Journal of the Korean Society of Safety
Vol. 38, No. 5, pp. 51-57, October 2023
https://doi.org/10.14346/JKOSOS.2023.38.5.51

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

해양사고 예방을 위한 사전학습 언어모델의 순차적 레이블링 기반 복수 인과관계 추출

문기영[1] · 김도현[2] · 양태훈[3] · 이상덕[2†]

Sequence Labeling-based Multiple Causal Relations Extraction using Pre-trained Language Model for Maritime Accident Prevention

Ki-Yeong Moon[1] · Do-Hyun Kim[2] · Tae-Hoon Yang[3] · Sang-Duck Lee[2†]

†Corresponding Author
Sang-Duck Lee
Tel : +82-31-460-5764
E-mail : sdlee@krri.re.kr

Received : September 4, 2023
Revised : September 24, 2023
Accepted : October 4, 2023

Abstract : Numerous studies have been conducted to analyze the causal relationships of maritime accidents using natural language processing techniques. However, when multiple causes and effects are associated with a single accident, the effectiveness of extracting these causal relations diminishes. To address this challenge, we compiled a dataset using verdicts from maritime accident cases in this study, analyzed their causal relations, and applied labeling considering the association information of various causes and effects. In addition, to validate the efficacy of our proposed methodology, we fine-tuned the KoELECTRA Korean language model. The results of our validation process demonstrated the ability of our approach to successfully extract multiple causal relationships from maritime accident cases.

Key Words : natural language processing, maritime accident prevention, causal relation extraction, KoELECTRA, sequence labeling

1. 서 론

최근 5년간 발생한 해양사고는 총 14,381건이며, 이로 인한 사망·실종자는 545명, 사고 선박은 15,997척으로 심각한 인적, 물적 피해가 발생하고 있다[1]. 또한, 해양사고 예방을 위한 많은 연구와 정책 수행에도 불구하고, 해양사고는 증가하는 추세를 보이고 있다[2]. 따라서 해양사고를 예방할 수 있는 기술 개발의 중요성이 증대되고 있으며, 해양사고를 예방하기 위한 다양한 기술들의 개발이 수행되었다[3-6]. 대표적으로, 방대한 양의 해양사고 사례를 분석함으로써 사고의 주요 원인과 결과를 파악할 수 있는 기술들이 개발되었다[7-12].

데이터마이닝(data mining)은 대표적인 사고사례 분석 방법으로, 데이터마이닝을 활용하면 대규모 데이터를 처리하여 사고와 관련된 주요 키워드를 신속하게 추출할 수 있다[13,14]. 따라서, 다양한 데이터마이닝 기반의 해양사고 분석 연구들이 수행되었다[7,8]. 그러나 데이터마이닝은 TF-IDF(Term Frequency-Inverse Document Frequency)와 같은 문서 내 키워드의 동시 등장 빈도를 중심으로 키워드를 추출하기 때문에 등장 빈도수는 높지만, 사고와 무관한 키워드가 추출될 수 있으며, 키워드간 인과관계를 포함하는 문맥 정보를 고려할 수 없다.

데이터마이닝 기반 사고 분석의 한계를 극복하기 위하여 자연어처리의 사전학습 언어모델과 전이학습을 활용한 연구들이 수행되었다[9-11]. 이러한 사전학습 언어모델은 방대한 텍스트 데이터를 학습하여 텍스트 내의 문맥 정보를 포착하고 이를 활용할 수 있다. 그러나 위 연구들은 미리 지정한 관심 키워드에 초점을 맞추어 인과관계를 확인한다. 즉, 미리 지정되지 않은 키워드가 등장할 경우, 인과관계 추출 성능이 저하될 수 있다.

이를 극복하기 위해 사전학습 언어모델을 미세조정(fine-tuning)하여 사고의 인과관계에 해당하는 키워드를

[1]한국철도기술연구원 첨단물류시스템연구실 연구원 (Logistics System Research Team of Korea Railroad Research Institute)
[2]한국철도기술연구원 첨단물류시스템연구실 선임연구원 (Logistics System Research Team of Korea Railroad Research Institute)
[3]인하대학교 데이터사이언스학과 학부생 (Department of Data Science Engineering, INHA University)

자동으로 추출하는 연구가 수행되었다[12]. 위 연구는 자연어처리 모델인 KoELECTRA와 레이블링(labeling) 기법을 이용하여 안전사고 사례에서 사고의 인과관계를 자동으로 추출하였다. 하지만 위 연구에서 적용한 레이블링(labeling) 방법은 'C:원인', 'CE:원인과 결과를 잇는 사건', 'E:결과'로, 사고의 최초 원인부터 최종 결과까지가 순차적으로 연결되는 경우 적합하나, 사고사례 내에서 복수의 인과관계를 추출하기에 적합하지 않다.

