creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학전문석사 학위 연구보고서

BERT 기반 언어 모델링을 활용한 북한 도발 징후 포착

BERT-Based Catching Signs of Provocation of North Korea Using Language Modeling

2021년 2월

서울대학교 공학전문대학원

응용공학과 응용공학전공

김 동 훈

BERT 기반 언어 모델링을 활용한 북한 도발 징후 포착

지도교수 윤 성 로

이 프로젝트 리포트를 공학전문석사 학위
연구보고서로 제출함
2021년 2월

서울대학교 공학전문대학원
응용공학과 응용공학전공
김 동 훈

김동훈의 공학전문석사 학위 연구보고서를 인준함
2021년 2월

위 원 장 _____ 구 윤 모 _____ (인)

위 원 _____ 곽 우 영 _____ (인)

위 원 _____ 윤 성 로 _____ (인)

국문초록

북한은 6.25 전쟁 이후에도 대한민국의 안보와 국제사회를 위협하는 전략·전술적 도발을 끊임없이 자행해왔다. 이로 인해, 무수히 많은 국군 장병이 희생되었으며, 피해복구 비용과 외국 투자자본 이탈 등의 경제적 손실 또한 발생하였다. 하지만, 북한은 자신들의 이러한 도발 행위를 인정하지 않고 있음은 물론, 진정성 있는 사과 한번 제대로 하지 않고 있다. 우리 군은 언제 또 다시 기습적으로 발생할지 모르는 북한의 도발 위협으로부터 대한민국의 안보와 국민의 생명, 재산을 지키기 위해 다양한 정보감시 자산을 운용하여 북한의 특이 동향과 이상 징후를 식별하고 있으나, 모든 위협을 예측하고 대응하기에는 한계가 있는 것이 사실이다. 따라서 본 연구에서는 Digital Transformation에 발맞춰 북한의 도발을 예측하는데 있어 최신 딥러닝 기반 자연어 처리 모델인 BERT 모델을 활용하여 북한 도발과 관련된 텍스트 데이터로부터 북한의 도발 징후를 포착하는 모델을 제안하였다. 과거 북한 도발 사례 11건을 선정하고, 각 도발 발생일 이전 1개월간의 데이터와 도발이 발생하지 않았던 평시 데이터를 비교 학습한 후 학습 단계와 평가 단계의 정확도를 Fine-tuning을 통해 최적화하고, 새로운 임의의 문장을 모델에 대입해봄으로써 성능을 검증해보았다. 궁극적으로 정보감시 자산의 활용 외에 최신 과학기술을 적용하여 북한 도발 징후를 포착함으로써 우리 군의 감시능력 강화를 통해 아군의 피해를 최소화하고 전투에서 승리하는데 기여하고자 함이다.

주요어 : 버트, 딥러닝, 텍스트 마이닝, 언어 모델, 자연어 처리, 북한 도발
학 번 : 2019-28970

목 차

제 1 장 서 론 ····· 1

제 1 절 연구동기 ····· 1

제 2 절 연구내용 ····· 3

제 3 절 논문구성 ····· 4

제 2 장 배경지식 및 관련연구 ····· 5

제 1 절 언어 모델 ····· 5

1. N-gram ····· 5

2. RNN ····· 7

3. Attention ····· 8

제 2 절 BERT 모델 ····· 11

1. Input Embeddings ····· 11

2. Structure of Pre-trained BERT ····· 13

3. 한국어 BERT 모델 ····· 14

제 3 절 주요 북한 도발 사례 ····· 16

제 3 장 실험 설계 및 데이터 ····· 19

제 1 절 실험 모델 ····· 19

제 2 절 실험데이터 수집 ····· 20

제 4 장 실험결과 ····· 22

제 1 절 실험데이터 탐색적 분석 ····· 22

제 2 절 학습 및 평가 단계 결과 ····· 25

제 3 절 새로운 문장 입력 결과 ····· 27

제 5 장 결 론 ····· 30

참고문헌 ····· 32

Abstract ····· 36

표 목 차

[표 2-1] Multilingual BERT 모델과 KoBERT모델 사전 구성 비교 ····· 15
[표 2-2] 주요 북한 도발 사례 선정 ····· 18
[표 3-1] 총 데이터 중 학습 및 평가 데이터 비율 ····· 21
[표 3-2] 학습 데이터 중 도발 징후 포착기 및 평시 데이터 비율 ····· 21
[표 3-3] 평가 데이터 중 도발 징후 포착기 및 평시 데이터 비율 ····· 21
[표 4-1] epoch, batch size 조정 결과 비교(학습 데이터/평가 데이터) ····· 26
[표 4-2] epoch, batch size 조정 후 learning rate 조정 결과 비교(학습 데이터/평가 데이터) ····· 26

그 림 목 차

[그림 2-1] RNN 모델 아키텍처 ····· 7
[그림 2-2] Attention 모델 아키텍처 활용 예시 ····· 10
[그림 2-3] BERT 모델 입력 구조 ····· 12
[그림 2-4] Transformer 모델 아키텍처 ····· 14
[그림 4-1] 도발 징후 포착기 데이터 워드 크라우드 ····· 23
[그림 4-2] 도발 징후 포착기 데이터 상위 20개 단어 빈도 막대그래프 ····· 23
[그림 4-3] 평시 데이터 워드 크라우드 ····· 24
[그림 4-4] 평시 데이터 상위 20개 단어 빈도막대그래프 ··· 24
[그림 4-5] 학습 데이터 10개 출력 ····· 26
[그림 4-6] 도발 징후 포착기 데이터 새로운 문장 입력 테스트 ····· 28
[그림 4-7] 평시 데이터 새로운 문장 입력 테스트 ····· 29

제 1 장 서 론

제 1 절 연구동기

북한은 1953년 7월 27일 정전협정 이후에도 대남적화통일 야욕과 대내외 정치적 목적으로 판문점 도끼만행 사건(1978.8.18.), 아웅산 묘소 폭탄 테러(1983.10.9.), 강릉 무장공비 침투사건(1996.9.18.), 제1・제2 연평해전(1999.6.15., 2002.6.29.), 천안함 피격(2010.3.26.), 연평도 포격도발(2010.11.23.), DMZ 목함지뢰 매설(2015.8.4.), 여섯 차례의 핵실험(2006.10.9.～2017.9.3.), 최근 남북공동연락사무소 폭파(2020.6.16.), 수십 차례의 탄도미사일 발사와 이외에도 무수히 많은 전략・전술적 도발을 끊임없이 자행해왔다. 북한은 자신들의 도발행위를 인정하지 않고 있으며, 적반하장 격으로 대한민국이 자신들을 국제사회에서 고립하기 위해 거짓을 말하고 있다며 억지 주장을 해오고 있다. 이러한 천인공노할 북한의 기습적 도발로 수많은 국군 장병이 목숨을 잃었을 뿐만 아니라, 심지어 연평도 포격도발 사건에서는 무고한 민간인 사망자가 2명이나 발생하는 비통함도 겪어야했다. 우리 군은 이러한 북한의 도발행위를 사전에 감지하고 대응하기 위해 다양한 정보감시 자산을 운용하고 있으나 여러 변수들을 예측하기에는 일부 운용상의 한계가 존재한다[1]. 따라서 북한의 도발 징후를 정확히 포착할 수 있도록 감시능력을 향상시키는 것 또한 군이 해결해야할 과제일 것이다. 한편, 우리 군도 4차 산업혁명 시대에 부합하기 위해 AI, 빅데이터 분석, 5G 통신 등 첨단 기술을 국방과학기술에 접목하기 위한 체계적인 노력을 하고 있다[2]. 특히, 2019년 1월 국방부차관을 단장으로 '스마트 국방혁신 추진단'이 출범하였으며, 4차 산업혁명 기술 기반 스마트 국방혁신을 추진하고 있다. 스마트 국방혁신의 핵심 과제는 '지능형 스마트부대'를 구축하는 것으로 최신 과학기술을 적용해 실시간으로 작전 상황과 모든 부대 현황을 파악하고 병력 및 군

수, 인사, 재정 관리의 효율성을 강화하는 것이다. AI와 빅데이터 기술을 통해 지휘관의 상황판단-결심-대응 능력을 지원하는 또 하나의 참모 역할을 수행하도록 하는데 초점을 맞추고 있다. 방위사업청은 최근 공개한 군사형 무기 드론 체계의 도입을 위해 올해부터 향후 10년간 2조 7천억원을 투자할 계획이며, 국군의무사령부는 의료데이터와 AI 의료분야 기업의 기술을 접목시켜 AI 융합 의료영상 진료·판독을 지원해주는 솔루션을 의무부대에 구축하는 등 가시적 성과도 달성하고 있다. 이렇듯 기존의 정보감시 자산 운용과 군사전문가의 경험적 판단 외에도 최신 과학기술을 이용하여 북한의 도발 징후를 포착하고 이를 통해 궁극적으로 현행 작전에 활용할 수 있다면 국민의 생명과 재산, 곁의 전우를 지킬 수 있는 효과적인 방안이 될 수 있을 것이라는 생각에 이번 연구를 진행하게 되었다.

