron, Manhole	Thumb*	Fracture*	Fall*	Fall due to a fracture of the thumb while dismantling the steel frame from the railing
3		Roof, Construction site	Rebar, Formwork, Saw	Wrist	Fracture*		Fracture due to rebar in construction site
4	Arrangement, Fixing, Dismantling, Welding, Repair	New Construction	Tree, Crane, Glove	Knee, Shoulder			Shoulder injury during crane dismantling in new construction
5	Check, Cut	Site, Housing	Grindstone, Pipe*			Shock	Received a shock while cutting the grinder at the residential construction site
6			Scaffolding, Mold, Stone	Right hand, Face			Right hand injury due to footrest and mold
7	Move		Vehicle, Electricity, Wood, Wind	Leg, Left hand, Instep of a foot		Fall*	Leg injuries caused by falling while moving the vehicle
8	Carrying	Stair, Rooftop	Material, Plywood, Machine				Injury while transporting material on stairs
9	Install, Assembly, Plasterer	Floor	Pipe*, Scaffold*	Waist			Waist injury during pipe installation and assembly
10	Salvage	Outer	Ladder, Sacffold, Wall, Poclain	Head, Chest			Head injury while lifting scaffold to outer wall

리는 작업 종류와 사고 장소, 기인물, 사고 부위, 사고 종류, 발생형태로 카테고리에 포함하기 힘든 키워드를 제외하고 82개의 키워드에 대해 Table 2와 같이 정리하였다.

먼저, 각 카테고리에 대하여 도출된 키워드들로 문맥을 형성하여 사고 프로세스를 작성하였다. 대표적으로 클러스터 2는 모든 카테고리에 키워드가 고르게 분포되었으며, 사고 프로세스는 “난간에서 철골 해체작업 중 엄지손가락 골절상으로 인해 추락”으로 작성할 수 있다. 이와 같은 분석을 통하여 각각의 키워드들과 관련된 사고를 예상할 수 있으며, 특히 특정 작업 중 주의해야 될 장소나 기인물을 확인하여 사고를 예방하거나, 부상 부위에 따라 작업 시 착용할 보호구를 파악할 수 있다.

다음으로, 다른 클러스터에 포함된 유사한 키워드를 살펴보면, 클러스터 1의 손가락, 엄지와 클러스터 2의 엄지손가락, 클러스터 2의 골절상과 클러스터 3의 골절은 정확하게 같은 단어로 전처리되지 않아 다른 클러스터에 도출되었다. 또한 클러스터 9의 우마의 경우 높이 1 m 내외의 말비계를 의미하며, 클러스터 10의 비계에 포함될 수 있다. 이와 같은 키워드는 동일한 클러스터로 분류되지 않았지만, 시각화 지도에서 살펴본

키워드의 위치는 인접한 것으로 나타났다. 현장에서 동의어로 사용되는 단어나 텍스트마이닝 전처리과정에서 정확하게 분류되지 않는 단어를 활용하여 건설업의 사고 키워드 사전을 작성할 수 있다.

이와 같은 결과는 기존의 키워드 빈도 위주의 텍스트마이닝으로 분석한 연구와 상이함을 알 수 있다. 기존의 분석방법의 경우, 문서 내에 존재하는 키워드의 빈도에 따라 분석하여 많이 도출된 단어를 중심으로 키워드를 분석하여 사고 프로세스를 도출하거나 현장에서 사용되는 단어에 따라 정확한 사고분석이 쉽지 않다. 이와 같은 어려움을 Word2Vec을 통해 문맥에서 나타나는 인접단어를 활용하여 키워드가 갖고 있는 의미를 파악할 수 있고, 문맥상 비슷한 위치에 도출되는 keyword set을 활용하여 사고개요를 더욱 정확하게 분석할 수 있다. 또한 산업재해통계에서 나타나는 사고에 영향을 미치는 재해발생형태나 기인물 등에 대한 인접단어의 분석이 가능하다. 건설업의 경우, 재해발생형태는 클러스터 2에 해당되는 떨어짐으로 인한 사고가 가장 많이 발생하였으며, 기인물의 경우는 클러스터 4에 해당되는 크레인과 클러스터 8에 해당되는 건설용 기계에서 사고가 많이 발생하였다. 이와 같이 위험요인에 대한 인접단어의 분석을 활용하여 안전관리자가 작성하는 사고개요를 단어 수준에서 관리할 수 있을 것으로 사료된다.