이런 한계점들을 극복하기 위해, 순차적 레이블링 기반 복수 인과관계 추출 방법을 제안한다. 이를 위해 898건의 한국어 해양사고 비정형 데이터를 수집하고, 해양사고의 인과관계를 분석하여 복수의 원인과 결과의 연결정보를 고려하여 레이블링을 수행하였다. 또한, 복수 인과관계 추출 방법의 성능 검증을 위해 한국어 언어모델인 KoELECTRA를 활용하여 미세조정 하였다.

본 연구에서 제안하는 순차적 레이블링 기반 복수 인과관계 추출 방법의 주요 특징은 다음과 같다. 첫째, 문장 내 복수의 인과관계가 등장해도 식별할 수 있는 레이블링 방법을 사용하므로 문장에서 복수의 인과관계를 추출할 수 있다. 둘째, 복수의 원인 및 결과의 순서와 연결정보를 고려하여 레이블을 부여하므로, 예측된 레이블을 통해 인과관계를 구조화할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 순차적 레이블링 기법과 사전학습 언어모델인 KoELECTRA 모델에 대한 이론적 배경을 설명한다. 3장에서는 학습 데이터, 순차적 레이블링 기반 복수 인과관계 추출 방법을 제시하고 4장에서는 실험을 통해 제안하는 방법론의 성능을 검증한다. 마지막으로, 5장에서 결론을 도출한다.

2. 관련 연구

본 연구에서는 순차적 레이블링 방법을 이용하여 한국어 언어모델인 KoELECTRA를 미세조정하고 해양사고의 복수 인과관계를 추출한다. 순차적 레이블링과 KoELECTRA에 대한 자세한 설명 및 선정 근거는 아래와 같다.

순차적 레이블링은 자연어처리 분야에서 주요한 레이블링 방법으로, 문장 내 각 토큰(token)에 특정 레이블을 할당하는 방법이다. 이 방법은 개체명 인식(Named Entity Recognition), 품사 태깅(Part-of-Speech tagging) 등 토큰 분류 작업에서 주로 사용된다[15]. 사전학습 언어모델은 Transformer[16] 기반 아키텍처를 활용하여 순차적 레이블링에 대한 미세조정을 수행한다. 이 과정에서 사전학습 언어모델은 주변 토큰의 임베딩(embedding) 정보를 반영하여 현재 토큰의 레이블을 학습한다. 이러한 원리로 인해 문장 내 토큰의 관계나 의존성을 고려하여 더 정확하고 일관된 토큰 분류가 가능하다. 그러나 대부분의 기존 연구는 단일 원인-결과 관계 추출에 초점을 맞추고 있다. 해양사고는 복수의 원인이 결합 되어 발생하는 경우가 많으므로, 복수의 인과관계 추출이 필요하며, 이를 위해서는 새로운 접근방식이 필요하다. 따라서 본 연구에서는 순차적 레이블링을 기반으로 복수의 인과관계를 추출할 수 있는 새로운 방법을 제안한다.

ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)[17]는 기존의 Transformer 기반 모델들[18,19]과는 다르게 RTD(Replaced Token Detection)를 통해 학습된다. RTD는 ELECTRA의 모델 구조인 생성자(generator)와 로 토큰을 마스크(mask)하고 판별자(discriminator) 마스크된 토큰을 판별하는 방법이며, 기존 MLM(Masked Language Model) 방식보다 더 효율적인 학습을 가능하게 한다. ELECTRA는 더 적은 파라미터(parameter)로 자연어처리 분야의 SOTA(State-of-the-art) 모델들보다 더 높은 성능을 달성하였다[20]. KoELECTRA[21]는 한국어 토큰에 특화된 ELECTRA의 변형 모델이며, 대량의 한국어 텍스트를 기반으로 학습되어 한국어 문장 내의 각 토큰의 의미와 그사이의 관계를 정밀하게 인식하므로 한국어로 기술된 사고사례의 자연어처리에 활용할 수 있다. 따라서 KoELECTRA를 미세조정 하여 제안하는 순차적 레이블링 기반 복수 인과관계 추출 방법의 성능을 검증한다.

3. 복수 인과관계 추출 방법

3.1 복수 인과관계 추출 방법의 개요

제안하는 복수 인과관계 추출 방법의 구현 과정은 Fig. 1에 도시한 바와 같이 3가지 단계로 구성된다. 첫 번째 단계인 데이터 수집 단계에서는 해양사고 사례를 수집하여 인과관계 추출을 위한 데이터셋(dataset)을 구축한다. 두 번째 단계인 데이터 전처리 단계에서는 학습에 활용할 문장을 추출하고 자연어처리의 입력 데이터로 변환하기 위해 분절화(tokenizing)를 수행한다. 또한, 해양사고에서 복수의 원인과 결과를 추출하기 위해 토큰별로 레이블링을 수행한다. 마지막으로 학습단계에서는 전처리가 완료된 데이터셋으로 KoELECTRA를 미세조정 한다. 각 단계에 대한 자세한 설명은 다음과 같다.