제 2 절 연구내용

이번 연구는 최신 자연어 처리(NLP) 딥러닝(Deep learning)[3] 기술인 BERT 모델을 기반으로 북한의 도발 징후를 포착해보고자 시도한 것이다. 북한 도발과 관련된 인터넷 언론보도와 정부부처의 공개자료 등의 텍스트데이터를 수집하여 텍스트 마이닝[4] 기법을 바탕으로 연구를 진행하였다. 텍스트 마이닝은 수집한 데이터에서 숨겨진 문맥적 관계 또는 패턴을 추출해내어 유의미하고 유가치한 정보를 찾아내는 기법이다. 데이터 수집을 위한 기준을 선정한 후 주요 북한 도발 사례 총 11건을 선정하였다. 수집한 데이터는 두 개의 유형으로 분류하였는데, 하나는 도발 징후가 있는 것으로 label을 분류하였고, 다른 하나는 도발 징후가 없는 것으로 label을 분류하였다. 분류한 데이터의 유가치성을 확인하기 위해 R코드를 통해 상위 빈출 단어를 분류해내고 이를 워드 크라우드와 빈도 막대그래프로 구현하여 유의미한 차이가 있는지 탐색적 데이터 분석(EDA, Exploratory Data Analysis)을 통해 확인해보았다. Pre-trained BERT 모델[5]은 최초 구글에서 공개한 Multilingual BERT 모델에 비해 한국어에서 더 우수한 성능을 보이는 SK-T Brain에서 공개한 KoBERT 모델을 기반으로 하였으며, 수집한 데이터를 도발 징후가 있는지와 없는지 이진분류(Binary classification) 하도록 모델을 구성하였다. 모델은 학습 단계와 평가 단계에서의 정확도를 Fine-tuning 과정을 통해 최적화하였으며, 최종적으로 새로운 텍스트데이터에서 북한의 도발 징후를 식별할 수 있는지를 검증하기 위한 목적으로 임의의 새로운 문장을 입력, 분류하는 실험을 진행하였다. 최근 자연어 처리 분야에서 BERT 모델을 이용한 다양한 연구가 활발히 이루어지고 있으며, 좋은 성과를 보이고 있다[6, 7, 8]. 궁극적으로 본 연구를 통해 우리 군의 감시능력 향상에 기여함으로써 북한 도발 징후를 사전에 포착하고 선제적으로 대응하여 아군의 피해를 최소화하고 성공적인 작전을 수행하는데 도움이 되고자 함이다.

제 3 절 논문구성

제1장에서는 연구를 시작하게 된 동기와 배경, 대략적인 연구내용에 대해 서술하였다.

제2장에서는 배경지식 및 관련연구로 통계적 언어 모델과 딥러닝 기반의 BERT 모델에 대해 설명하고, 북한의 주요 도발 사례 선정 기준 및 각 도발 사례에 대해 소개하였다.

제3장에서는 실험 설계 및 데이터에 대해 기술하였으며, 실험모델의 동작 원리와 수집한 데이터에 대해 구체적으로 설명하였다.

제4장에서는 실험 결과로 우선 실험데이터에 대해 전반적으로 분석한 결과를 제시하였으며, 학습 단계와 평가 단계의 정확도를 Fine-tuning을 통해 조정한 과정과 새로운 임의의 문장을 모델에 대입해봄으로써 모델의 성능을 검증해보는 과정을 기술하였다.

마지막 제5장에서는 본 연구의 결론과 의의에 대해 논의하였다.

제 2 장 배경지식 및 관련연구

제 1 절 언어 모델

특정 연구를 수행하거나 과업을 달성하기 위해 텍스트 데이터를 수집하고, 수집한 데이터에서 새로운 문맥적 패턴과 관계를 분석하여 유의미한 내용을 찾아내는 연구 분야를 텍스트 마이닝(Text mining)이라고 한다. 인간이 사용하는 언어인 자연어(Natural language)는 컴퓨터가 이해할 수 없는 구조이므로 자연어 처리(Natural language processing)[9]를 통해 컴퓨터가 연산을 수행할 수 있도록 해야 한다. 이를 위해 자연어를 의미 공간에 숫자 형태의 벡터로 변환하는 작업이 필수적인데 이를 단어 임베딩(Word embedding)이라 한다. 하지만 단어 임베딩의 대표적 방법론인 Word2vec의 경우 중심 단어의 의미를 주변 단어를 근거로 예측하므로 단어가 가지는 또 다른 중의적 의미는 고려하지 못하는 단점이 있다. 다시 말해, 여러 의미를 가지는 동일 형태의 단어가 문맥에 따라 다른 의미를 가짐에도 Word2vec은 이를 해결하기에 제한적이다[10]. 그러나 딥러닝 기술이 발전하면서 이러한 문제를 해결할 수 있게 되었다. 이번 연구에 활용한 방법인 BERT 모델 또한 최신 자연어 처리 딥러닝 기술로써 BERT 모델에 대한 체계적 이해를 위해 과거 대표적인 통계적 기반의 언어 모델부터 딥러닝 기반의 언어 모델까지 논의해보고자 한다.

1. N-gram

초기 언어 모델은 통계적인 방법으로 접근하여 개발되었으며 이를 통계적 언어 모델(Statistical Language Model, SLM)이라고 한다. 텍스트 데이터의 토큰들에 조건부 확률을 적용하여 주어진 문장에서 다음 단어

가 올 수 있는 예측 확률을 구하는 방식이다. 후보 단어열 W의 출현확률을 $P(W)$라고 했을 때 (1)과 같이 정의할 수 있다[11]. 그러나 이러한 방법은 한 번도 학습에 등장하지 않은 패턴에 대해서는 확률값이 0이 되어 예측이 불가능 하며, 통계적 분석을 바탕으로 하므로 모든 문장에 대한 경우의 수를 파악할 수는 없다는 단점이 있다.

$$\begin{aligned} P(W) &= P(w_1, w_2, \cdots, w_T) \\ &= P(w_1)P(w_2|w_1)\cdots P(w_T|w_1, w_2, \cdots, w_{T-1}) \\ &= \prod_{i=1}^{T} P(w_i|w_1, \cdots, w_{i-1}) \end{aligned} \tag{1}$$

이를 보완하기 위해 등장한 또 다른 통계 기반의 언어 모델이 바로 N-gram이다. N-gram은 문장에 등장한 모든 단어들을 고려하여 계산하는 것이 아니라 일부 단어만을 고려하여 계산에 활용하는 방법으로 이때 몇 개의 단어를 활용하느냐가 N-gram에서의 N을 의미한다. 모든 가능한 단어열 W에 대한 조건부 확률을 구하는 것이 아니라 (2)와 같이 근사한 N-gram을 사용한다. 이때 활용하는 단어의 수를 의미하는 N은 N = k + 1의 값을 가지며 (3)은 N = 1인 Uni-gram의 경우, (4)는 N = 2인 Bi-gram의 경우, (5)는 N = 3인 Tri-gram의 경우이다. N이 커지게 되면 계산량과 학습에 필요한 자료가 많아지게 되므로 일반적으로 N = 2 또는 3을 이용하여 연산을 한다[12].