5. 결 론

본 연구는 워드 임베딩 방법을 활용하여 서술형으로 작성된 사고문서를 분석할 수 있는 기법을 제시하였으며, 문장 구조에서 인접단어의 기계학습을 통해 키워드를 분석하였다. Word2Vec을 활용하여 건설업에서 발생하는 사고들의 주요 요인을 분석하고 문맥으로 구성되는 사고 과정을 제시하였다. 그 결과 인접단어의 기계학습을 통해 작성자에 따라 다르게 사용될 수 있는 사고개요의 특성에 맞게 작업 종류와 사고 장소, 기인물, 사고 부위, 사고종류, 발생형태와 같이 사고를 구성하는 키워드들을 도출하였다. 이를 통해 현장에서 발생할 수 있는 사고에 대해 위험점을 안전관리자가 쉽게 파악할 수 있으며, 같은 의미로 다양하게 사용되고 있는 단어들에 대한 keyword set을 작성할 수 있을 것이라 생각된다.

그러나 본 연구는 향후 개선되어야 할 필요가 있다. 먼저, 2012년부터 2017년까지 건설업에서 발생한 사고개요로 분석하였다. 건설업을 포함하여 다른 업종의 사고개요를 수집하여 사고 과정을 분석할 필요가 있다. 이를 통해 다양한 업종에서 발생하는 사고에 대한 안전관리에 기여할 수 있다. 다음으로, 시각화한 키워드의 개수를 100개로 한정하여 분석하였다. 더 많은 키워드를 시각화하면 시각화 지도에서 각각의 키워드를 파악하기 어려운 문제가 있었으며, 전체 키워드에 대한 인접단어를 탐색할 수 있는 방법이 요구된다. 이는 사고개요의 텍스트마이닝의 전처리과정에 활용하여 더욱 정확한 사고분석을 진행할 수 있으며, 추후 현장에서의 안전관리에 활용될 것이라고 사료된다.

Acknowledgement: This work was supported by a Research Grant of Pukyong National University(2019)

References

1) KOSHA, “Statistical Survey and Analysis of Industrial Disasters”, 2018.
2) Y. Suh, “Data Analytics for Social Risk Forecasting and Assessment of New Technology”, J. Korean Soc. Saf., Vol. 32, No. 3, pp. 83-89, 2017.
3) C. D. Manning and H. Schutze, “Foundations of Statistical Natural Language Processing”, MIT Press, 1999.
4) B. Kim, S. Chang and Y. Suh, “Text Analytics for Classifying Types of Accident Occurrence Using Accident Report Documents”, J. Korean Soc. Saf., Vol. 33, No.3, pp. 58-64, 2018.
5) S. Kang and Y. Suh, “On the Development of Risk Factor Map for Accident Analysis using Textmining and Self-Organizing Map(SOM) Algorithms”, J. Korean Soc. Saf., Vol. 33, No. 6, pp. 77-84, 2018.
6) G. Ahn, M. Seo and S. Hur, “Development of Accident Classification Model and Ontology for Effective Industrial Accident Analysis based on Textmining”, J. Korean Soc. Saf., Vol. 32, No. 5, pp. 179-185, 2017.
7) T. L. Bunn, S. Slavova and L. Hall, “Narrative Text Analysis of Kentucky Tractor Fatality Reports”, Accid. Anal. Prev., Vol. 40, No. 2, pp. 419-425, 2008.
8) T. Mikolov, K. Chen, G. Corrado and J. Dean, “Efficient Estimation of Word Representations in Vector Space”, arXiv preprint, arXiv:1301.3781, 2013.
9) X. He, D. Cai, S. Yan and H. Zhang, “Neighborhood Preserving Embedding”, Tenth IEEE International Conference on Computer Vision, 2005.
10) K. Toutanova, D. Chen, P. Pantel, H. Poon, P. Choudhury and M. Gamon, “Representing Text for Joint Embedding of Text and Knowledge Bases”, Conference on Empirical