3.2 데이터 수집

해양안전심판원은「해양사고의 조사 및 심판에 관한 법률 제11690호)」에 의거 하여 해양사고를 조사 및 분석하고 있다. 조사 및 분석 결과는 재결서의 형태로

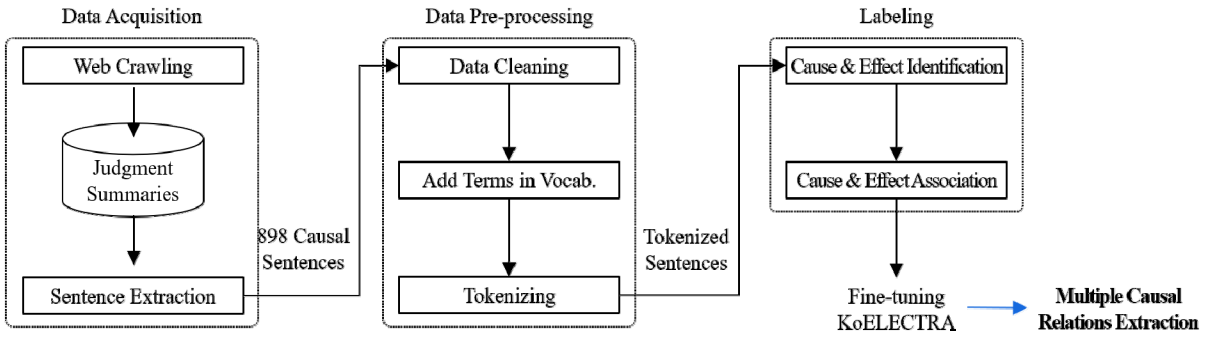

Fig. 1. Procedure of multiple causal relations extraction.

해양사고조사심판정보포털[22]에서 확인할 수 있으며, 이때 재결서를 3쪽 이내로 요약한 재결요약서도 함께 제공된다.

해양사고의 인과관계를 도출하기 위해서는 사고의 상세한 발생 원인과 결과가 학습 데이터에 포함되어야 하며, 재결요약서는 사고 발생 원인과 결과에 대한 자세한 분석을 포함하므로 해양사고의 인과관계 추출에 적합하다. 따라서, 본 연구에서는 2017년부터 2022년까지 해양안전심판원에서 발행한 총 898건의 재결요약서를 수집 및 활용하였다.

3.3 데이터 전처리

재결요약서는 사건명, 관련자, 원인판단 주제어, 주문(판결의 결론) 등의 항목으로 구성되어 있다. 따라서 본 연구는 재결요약서를 활용하여 인과관계 학습 데이터로 활용하였다. 재결요약서의 내용 중 해양사고의 원인 및 결과와 무관한 판결사항 등을 제외한 문장을 추출 및 활용하였으며, Table 1에 본 연구에서 활용한 학습 데이터의 예시를 나타내었다. 추출된 문장은 “지피에스(GPS)”, “접현(接舷)”과 같이 영어 및 한자가 괄호 안에 병기되어있다. 이와 같은 병기는 제한된 단어사전에서 문맥 이해 성능을 떨어뜨린다 판단하여 제거하였으며, 총 342개의 영어 및 한자 단어를 제거하였다.

Table 1. Example of sentence in a summary of a decision

Example of sentence (in English)
“이 **충돌사건**은 시정이 양호한 주간에 포항항 항계 내 영일만항 입구에서, 영일만항 항로를 횡단하던 일성호가 **경계를 소홀히** 하여 포항파일럿2호의 **진로를 피하지 아니**하여 발생한 것이나, 항로를 항행하던 포항파일럿2호가 **경계 소홀**로 **피항협력동작을 취하지 아니**한 것도 일부 원인이 된다.”
(“This **collision accident** occurred with favorable conditions, at the entrance of Yeongil Bay Harbor within the Pohang Harbor route. The Ilsungho, which was crossing the Yeongil Bay Harbor route, **negligected of lookout** and **did not avoid the path** of Pohang Pilot 2. However, the Pohang Pilot 2, which was navigating the route, also contributed to the cause by **neglecting to lookout** and **not taking evasive actions**.”)