$$P(w_1, w_2, w_3, \cdots, w_T) \approx \prod_{i=1}^{T} P(w_i|w_{i-k}, \cdots, w_{i-1}) \tag{2}$$

$$\approx \prod_{i=1}^{T} P(w_i) \tag{3}$$

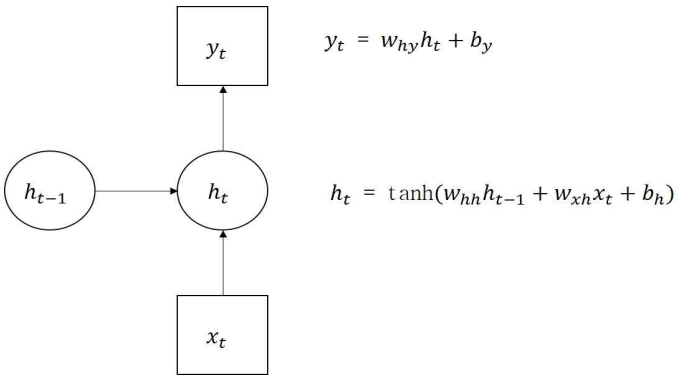
$$\approx \prod_{i=1}^{T} P(w_i|w_{i-1}) \tag{4}$$

$$\approx \prod_{i=1}^{T} P(w_i|w_{i-1}, w_{i-2}) \tag{5}$$

하지만, N-gram은 SLM 모델에 비해 단어를 예측할 수 있는 확률이 높아지긴 했지만 여전히 방대한 양의 데이터를 가지고 학습하여도 모든 경우의 수를 예측할 수 없으며, N 값이 커질수록 예측률은 높아지나 모델 사이즈도 커져 연산이 느려진다는 단점이 있다[13].

2. RNN

RNN(Recurrent Neural Network)은 음성이나 문자 등 순서를 가지고 순차적으로 이루어져있는 데이터를 처리하는데 알맞은 모델로 이전의 정보가 다음 단계에 영향을 줄 수 있도록 순환구조를 가지고 있다[14]. 그림 2-1에서 보는 것과 같이 RNN은 인풋 데이터 x_t가 들어가고 현재 step의 hidden layer h_t는 이전 stpe의 hidden layer h_{t-1}을 받아 갱신되는 형태이며, 아웃풋 y_t는 현재 step의 결과로 나타나게 되는 구조이다. 이러한 순환구조를 거쳐 최종 아웃풋은 단어가 가지는 문맥적 의미를 고려하여 만들어지게 된다. 그러나 RNN은 문장의 길이가 길어질수록 초반에 등장했던 단어의 의미가 희석되는 문제(Long-term dependency problem)가 발생하게 되며[15], 모든 단어들이 최종 아웃풋에 영향을 미치기 때문에 중요하지 않은 단어들도 결과에 영향을 미치며, 순차적 연산을 진행함으로 학습 속도가 느리다는 단점이 있다.