Methods in Natural Language Processing, pp. 1499-1509, 2015.
11) J. H. Jo, “A study on the Causes Analysis and Preventive Measures by Disaster types in Construction Fields”, KSMS, Vol. 14, No. 1, pp. 7-13, 2012.
12) S. K. Kang, H. Yu and Y. J. Lee, “Analyzing Disaster Response Terminologies by Text Mining and Social Network Analysis”, Information Systems Review, Vaol. 18, No. 1, pp. 141-155, 2016.
13) W. Jang and Y. Suh, “Identifying Abnormal Accidents Using Local Outlier Factor and Decision Tree Algorithms”, Journal of the Korean Institute of Industrial Engineers, Vol. 45, No. 4, pp. 329-340, 2019.
14) Y. Goldberg and O. Levy, “Word2vec Explained: Deriving Mikolov et al.‘s Negative-sampling Word-embedding Method”, arXiv preprint, arXiv:1402.3722, 2014.
15) L. Ma and Y. Zhang, “Using Word2Vec to Process Big Text Data”, IEEE International Conference on Big Data, 2015.
16) Sanghyuk Choi, Jinseok Seol and Sang-goo Lee, “On Word Embedding Models and Parameters Optimized for Korean”, Korean Language information Science Society, pp. 252-256, 2016.
17) L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE”, J Mach Learn Res, Vol. 9, pp. 2579-2605, 2008.
18) A. Likas, N. Vlassis and J. J. Verbeek, “The Global k-means Clustering Algorithm”, Pattern Recognition, Vol. 36, No. 2, pp. 451-461, 2003.

부록 : 영한 사전(알파벳순)

영어	한글	영어	한글
Apartment	아파트	Left hand	왼손
Arrangement	정리	Leg	다리
Assembly	조립	Machine	기계
Bottom	하부	Manhole	맨홀
Brick	벽돌	Material	자재
Building	건물	Mold	형틀
Carrying	운반	Move	이동
Ceiling	천정	New Construction	신축공사
Center	중심	Out	외부
Check	확인	Outer	외벽
Chest	가슴	Pipe	배관
Concrete	콘크리트	Pipe	파이프
Connect	연결	Plasterer	미장
Construction	시공	Plywood	합판
Construction site	건설현장	Poclain	포크레인
Crane	크레인	Preparation	준비
Cut	절단	Price	가격
Degree	정도	Rebar	철근
Demolition	철거	Remove	제거
Dismantling	해체	Repair	보수
Dismantling	해체작업	Right hand	오른손
Edge	모서리	Roof	지붕
Electricity	전기	Rooftop	옥상
Equipment	장비	Salvage	인양
Erection	건설	Saw	톱날
Face	얼굴	Scaffold*	비계
Facility	시설	Scaffold*	우마
Fall*	추락	Scaffolding	발판
Fall*	낙하	Shock	충격
Falling	낙상	Shoulder	어깨
Finger	손가락	Site	공사현장
Fixing	고정	Slabs	슬라브
Floor	바닥	Stair	계단
Formwork	거푸집	Stone	석재
Fracture	골절상	Structure	구조물
Fracture	골절	Thumb*	엄지
Glove	장갑	Thumb*	엄지손가락
Grindstone	그라인더	Tile	타일
Ground	지상	Top	상부
Hammer	망치	Tree	나무
Handrail	난간	Use	사용
Head	머리	Vehicle	차량
Height	높이	Waist	허리
Housing	주택	Wall	벽체
Install	설치	Welding	용접
Instep of a foot	발등	Wind	바람
Iron	철골	Wood	나무
Knee	무릎	Wound	상해
Ladder	사다리	Wrist	손목