Table 2. Examples of tokenizing

Sentence (in English)	Token (in English)	
“이 **충돌사건**은 시정이 양호한 주간에 포항항 항계 내 영일만항 입구에서, 영일만항 항로를 횡단하던 일성호가 **경계를 소홀히** 하여 포항파일럿2호의 **진로를 피하지 아니**하여 발생한 것이나, 항로를 항행하던 포항파일럿2호가 **경계 소홀**로 **피항협력동작을 취하지 아니**한 것도 일부 원인이 된다.”	충돌	(collision accident)
	##사건	
	경계	(neglected of lookout)
	##를	
	소홀히	
	진로	(did not avoid the path)
	##를	
	피하	
	##지	
	아니	
(“This **collision accident** occurred with favorable conditions, at the entrance of Yeongil Bay Harbor within the Pohang Harbor route. The Ilsungho, which was crossing the Yeongil Bay Harbor route, **negligected of lookout** and **did not avoid the path** of Pohang Pilot 2. However, the Pohang Pilot 2, which was navigating the route, also contributed to the cause by **neglecting to lookout** and **not taking evasive actions**.”)	경계	(neglecting to lookout)
	##소홀	
	피항	(not taking evasive actions)
	##협력	
	##동작	
	##을	
	취하	
	##지	
	아니	

영어와 한자가 제거된 문장을 KoELECTRA 토크나이저(tokenizer)를 이용하여 자연어처리 입력 단위인 토큰으로 분절화하였다. 이 과정에서 468개의 전문용어를 수집 후 단어사전에 추가하여 분절 오류를 방지하였다. 추출한 문장은 최종적으로 총 69,581개의 토큰으로 분절되었고, 한 문장은 평균 77개의 토큰으로 구성된다. Table 1 사례의 토큰화 예시는 Table 2와 같으며, 사례를 구성하는 토큰이 많아 굵게 표시된 부분만 나타내었다. 이때, (meaning)단위로 묶인 토큰은 해양사고 사례의 주요 원인과 결과이다.

3.4 인과관계 레이블링

분절된 토큰에 레이블링 작업 시, 작업자의 주관성을 배제하고자 다음과 같은 5가지 원칙을 토대로 인과관계 레이블을 부여하였다.

1) 레이블은 토큰 단위로 부여하며, 사고의 원인이나 결과를 구성하는 토큰을 그룹화하여 토큰을 부여함

2) 인과가 이어져 일어난 말단 사건 바로 뒤에는 항상 조건설에 위배 되지 않는 강한 인과관계로 이어져야 함
3) 지엽적인 정보(행위 주체, 대상 장비, 장소, 시간) 등은 제외함. 단, 제외 시 사건의 원인과 결과를 이해할 수 없는 경우에는 포함함.
4) 어미(“하여”, “해서” 등), 조사(서술격 조사: “이다” 등), 의존 명사 등은 레이블링하지 않으며, 주격 조사는 레이블링함. 단, 제외 시 사건의 원인과 결과를 이해할 수 없는 경우에는 포함함.
5) 인과관계를 파악할 수 있는 표현(‘-로’, ‘-로 인해’, ‘-하여’ 등)을 기준으로 선행하는 원인과 결과를 분리하여 레이블링함

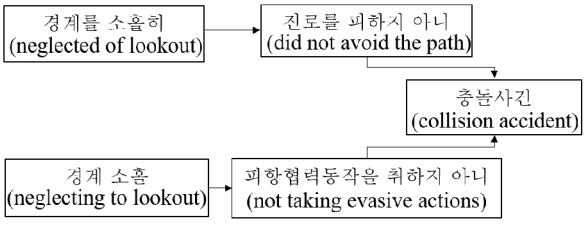

Fig. 2. Causal relation diagram.

위 규칙을 적용하여 Table 2 사례의 인과관계를 도식화하면 Fig. 2와 같다. 최초 원인인 “경계를 소홀히”와 “경계 소홀”은 각각 이어진 사건인 “진로를 피하지 아니”, “피항협력동작을 취하지 아니”와 연결되며, 이어진 사건들은 사건의 결과인 “충돌사건”과 연결된다.

본 연구의 레이블링 방법은 Fig. 2의 도식화 모형을 기반으로 원인-결과의 연결정보를 표현하고자 하였다. 이를 위해, 원인-결과의 식별 단계와 원인-결과 연결 단계로 구분하여 진행하였다. 첫 번째 단계인 원인-결과의 식별 단계에서는 각 토큰이 사고 사례에서 원인인지, 결과인지를 판별하며, i에 사례 내 사건의 출현 순서를 표현한다. 자세한 규칙은 다음과 같다.