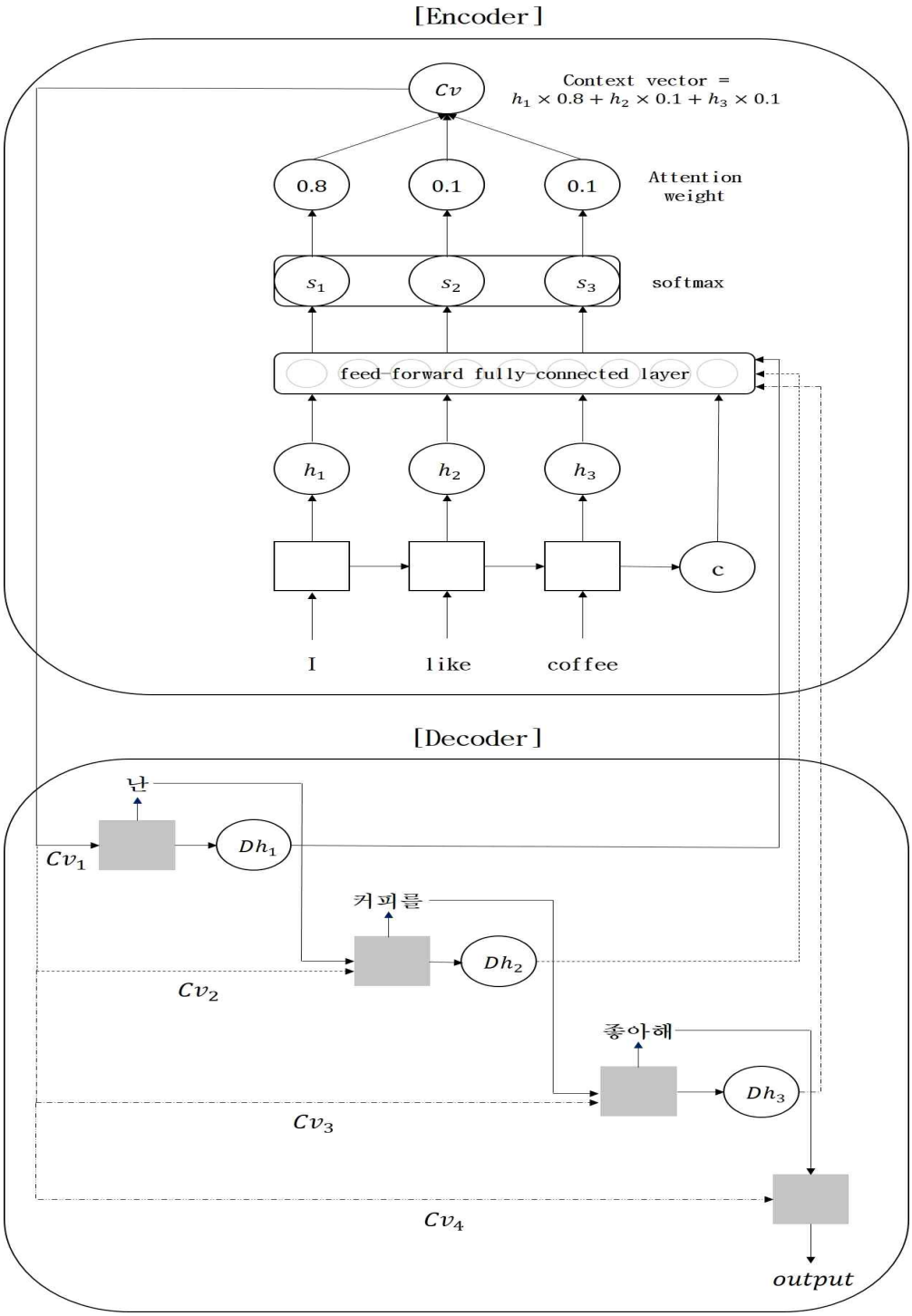


[그림 2-1] RNN 모델 아키텍처

3. Attention

Attention 모델은 앞서 설명한 RNN의 단점을 보완하기 위해 개발되었다. 인간이 데이터 정보를 처리할 때 모든 단어들에 대해 하나하나 의미를 두고 읽지 않고, 핵심 단어와 이를 부연하는 보충 단어들 위주로 읽으며 문서 정보를 이해하는 논리적 사고 체계와 같은 개념이다[16]. 기존 RNN에서는 각 RNN 셀에서 계산된 아웃풋 vector를 사용하지 않고 이전의 hidden state만 다음 hidden state로 전달하며 오로지 학습에만 사용하고, 최종 아웃풋인 Context vector를 생성해내는 반면, Attention은 각 RNN셀의 아웃풋 vector를 사용하여 학습 과정에서 역동적인 정보들을 이용하여 최종 Context vector를 생성해보자는 접근이다[17]. 그림 2-2에서 h_1은 “I”의 아웃풋, h_2는 “like“의 아웃풋, h_3는 “coffee”의 아웃풋, c는 기존 RNN 모델이 최종 아웃풋인 Context vector이다. 여기서 기존 RNN 모델의 각 셀에서 나온 아웃풋 h_1, h_2, h_3, c를 입력으로 하는 feed-foward fully-connected layer를 부착하고 이를 거쳐 나온 아웃풋을 각 RNN 셀의 score라고 했을 때, 이들 score에 softmax를 취하여 0에서 1사이의 값으로 변환하는 과정을 거치고 변환된 값들을 Attention 모델의 Attention weight로 결정한다. Attention 모델에서의 Context vector는 앞서 구한 Attention weight에 각 hidden state를 곱하여 얻게 된다[18]. 여기까지가 Attention 모델의 Encoder 수행 과정이며, Decoder에서는 Attention 모델의 Context vector를 인풋으로 넣어 나온 출력값인 Dh_1, Dh_2, Dh_3이 정답과 많은 차이를 보인 경우 Attention weight를 조정하여 Context vector를 보정하는 작업을 수행하게 된다. 다시 말해, hidden state와 fully-connected layer를 거쳐 나온 아웃풋 값들에 대해 Decoder에서 어느 정도의 정확도를 보였는지를 바탕으로 역전파(Backpropagation) 과정을 통해 Attention weight를 조정함으로써 최적의 결과를 찾아나가는 원리로 학습하게 되는 것이다. 이러한 Attention 모델의 장점은 인풋 데이터에 대한 Attention weight를 사용함으로써 인풋 데이터 중 중요한 의미를 갖는 데이터에 집중함으로써 성능을 향상시

킬 수 있다는 것이다. 하지만 Attention 모델에도 단점이 있는데, 여전히 RNN을 기반으로 함으로써 연산 속도가 느리다는 것이다. 이러한 RNN의 단점을 보완하기 위해 인코더와 디코더에서 RNN을 제거한 것이 Self-Attention 모델로 디코더 과정에서 발생하는 역전파를 통해 Attention weight를 조정할 필요 없이 최초 인풋 단계에서 Attention을 통해 인풋 값에 대한 상당히 정확도 높은 임베딩 값을 만들어보자는 개념이다. Self-attention 모델은 인풋을 임베딩한 후 임베딩된 벡터를 query, key, value 3개의 weight vector로 나누어주고, weight vector들의 상호 연산을 통해 Attention layer의 최종 아웃풋인 Context vector를 구하게 된다[19]. BERT 모델은 Self-Attention을 동시에 여러 개 수행하는 Multi-head Self-Attention 모델을 사용하고, 학습 과정에서는 Multi-head Self-Attention 모델로 이루어진 인코더를 여러 층 쌓은 Transformer 모델로 학습을 진행한다.



[그림 2-2] Attention 모델 아키텍처 활용 예시

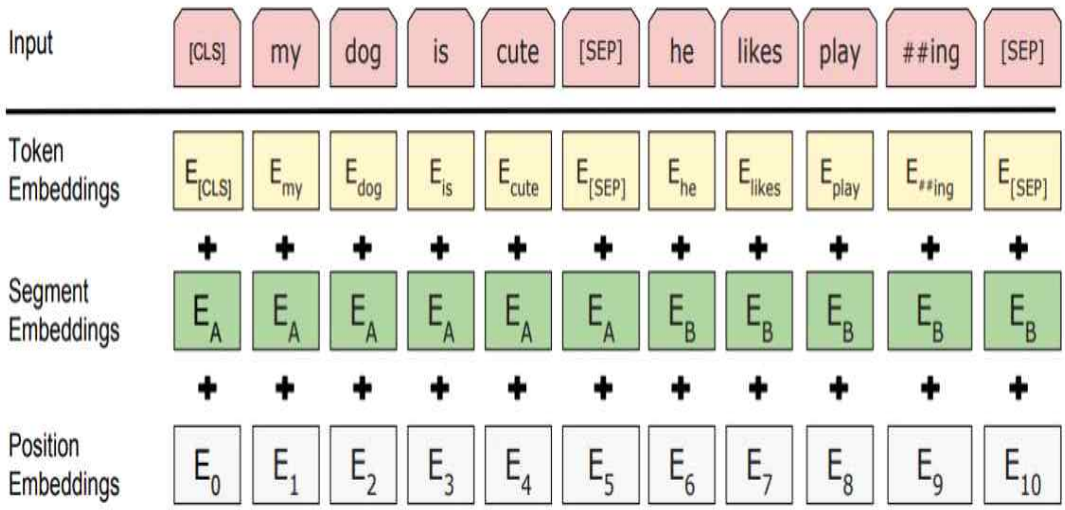
제 2 절 BERT 모델

BERT(Bidirectional Encoder Representations from Transformers) 모델[5]은 2018년 10월 구글에서 공개한 최신 자연어 처리 딥러닝 모델로 Multi-head Self-Attention 모델로 이루어진 인코더를 여러 층 쌓은 Transfomer를 기반으로 한 모델이다. 처음 등장했을 당시 자연어 처리 분야에서 인간의 퍼포먼스를 능가하는 결과를 보여주어 크게 주목받았고 현재도 자연어 처리 분야에서 매우 뛰어난 성능을 보여주고 있다. BERT 모델의 큰 특징은 학습 데이터가 부족하면 모델의 성능 또한 저하될 수 밖에 없는 딥러닝의 단점을 대용량 코퍼스를 통해 사전 잘 학습된 Pre-trained model을 사용하여 극복하였다는 것이다. 그리고 Pre-trained model에 최종적으로 사용 목적에 맞는 classification layer 1개 층을 추가하고 Hyperparameter를 조정하는 Fine-tuning 과정을 통해 우수한 성능을 보일 수 있다. 이번 절에서는 본 연구에 활용된 BERT 모델의 임베딩과 Pre-trained BERT 모델의 구조, 한국어 BERT 모델 중 KoBERT 모델[20]에 대해 논의하고자 한다.

1. Input Embeddings

BERT 모델은 그림 2-3에서 보는바와 같이 Token Embbedings, Segment Embeddings, Position Embeddings라는 세 가지 입력 절차로 구성된다[5]. 먼저 Token Embeddings는 BERT 모델은 기존의 단어 임베딩 방식을 사용하지 않고 WordPiece tokenizing 이라는 방식을 사용하였는데, 이로 인해 최초 구글에서 공개한 104개 언어의 Wikipedia 데이터로 학습한 Multilingual BERT가 등장할 수 있었고, 각 나라 언어들의 복잡한 문법적 특성을 고려하지 않았음에도 해당 국가의 언어 데이터에 적용이 가능하였다. WordPiece tokenizing은 단어가 등장한 빈도수에 기반하여 자주 등장하는 가장 긴 단어는 하나의 sub-word 그 자체로 분류