- C_i : 원인 또는 원인이자 결과를 나타내는 사건
- E: 사고 최종 결과를 나타내는 사건
- O: 원인 또는 결과가 아님

위 규칙을 적용해 Table 2의 사례에 레이블링을 부여하면 Fig. 3과 같다. “경계를 소홀히”는 사고의 원인이므로 C_1을 부여한다. “진로를 피하니 아니”는 C_1의 결과이자 “충돌사건”의 원인이므로 C_2를 부여한다. 사건의 유일한 결과인 “충돌사건”은 E를 부여한다. 또한, “경계 소홀”은 “피항협력동작을 취하지 아니”의 원인이므로 C_3을 부여한다.

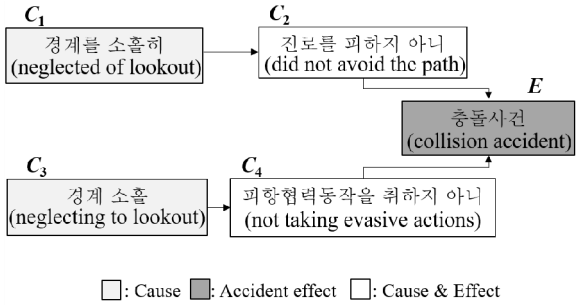

Fig. 3. Causal relation diagram and examples of labeling for cause–effect identification.

두 번째 단계인 원인-결과의 연결 단계에서는 인과관계 키워드 사이의 연결정보를 표현하기 위해 아래와 같이 원인과 원인 혹은 원인과 결과의 연결정보를 표현하는 레이블을 부여한다.

- C_iC_j : 원인 C_i가 원인 C_j를 직접적으로 초래했음을 나타냄
- C_iE: 원인 C_i가 사건 결과 E를 직접적으로 초래했음을 나타냄

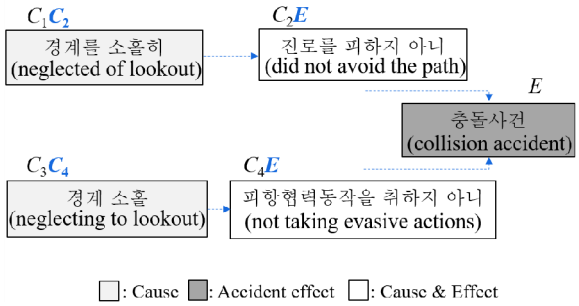

Fig. 4. Causal relation diagram and examples of labeling for cause–effect association.

위 규칙을 적용하여 Table 2 사례의 원인-결과의 연결정보를 나타내면 Fig. 4와 같다. Fig. 4에서 원인 “경계를 소홀히”가 원인 “진로를 피하니 아니”를 유발했을 경우, “경계를 소홀히” 토큰에 C_1C_2 레이블을 부여한다. 이는 원인 C_1이 원인 C_2를 직접적으로 초래했음을 나타낸다. 이와 같은 방식은 원인과 원인, 또는 원인과 결과 사이의 직접적인 인과관계를 표현하는 레이블 체계이다. 따라서, 원인-결과의 식별 단계와 원인-결과 연결 단계를 거친 레이블링 방식을 통해 사건의 복수 인과관계 정보를 포함하는 학습 데이터셋 구축이 가능하다. 레이블링 과정에서 출현순서와 연결정보를 고려한 총 38개의 레이블로 학습 데이터를 생성하였다. 전체 사례인 898개의 사례에서 사례당 평균 5.63종류의 원인이나 결과를 레이블링하였다.

Table 3. Average label usage and ratio per case

Label	Label uses	Label ratio(%)
C_iC_j	1.71	14.47
C_iE	1.92	3.32
E	1.00	9.37
O	1.00	72.84

또한, 본 연구에서 사용한 레이블 체계인 C_iC_j, C_iE, E, O의 사례당 평균 사용 횟수와 비율은 Table 3과 같다. 본 연구에서는 해당 레이블을 바탕으로 복수 인과관계 추출을 위한 사전언어모델을 학습하였다. 학습 과정에서 사전학습 언어모델은 토큰의 주변 문맥 정보를 학습하므로 등장순서와 연결정보를 학습할 수 있다.

4. 복수 인과관계 추출 성능 평가

4.1 미세조정

3장에서 설명한 방법의 성능을 검증하기 위하여 3장에서 설명한 방법으로 생성된 학습데이터를 바탕으로 KoELECTRA를 미세조정 하였다. 레이블링이 완료된 데이터를 8대 2의 비율로 학습과 검증 데이터셋으로 분할 하여 KoELECTRA를 미세조정 하였으며, 문장 내 각 토큰이 38개의 레이블 중 어느 하나에 해당하는지 학습한다. 토큰 분류를 위해 KoELECTRA의 출력층에 노드 수가 38개인 완전 연결 층(fully-connected layer)를 추가하였으며, 과적합 방지를 위해 Dropout을 0.25로 설정하였다.

미세조정 과정에서 딥러닝 모델의 하이퍼 파라미터(hyper-parameter)는 인과관계 추출 성능에 영향을 끼치므로 사전실험을 통해 최적화해야 한다. 주요한 하이퍼 파라미터로 배치 크기, 학습률 등이 있으며, 본 연구에서 사용한 하이퍼 파라미터를 Table 3에 정리하였다. Table 4를 통해 확인할 수 있는 바와 같이 배치 크기는 64로 설정하였으며, 학습률은 사전실험을 통해 각각 5e-5로 설정하였다. 최적화 함수는 안정성과 빠른 수렴 속도로 인하여 일반적으로 사용되는 아담(Adam)을 사용하였고, 손실 함수 또한 다중 분류에서 널리 사용되는 크로스 엔트로피(cross-entropy)를 사용하였다.

Table 4. Hyper-parameter settings

Type	Value
Batch size	64
Learning rate	5e-5
Optimizer	Adam
Loss function	cross-entropy

하이퍼 파라미터를 기반으로 미세조정을 진행한 환경은 다음과 같다. 먼저, 하드웨어로 GPU NVIDIA 3090 Ti, CPU Intel i9-12900KF를 사용하였다. 소프트웨어는 Windows11 운영체제에서 Python 3.9.1, CUDA 11.8, cuDNN 8.0.5를 사용하였다.

4.2 복수 인과관계 추출 성능 평가

제안된 복수 인과관계 추출 모델은 다중 분류 모델이므로, 다중 분류 모델의 대표적 성능 지표인 F1 점수를 사용하여 미세조정된 KoELECTRA의 성능을 측정하였으며, F1 점수를 산출하는 식은 아래와 같다.

$$Recall = \frac{TP}{(TP+FN)} \tag{1}$$

$$Precision = \frac{TP}{(TP+FP)} \tag{2}$$

$$F1-score = \frac{2*TP}{(2*TP+FN+FP)} \tag{3}$$

여기서, TP는 참 양성(true positive)이며, FP는 거짓 양성(false positive)이다. FN은 거짓 음성(false negative)이다. 식(1)의 재현율(recall)은 실제 양성 중 참 양성이라고 분류된 사례의 비율이다. 식(2)의 정밀도(precision)는 예측 양성과 참 양성 중 참 양성의 비율이다. F1 점수는 식(3)과 같이 정밀도와 재현율의 조화평균이다.

수렴 상태와 과적합 여부를 판단하기 위하여 에포크별 학습 손실과 검증 손실을 사용하였으며, 검증 손실이 최소화되는 지점에서 모델 가중치는 과적합이 되지 않은 최적의 성능을 나타내는 것으로 판단하였다.

Fig. 5는 에포크의 변화에 따른 학습 손실과 검증 손실 곡선, F1 점수를 나타낸 그래프이다. Fig. 5를 통해

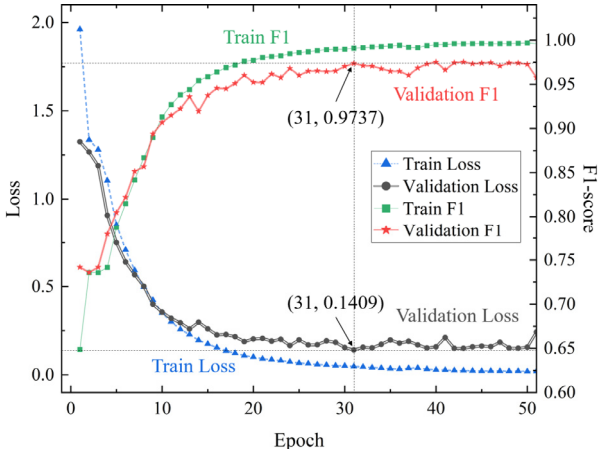

Fig. 5. Loss curve and F1-score by epoch.

확인할 수 있듯이 에포크 31을 지나면서 검증 손실과 F1 점수가 수렴하는 경향을 보인다. 본 연구에서 제시한 복수 인과관계 추출 방법은 에포크 31에서 F1 점수 0.9737을 기록하였다. 따라서, 본연구에서 제시한 인과관계 추출 방법을 바탕으로 해양사고 사례 문장 내 복수의 인과관계를 추출할 수 있다.