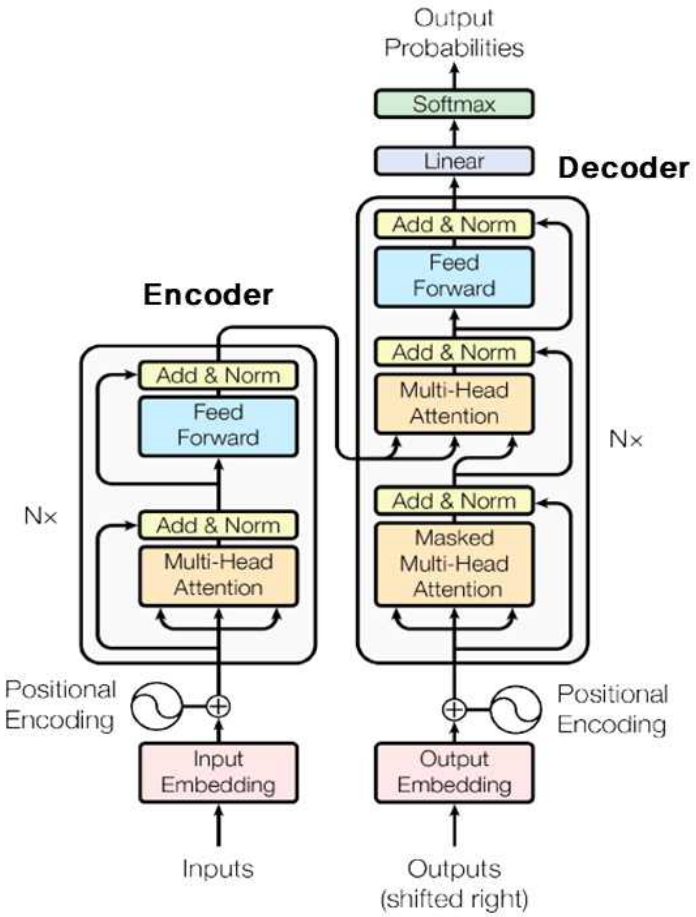
되고, 자주 등장하지 않는 단어는 sub-word로 쪼개지게 되는 방식으로 Tokenizing 하는 것으로 이름, 희귀단어 등 단어장에 없는 단어에 대한 학습이 이루어지지 않았던 Out-of-vocabulary(OOV) 문제 해결에 효과적이다. 다음으로 Segment Embeddings는 입력된 문장을 구분하기 위한 임베딩 방식으로 두 개의 문장 사이에 이를 구분하기 위한 [SEP]이 입력된 후 하나로 합쳐져 임베딩 된다. 적정 메모리 실행 수준 유지를 위해 입력 길이는 합쳐진 두 문장을 기준으로 512 sub-word 이하로 제한되며, 한국어 문장의 경우에는 통상 한 문장이 20~60 sub-word로 구성되므로 두 문장을 합쳤을 경우를 고려하여 입력 최대 길이를 128로 설정하여 학습을 진행하는 것이 일반적이다. 마지막 과정인 Position Embeddings는 BERT 모델에 사용된 Self-Attention 모델이 토큰의 입력 위치에 대한 정보를 고려하지 못하기 때문에 이를 보완하기 위해 고안된 임베딩 방식으로 입력 토큰의 위치 정보를 입력하기 위한 목적으로 사용하며, 단순히 토큰의 순서대로 0, 1, 2, 3,…과 같이 순서대로 번호를 부여하는 방식을 사용한다.



[그림 2-3] BERT 모델 입력 구조[5]

2. Structure of Pre-trained BERT

BERT 모델은 Transformer 모델을 기반으로 하며, Transformer 모델은 RNN의 long-term dependency problem을 해결하기 위해 탄생한 Attention 모델과 Attention 모델이 기존 RNN을 기반으로 하여 연산 속도가 느리다는 단점을 보완하기 위해 Attention 모델에서 RNN을 제거한 Self-Attention 모델을 동작 원리로 한다. Transformer 모델은 그림 2-4와 같이 feed-forward neural network를 이용한 인코더와 디코더 구조 모델이다[21]. BERT 모델은 Transformer의 인코더와 디코더 중 인코더만을 사용하여 학습한 언어 모델이다. 그리고 자연어 처리 분야에서 BERT 모델이 우수한 성능을 보이게 하는 또 다른 특징은 MLM(Masked Language Model)과 NSP(Next Sentence Prediction) 이다. MLM은 인풋 토큰을 일정 확률로 masking 처리하는 것으로, 일부 단어를 빈칸으로 만들고 원래 단어를 예측하도록 학습을 수행한다. 문장마다 15%의 단어를 선정하고 이중 80%만 masking 처리하며 나머지 10%는 상관없는 임의의 단어로 치환하고, 나머지 10%는 원래 단어 그대로 유지한다[22]. NSP는 두 문장이 주어졌을 때 두 번째 문장이 첫 번째 문장 다음에 오는 문장인지 아닌지를 예측하는 문제를 말한다. Q&A 관련 Task 수행처럼 순차적 특성을 가지는 텍스트 데이터의 특성을 고려하여 앞, 뒤 문장 상호 간에 얼마나 관련성이 있는지를 학습하기 위해 고안된 방식이다. 앞서 설명했듯이 두 문장을 하나의 임베딩으로 합치는 Segmentation embeddings에서 문장 구분 입력자 [SEP]을 사용하였는데, 두 문장을 50% 비율로 실제 연이어 등장하는 문장으로 두어 참인지(IsNext) 또는 50% 비율로 무작위로 추출된 문장으로 치환하여 거짓인지(NotNext) 까지도 학습하게 되는 방식이다[23]. MLM과 NSP는 BERT 모델이 양방향성(Bidirectional)이라는 특징을 잘 보여주는 기법이며, 이를 통해 각 토큰이 문맥을 고려할 수 있게 된다.



[그림 2-4] Transformer 모델 아키텍처[17]

3. 한국어 BERT 모델

최초 구글이 공개한 Pre-trained BERT 모델인 Multilingual BERT 모델은 104개 언어의 Wikipedia 데이터로 학습을 하였으며, 104개 언어에 한국어도 포함되어 있어 한국어 자연어 처리에도 적용이 가능하다.

그러나 영어로 된 자연어 처리 분야에서는 우수한 성능을 보였으나, 기타 언어에 있어서는 각 언어들이 갖는 특징을 모두 고려하지 못하였기 때문에 영어만큼 우수한 성능을 보여주지는 못하였다[24]. 따라서 이번 연구에서는 Multilingual BERT 모델 대신 SK T-Brain에서 공개한 한국어 BERT 모델인 KoBERT 모델[20]을 이용하였다. KoBERT 모델 사전의 크기는 8,002이고, 한국어 Wikipedia에서 500만 개의 문장과 5000만 개의 단어를 학습하고, 한국어 뉴스에서 2000만 개의 문장과 27,000만 개의 단어를 가지고 학습되었다. 표 2-1은 Multilingual BERT 모델과 KoBERT 모델이 갖는 사전의 구성을 정리[25]한 것으로 사전을 구성하는 전체 토큰 중 단어와 하위단어에서 한글이 차지하는 비중만 보아도 KoBERT 모델이 Multilingual BERT 모델에 비해 더 우수한 성능을 보인다는 것을 알 수 있다.

[표 2-1] Multilingual BERT 모델과 KoBERT모델 사전 구성 비교[25]

구분	Multilingual BERT	KoBERT
단어(한글)	1,664(1.391%)	4,489(56.098%)
하위단어(한글)	1,609(1.346%)	2,678(33.467%)
기호, 기타 언어	116,170(97.175%)	830(10.372%)
special tokens	5(0.004%)	5(0.062%)
tokens	119,547	8,002