4.3 복수 인과관계 추출 방법의 활용

Table 5는 복수의 인과관계를 포함하는 해양사고 사례의 토큰별 인과관계 예측 결과이다. 해당 사례는 해양안전심판원에서 발행한 재결요약서에서 발췌하였으며, 4.1절에 설명한 학습 및 검증 데이터셋에 포함되지 않은 사례이다. Table 5를 통해 확인할 수 있듯이, 3.4절의 인과관계 레이블링 기준과 동일하게 인과관계를 예측하였다. 사건의 최종 결과인 "충돌사건" 토큰을 E로 예측하였으며, 최초원인인 "정비점검을 소홀히"부터 사례 내 사건의 출현순서에 따라 번호가 부여되었음을 확인할 수 있다. 또한, 출현순서와 연결정보를 고려하여 올바르게 예측했음을 확인할 수 있다. 따라서, 본 연구에서 제안하는 방법을 적용하면 복수의 인과관계를 추출할 수 있다. 추출된 레이블 정보를 바탕으로 Table 5 사례의 인과관계를 Fig. 6에 도식화하였다. Fig. 6을 통해 확인할 수 있듯이

Table 5. Application for causal relation extraction

Sentence (in English)	Token (in English)		Prediction
이 **충돌사건**은 126득명호가 발전기의 **정비점검을 소홀히** 하여 출항 중 갑자기 **발전기가 정지**되어 조타장치 등에 **전원이 공급되지 않**아 발생한 것이나, 선장이 **조타불능상태**에서 적절한 **비상조치를 하지 않**는 것도 일인이 된다. (This **collision accident** occurred because the ship, named 126 Deukmyeongho, **neglected maintenance** of its generator. This led to a sudden **generator failure** while sailing, resulting in a **loss of power** to the steering system and other equipment. Additionally, the captain's **failure to take appropriate emergency actions** during the loss of steering control also contributed to the accident.)	충돌	(collision accident)	E
	##사건		
	정비	(neglected maintenance)	C_1C_2
	점검		
	을		
	소홀히		
	발전기	(generator failure)	C_2C_3
	#가		
	정지		
	전원	(loss of power)	C_3E
	이		
	공급		
	#되		
	#지		
	않		
	조타	(the loss of steering control)	C_4C_5
	불능		
	상태		
	적절	(failure to take appropriate emergency actions)	C_5E
	#한		
	비상		
	조치		
	#를		
	하지		
	않		

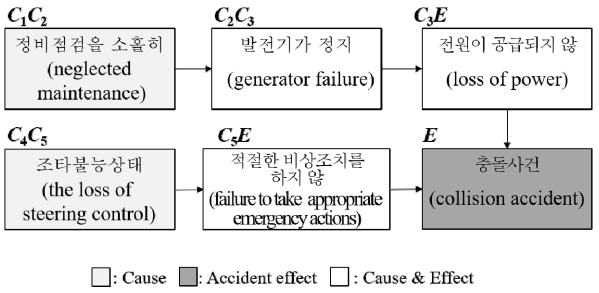

Fig. 6. Causal relation diagram of verification case.

레이블을 활용하여 원인과 결과를 연결함으로써 사건의 전체적인 인과 구조를 파악할 수 있다.

5. 결 론

본 연구에서는 해양사고의 복수 인과관계를 추출하기 위해 순차적 레이블링 기반 복수 인과관계 추출 방법을 제안하였다. 인과관계를 분석하여 복수의 원인과 결과의 연결정보를 고려하여 레이블링을 수행하였다. 제안된 방법의 성능은 FI 점수를 통해 검증하였으며, 이를 바탕으로 다음과 같은 결론을 도출하였다.

첫째, 본 방법은 문장 내 복수의 인과관계가 등장해도 식별할 수 있는 레이블링 방법을 사용하므로 문장에서 복수의 인과관계를 추출할 수 있다. 따라서, 복수의 원인이 결합되어 발생하는 해양사고 사례를 대상으로 인과관계를 추출할 수 있고 이를 사고 예방을 위한 대응 전략 수립에 활용할 수 있다.

둘째, 본 연구에서 제안된 방법은 복수의 원인 및 결과의 순서와 연결정보를 고려하여 레이블을 부여하므로, 예측된 레이블을 통해 인과관계를 구조화할 수 있다. 따라서 이를 활용하여 해양사고의 인과관계를 사용자에게 직관적으로 전달할 수 있다.

하지만 본 연구에서 제안하는 방법은 재결요약서의 주문을 대상으로 학습하므로 다른 유형의 문서나 데이터에서의 일반화 능력에 대한 검증이 추가로 필요하다. 또한, 일부 사례에서 인과관계가 명확하지 않거나 생략된 경우, 원인과 사고 최종 결과 사이의 사건을 파악하기 어렵다는 한계가 존재한다.

향후, 제안하는 모델을 통해 추출된 사고의 원인 · 결과 키워드를 사용자에게 제공하기 위한 시각화 네트워크 기술 및 안전사고 관리 UI를 개발할 계획이다.