제 3 절 주요 북한 도발 사례

앞서 서술했던 바와 같이 북한은 6·25 전쟁 이후 한반도 적화통일 야욕을 버리지 못하고 무수히 많은 전략·전술적 도발을 자행해왔다. 본 연구의 목적은 북한 도발 징후를 포착하는 것으로 이를 위해서는 과거 북한 도발 사례를 선정하고 관련 텍스트데이터를 수집하여 모델을 학습시켜야 한다. 연구에서는 표 2-2에서 보는 바와 같이 주요 북한 도발 사례 총 11건을 선정하였다. 선정 기준은 원활한 데이터 수집을 위해 인터넷 포털 사이트와 온라인 뉴스를 이용하기 시작한 1990년대 후반 이후의 도발 사례, 우리 정부와 군의 대응이 있었으며 언론에서 집중 보도한 도발 사례를 선정하였고, 전략·전술적 도발을 모두 포함하였다. 선정한 주요 북한 사례에 대해 소개하면 다음과 같다. 먼저 제1연평해전은 1999년 6월 15일 대한민국 서해 연평도 부근 해상에서 벌어진 우리 해군과 북한 해군과의 교전으로, 북한의 해군 경비정이 연평도 서쪽 북방한계선(NLL, Northern Limit Line)을 침범하면서 일어났다[26]. 교전 결과 우리 해군은 10척의 북한 함정 중 어뢰정 1척을 침몰시키고 대형 경비정 1척 대파, 중형 경비정 2척 기동 불능 상태로 반파, 소형 경비정 2척을 파손했으며 적 30여 명을 사살하고 70여 명을 부상시키는 큰 전과를 거뒀다. 반면, 아군은 4척의 경비정과 1척의 초계함이 기관실·선체 일부가 파손됐고 9명의 장병이 비교적 경미한 부상을 입는 데 그쳤다. 두 번째 사례인 제2연평해전은 2002년 한·일 월드컵이 진행 중이던 2002년 6월 29일에 발생하였다. 당시 연평도 서남방 해상에서 북한경비정이 북방한계선을 불법 침범하여 이를 저지하기 위해 출동한 우리 해군 함정과 북한 함정 간에 벌어진 해전이다. 북방한계선을 불법 침범한 북한경비정 1척이 먼저 85mm 대구경포 등으로 우리 고속정인 참수리-357정에 기습사격을 가해 옴으로써 우리 함정도 즉각 대응사격을 가해 북한경비정을 격퇴하였다. 교전 결과 우리 해군은 6명(윤영하 대위, 한상국 중사 서후원 하사, 조천형 하사, 황도현 하사, 박동혁 병장)이 전사하고, 고속정 1척이 침몰하는 참극이 일어났다[27]. 고속정 승조원들의 관측에 의하면 북한

경비정도 30여명의 많은 인명피해를 입고 외부에 장착된 각종 화기가 무력화한 것으로 판단되며, 침몰직전에 다른 경비정에 의해 예인되어 북상하였다[28]. 세 번째와 네 번째, 일곱 번째, 아홉 번째부터 열한 번째 사례는 북한의 여섯 번의 핵실험을 선정하였다[29]. 북한은 국제사회의 제제에도 불구하고 핵보유국 지위를 인정받기 위해 여섯 차례나 핵실험을 감행하였고 이는 우리나라 안보 뿐만 아니라 국제사회에도 커다란 위협이 되고 있다. 2017년 9월 3일 북한 함경북도 길주군 풍계리에서 단행한 6차 핵실험은 2016년 9월 9일 5차 핵실험 후 1년 만에 감행된 것이다. 1차 핵실험(2006.10.9)과 2차 핵실험(2009.5.25)이 2년 7개월, 2차 핵실험과 3차 핵실험(2013.2.12)이 3년 8개월, 3차 핵실험과 4차 핵실험(2016.1.6)이 2년 11개월 간격으로 3년 안팎으로 발생한 것에 비해 4차 핵실험과 5차 핵실험, 5차 핵실험과 6차 핵실험은 각각 1년도 되지 않은 간격으로 발생한 것으로 비추어 볼 때 핵 개발 기술이 상당히 진전을 이룬 것으로 보인다. 특히, 6차 핵실험의 위력은 약 50kt으로 5차 핵실험에 비해 약 5~6배로 추정된다. 다음으로 다섯 번째 선정 사례인 천안함(PCC-772) 피격 사건은 2010년 3월 26일에 발생하였다[30]. 당시 서해 북방한계선 근해에서 감시 및 대응태세를 유지하며 정상적인 해상경비와 어로활동 지원 임무를 수행 중이던 천안함은 백령도 연화리 서남방 2.5km 해상에서 피격되어 완전히 침몰되었으며, 승조원 104명 중 46명이 사망, 실종 장병 구조작전 중 숨진 한주호 준위 등 총 47명의 전우가 목숨을 잃었다. 국방부는 국내·외 전문가로 구성된 민·군 합동조사단을 편성해 침몰 원인을 조사한 결과, 북한에서 제조·사용 중인 어뢰에 의해 피격된 것으로 확인되었다. 여섯 번째 선정 사례는 2010년 11월 23일 발생한 연평도 포격도발 사건이다[31]. 당시 북한이 서해 연평도에 해안포와 방사포로 무차별적 포격을 가하였고, 이에 우리 해병대 연평부대가 K-9 자주포로 대응사격을 실시한 쌍방 포격전이다. 북한의 포격으로 우리 해병 2명이 전사하고, 민간이 2명이 사망하였으며, 수십 채의 가옥과 산림이 불타는 피해를 입었다. 북한의 연평도 포격은 1953년 정전협정 체결 이후 처음으로 우리 영토에 대해 직접 포격을 가한 무력도발이며, 무방비 상태인

민간인 거주지역에 까지 무차별적인 포격을 가한 비인도적 만행일 뿐만 아니라, 유엔헌장・정전협정・남북불가침합의를 정면으로 위반한 불법적 도발행위이다. 여덟 번째 도발 사례인 DMZ 목함지뢰 도발은 2015년 8월 4일 경기도 파주시 인근 DMZ 남측 GP 추진철책 통문하단에 북한측이 매설한 목함지뢰가 폭발함으로써 우리군 하사 2명이 다리가 절단되는 심각한 부상을 당한 사건이다[32].

[표 2-2] 주요 북한 도발 사례 선정

구분	주요 북한 도발 사례	발생일
1	제1연평해전	1999.06.15.
2	제2연평해전	2002.06.29.
3	북한 1차 핵실험	2006.10.09.
4	북한 2차 핵실험	2009.05.25.
5	천안함(PCC-772) 피격	2010.03.26.
6	연평도 포격	2010.11.23.
7	북한 3차 핵실험	2013.02.12.
8	DMZ 목함지뢰 매설	2015.08.04.
9	북한 4차 핵실험	2016.01.06.
10	북한 5차 핵실험	2016.09.09.
11	북한 6차 핵실험	2017.09.03.

제 3 장 실험 설계 및 데이터

제 1 절 실험 모델

문제 발생의 징후를 가장 잘 설명하는 통계학적 접근으로 1:29:300의 법칙이라고 불리는 하인리히의 법칙(Heinrich's law)이 있다[33]. 이는 어떤 대형사고가 발생하기 전에는 그와 관련된 수십 차례의 경미한 사고와 수백 번의 징후들이 반드시 나타난다는 것을 뜻한다. 북한 도발도 사건이 발생하기 전에 우리가 인지하지 못하거나 간과하는 몇 가지 징후들이 있을 것이고, 도발 발생시점이 다가오면 더 명확하고 뚜렷한 징후가 빈번하게 나타날 수 있다는 것을 전제로 모델을 구상하였다. Pre-trained BERT 모델은 한국어 자연어 처리에서 좋은 성능을 보이는 KoBERT 모델을 기반으로 하였으며, 북한 도발의 사전 징후들을 포착하기 위해 앞서 선정한 각 도발 발생일 직전 한 달간의 데이터를 수집하였으며, 비교・대조를 위하여 도발이 발생하지 않은 평시 한 달간의 데이터 또한 수집하여 이진분류(Binary classification) 하도록 모델을 설계하였다. 다시 말해, 수집한 데이터 중 도발 발생일 직전 한 달간의 데이터는 도발 징후가 있다고 label을 분류하는 것이고, 그렇지 않은 평시의 데이터는 도발 징후가 없다고 label을 분류하는 모델을 학습시키는 것이다. 최종적으로 모델이 학습을 통해 데이터에 내포된 의미적 관계와 특징적 패턴을 바탕으로 새로운 데이터를 넣었을 때 도발 징후가 있는지 없는지를 결과로 제시하는 모델을 제작하였다. 설계된 BERT 모델은 Fine-tuning시 Adam optimizer[34]를 사용했으며, 학습 횟수는 2, 3, 4 epoch, batch size는 16과 32를 적용하여 정확도를 비교해본 후 가장 높은 정확도를 보인 epoch과 batch size를 대상으로 learning rate를 2e-5, 3e-5, 4e-5, 5e-5로 바꿔가며 최적화된 모델을 제시하였다. 실험 환경은 Intel Core i7-8750H CPU 2.20 GHz, 8 GB RAM, GeForce RTX 2080 Ti이다.