Acknowledgement: This research was supported by Korea Institute of Marine Science & Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries (1525013138)

References

1) KMST, “Maritime Accidents Statistical Yearbook”, https://www.kmst.go.kr/web/stcAnnualReport.do?menuIdx=126, Retrieved on 08.31.2023.
2) J. Y. Choi, “A Study on the Causes of Marine Accidents and Prevention of Marine Accidents in Vessels”, Cultural Interaction Studies of Sea Port Cities, Vol. 25, pp. 337-359, 2021.
3) M. Luo and S. H. Shin, “Half-century Research Developments in Maritime Accidents: Future Directions”, Accident Analysis & Prevention, Vol. 123, pp. 448-460, 2019.
4) Y. Zhang, X. Sun, J. Chen and C. Cheng, “Spatial Patterns and Characteristics of Global Maritime Accidents”, Reliability Engineering & System Safety, Vol. 206, p. 107310, 2021.
5) R. J. Bye and P. G. Almklov, “Normalization of maritime accident data using AIS”, Marine Policy, Vol. 109, p. 103675, 2019.
6) W. Qiao, Y. Liu, X. Ma and Y. Liu, “Human Factors Analysis for Maritime Accidents based on a Dynamic Fuzzy Bayesian Network”, Risk Analysis, Vol. 40, No. 5, pp. 957-980, 2020.
7) S. Tirunagari, “Data Mining of Causal Relations from Text: Analysing Maritime Accident Investigation Reports”, arXiv preprint, arXiv:1507.02447, 2015.
8) B. Navas de Maya, O. Arslan, E. Akyuz, R. E. Kurt, and O. Turan, “Application of Data-mining Techniques to Predict and Rank Maritime Non-conformities in Tanker Shipping Companies using Accident Inspection Reports”, Ships and Offshore Structures, Vol. 17, No. 3, pp. 687-694, 2022.
9) J. I. Single, J. Schmidt and J. Denecke, “Knowledge Acquisition from Chemical Accident Databases using an Ontology-based Method and Natural Language Processing”, Safety Science, Vol. 129, p. 104747, 2020.
10) G. Liu, M. Boyd, M. Yu, S. Z. Halim and N. Quddus, “Identifying Causality and Contributory Factors of Pipeline Incidents by Employing Natural Language Processing and Text Mining Techniques”, Process Safety and Environmental Protection, Vol. 152, pp. 37-46, 2021.
11) G. Perboli, M. Gajetti, S. Fedorov, and S. L. Giudice, “Natural Language Processing for the Identification of Human Factors in Aviation Accidents Causes: An Application to the SHEL Methodology”, Expert Systems with Applications, Vol. 186, p. 115694, 2021.
12) Y. J. Lee, J. H. Park and S. D. Lee, “Named Entity Recognition and Causal Relation Extraction Based on Pre-trained Language Model for Safety Accident Analysis”, Journal of Korean Institute of Intelligent Systems, Vol. 33, No. 4, pp. 360-367, 2023.
13) J. M. Hwang and S. W. Shin, “Correlational Structure Modelling for Fall Accident Risk Factors of Portable Ladders Using Co-occurrence Keyword Networks”, J. Korean Soc. Saf., Vol. 36, No. 3, pp. 50-59, 2021.
14) Y. G. Yoon, J. Y. Lee, and T. K. Oh, “Text mining-based Data Preprocessing and Accident Type Analysis for Construction Accident Analysis”, J. Korean Soc. Saf., Vol. 37, No. 2, pp. 18-27, 2022.
15) A. Akbik, D. Blythe and R. Vollgraf, “Contextual String Embeddings for Sequence Labeling”, In Proceedings of the 27th International Conference on Computational Linguistics, pp. 1638-1649, 2018.
16) A. Vaswani et al., “Attention is All You Need”, Advances in Neural Information Processing Systems, Vol. 30, 2017.
17) J. Devlin, M. W. Chang, K. Lee, & K. Toutanova, “Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv Preprint arXiv:1810.04805, 2018
18) Y. Liu et al., “Roberta: A Robustly Optimized Bert Pretraining Approach“, arXiv preprint arXiv:1907.11692, 2019.
19) K. Clark, M. T. Luong, Q. V. Le, & C. D. Manning, “Electra: Pre-training Text Encoders as Discriminators Rrather than Generators”, arXiv Preprint arXiv:2003.10555, 2020.
20) J. M. Jang, J. O. Min and H. S. Noh, “KorPatELECTRA : A Pre-trained Language Model for Korean Patent Literature to Improve Performance in the Field of Natural Language Processing(Korean Patent ELECTRA)”, Journal of The Korea Society of Computer and Information, Vol. 27, No. 2, pp. 15-23, 2022.
21) J. Park, “Koelectra: Pretrained Electra Model for Korean”, GitHub Repository, 2020.
22) KMST, “Web site for the Investigation and Judgement Information Portal of Maritime Causalities”, https://www.kmst.go.kr/web/verdictList.do?menuIdx=121, Retrieved on 08.31.2023.