제 2 절 실험데이터 수집

데이터는 보안상의 이유로 군사비밀을 사용하지 않고, 접근이 용이한 인터넷 공개 자료인 언론사의 언론보도와 통일부에서 제공하는 북한동향 자료를 사용하였다. 언론사의 언론보도는 '빅카인즈(BIG KINDS, www.bigkinds.or.kr)'라는 인터넷 뉴스 검색 서비스를 사용하여 자료를 수집하였으며, 통일부의 북한동향 자료는 '통일부 북한정치포털(nkinfo.unikorea.go.kr)'에 게시된 주간북한동향 자료를 다운로드 하여 사용하였다. 이중 통일부에서 제공하는 북한동향 자료는 '노동신문, 평양신문, 민주조선, 조선중앙방송, 평양방송' 등 북한 일간지와 방송의 주요 보도내용을 요약하여 제공하고 있어 북한 내부 정세를 알 수 있는 자료로 연구에 유의미한 영향을 미칠 것이라 판단되어 사용하였다. 도발 징후 포착 관련 자료는 각 도발 발생일 기준 1개월 전부터 도발발생 하루 전날까지의 언론보도 자료와 통일부 북한동향 자료를 문장 단위로 수집하였으며 총 개수는 18,422개이다. 비교학습을 위해 도발이 발생하지 않았던 기간의 자료는 각 도발 발생일 사이의 평시 기간을 1개월 단위로 선정하여 자료를 문장 단위로 수집하였고 총 개수는 10,338개이다. 수집한 총 데이터는 28,760개이다. 딥러닝 학습을 위해서는 수집한 데이터를 학습 데이터(training data set)와 평가 데이터(test data set)로 구분하여야 한다. 이때의 비중은 학습 데이터 70%, 평가 데이터 30%로 하는 것이 일반적이다. 수집한 데이터를 학습 데이터와 평가 데이터로 구분하는 이유는 수집한 데이터의 100%를 학습 데이터로 사용한 후 이중 일부를 다시 평가 데이터로 활용한다면 이미 학습한 자료를 가지고 평가를 하게 되므로 예측률이 100%가 되어 평가의 의미가 사라지기 때문이다. 표 3-1에서 보는 바와 같이 이번 실험에서도 학습 데이터와 평가 데이터의 비율을 7:3으로 무작위로 나누어 연구를 진행하였다. 세부적으로 표 3-2와 표 3-3은 무작위로 분류한 학습 데이터와 평가 데이터에서 도발 징후 포착기 데이터와 평시 데이터가 차지하는 비율을 각각 나타낸 것이다.

[표 3-1] 총 데이터 중 학습 및 평가 데이터 비율

구분	학습 데이터	평가 데이터	총 데이터
개수	20,171(70%)	8,589(30%)	28,760(100%)

[표 3-2] 학습 데이터 중 도발 징후 포착기 및 평시 데이터 비율

구분	도발 징후 포착기	평시	계
학습 데이터 개수	12,975(64%)	7,196(36%)	20,171(100%)

[표 3-3] 평가 데이터 중 도발 징후 포착기 및 평시 데이터 비율

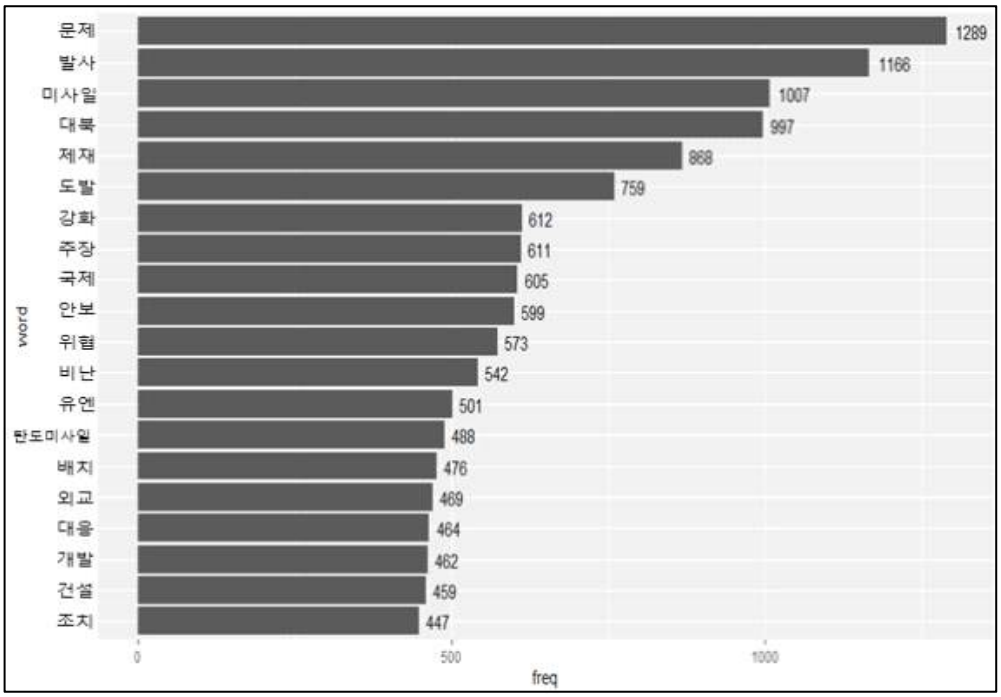
구분	도발 징후 포착기	평시	계
평가 데이터 개수	5,447(63%)	3,142(37%)	8,589(100%)

제 4 장 실험결과

제 1 절 실험데이터 탐색적 분석

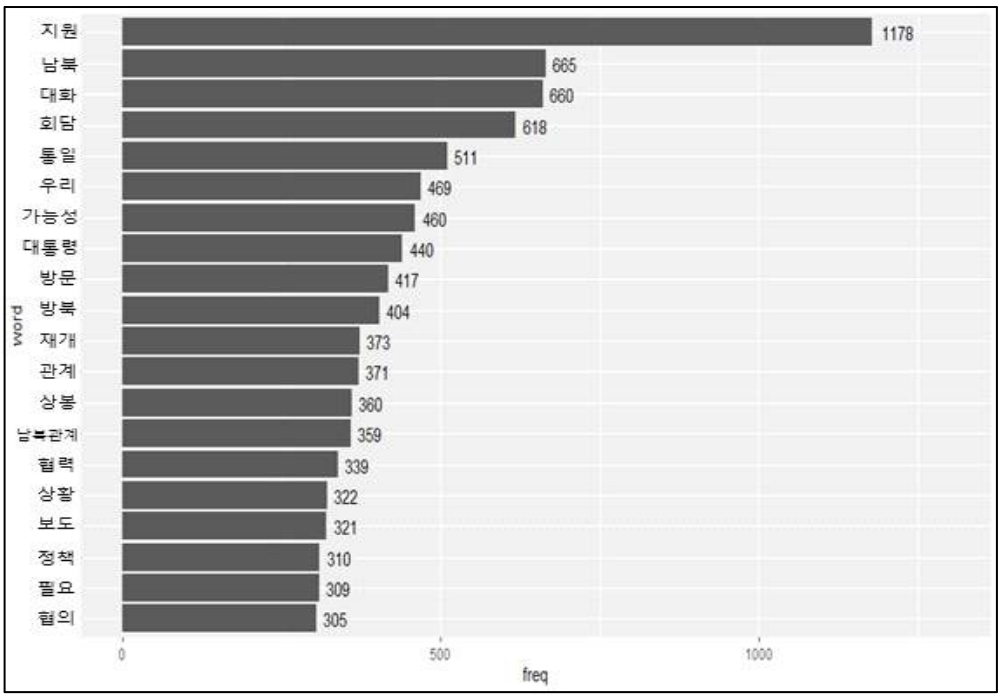
먼저 수집한 데이터가 유의미한 특징을 가지고 있는지 확인하기 위해 전반적 분석을 진행하였다. 데이터를 구성하는 상위 빈도수의 단어(Keyword)들로 워드 크라우드(Word cloud)를 구현해 보았고, 상위 20개 단어들의 출현 횟수를 빈도막대그래프로 나타내어 보았다. 그림 4-1과 그림 4-2는 도발 징후와 관련된 데이터를 분석한 결과로 '문제', '발사', '미사일', '제재', '도발', '위협', '비난' 등의 부정적이고 공격적인 단어가 주로 나타났다. 반면, 그림 4-3과 그림 4-4는 도발 징후가 없다고 분류한 평시 데이터를 분석한 결과로 '지원', '대화', '회담', '통일', '방북', '협력', '협의' 등의 긍정적이고 호의적인 단어가 상위 빈도 단어로 출현하였다. 이를 통해 연구자가 수집한 도발 징후 관련 데이터와 도발이 발생하지 않았던 평시 기간에 수집한 데이터 간에 비교가능한 차이가 존재하며, 이러한 차이는 실험의 결과에도 유의미한 영향을 미칠 수 있다는 것을 유추해볼 수 있다. 해당 실험에는 'R'이라는 통계 분석 및 시각화를 위한 프로그래밍 언어를 사용하였다.

[그림 4-1] 도발 징후 포착기 데이터 워드 크라우드



[그림 4-2] 도발 징후 포착기 데이터 상위 20개 단어 빈도막대그래프

[그림 4-3] 평시 데이터 워드 크라우드



[그림 4-4] 평시 데이터 상위 20개 단어 빈도막대그래프

제 2 절 학습 및 평가 단계 결과

본격적인 KoBERT 모델 적용을 위해 딥러닝 구현에 가장 많이 쓰이는 파이썬(python) 프로그래밍 언어를 사용하여 연구를 진행하였다. 그림 4-5와 같이 연구자의 학습 데이터에서 파이썬 명령어를 통해 상위 10개만 출력해 보았다. id는 데이터의 연번, documents는 데이터의 내용, label은 도발 징후 관련 데이터에는 0("도발 징후 있음"을 의미), 평시 관련 데이터에는 1("도발 징후 없음"을 의미)로 표기하여 분류되도록 하였다. 학습 데이터에 대한 학습 횟수인 epoch 값은 2, 3, 4로 설정하였고 메모리 과부하를 방지하기 위해 데이터를 나누어 학습시키는 것을 의미하는 batch size는 16, 32로 설정하였으며, 학습률을 의미하는 learning rate는 2e-5, 3e-5, 4e-5, 5e-5로 설정하여 실험을 진행하였다. 실험은 우선 epoch값과 batch size를 바꿔가며 가장 높은 정확도를 보이는 모델을 선정한 이후 learning rate 값을 변경해가며 정확도를 비교하였다. epoch과 batch size, learning rate는 정해진 값이 있는 것이 아니라, 연구자의 반복적인 실험을 통해 조정해가며 최적값을 찾아내야하는 Hyper parameter이다. 실험 결과 표 4-1, 표 4-2와 같이 epoch은 4, batch size는 16, learning rate는 2e-5로 했을 때 가장 높은 정확도를 보였으며, 학습 단계에서의 정확도는 88%, 평가 단계에서의 정확도는 83% 이었다.

	id	documents	label
0	1.0	북한 지하시설의 의혹 해소가 북미간 시급한 과제며 미국은 이 문제 해결에 비중을 두...	1
1	2.0	로스 차관보는 이날 워싱턴 포린 프레스 센터에서 가진 회견에서 북한의 핵무기정책에...	1
2	3.0	그는 한반도 평화정착을 위한 4자회담은 단기간내 돌파구를 마련할 수 있는 사안이 ...	1
3	4.0	로스 차관보는 한반도문제는 지하핵시설,미국 의회의 한반도에너지기구(KEDO) 예산...	1
4	5.0	그는 북한이 최근 북·미 핵협정의 준수 의도가 없다고 밝힌것은 위협용이라며 과도한...	1
5	6.0	13일로 1주일째를 맞은 북한의 서해 영해 침범 사태는 북한이 이 문제를 논의하기 ...	0
6	7.0	그러나 북한이 침범행위를 언제쯤에나 중단할지는 불투명하다	0
7	8.0	북한은 장성급회담 제의를 수용한 13일에도 어선 및 경비정을 북방한계선 (NLL) ...	0
8	9.0	북한의 침범행위에는 특정 또는 다양한 노림수가 있는 것만은 분명하다	0
9	10.0	나름대로 정한 목표치를 달성해야 도발행위를 그만둘 것이라는 추론도 가능하다	0

[그림 4-5] 학습 데이터 10개 출력

[표 4-1] epoch, batch size 조정 결과 비교(학습 데이터 / 평가 데이터)

구분	epoch 2	epoch 3	epoch 4
batch size 16	86% / 82%	87% / 83%	<u>88% / 83%</u>
batch size 32	84% / 81%	85% / 81%	86% / 81%

[표 4-2] epoch, batch size 조정 후 learning rate 조정 결과 비교(학습 데이터 / 평가 데이터)

구분	lr 2e-5	lr 3e-5	lr 4e-5	lr 5e-5
epoch 4 batch size 16	<u>88% / 83%</u>	86% / 82%	87% / 82%	87% / 83%

제 3 절 새로운 문장 입력 결과

이번에는 연구자가 수집한 데이터 외에 새로운 임의의 문장을 모델에 입력한 후 해당 문장이 도발 징후가 있는지(0), 없는지(1) 결과로 제시하는 실험을 진행해 보았다. 이번 실험에 적용해본 문장은 그림 4-6과 그림 4-7과 같이 총 20개이다. 이 20개 문장 중 ①~⑩번까지의 문장은 도발 관련 데이터인 지난 6월 16일 발생한 북한의 남북공동연락사무소 폭파 사건을 기준으로 발생일 이전 1개월 동안 보도된 북한 관련 언론보도 기사 중 일부 문장을 발췌한 것으로, 연구자가 설계한 모델대로 라면 모두 도발 징후가 있는 것으로 분류되어 결과값이 '0'으로 출력되면 올바르게 예측한 것이다. 한편, ⑪~⑳번까지의 문장은 우리나라와 북한이 화해, 협력의 국면으로 진입하였던 2018년 4월 27일 남북 정상회담을 기준으로 이전 1개월 동안 보도된 북한 관련 언론보도에서 문장을 발췌하여 입력한 것으로, 도발 징후가 없는 것으로 분류되어 결과값이 '1'로 출력되면 모델이 잘 작동하는 것이다. 실험 결과, ①~⑩번까지의 문장은 ⑧번 문장인 "특히 과거에도 북한이 대북 전단을 문제 삼은 적은 있지만 이번엔 대내적으로까지 공론화하고 있다는 점을 주목할 만하다."를 제외한 9개의 문장 모두 '0'으로 출력되었다. 주목할 만한 점은 ⑤번 문장인 "북한은 올해 들어 대중 수출입 규모가 전년 대비 90%나 급감해 북한 경제의 핏줄인 장마당이 크게 위축된 가운데 외화보유액도 고갈 조짐이다."인데 북한의 악화된 경제사정이 도발로 이어질 수도 있음을 예측한 의미 있는 결과라고 볼 수 있다. 또한, ⑩번 문장에서 남북 공동연락 사무소 폐쇄를 언급하였는데 실제 폭파로 이루어진 것은 주목할 만하다. ⑪~⑳번까지의 문장은 ⑯번 문장인 "김정은 북한 국무위원장이 트럼프 대통령과 한반도의 비핵화 문제를 기꺼이 논의할 의향이 있다는 것을 재확인했다."와 ⑱번 문장인 "한미 양국 군이 역사적인 남북정상회담을 하루 앞둔 26일 대규모 연례 군사훈련인 독수리(FE)연습을 사실상 종료했다."를 제외한 8개의 문장이 모두 '1'로 출력되었다. 이를 통해, 모델의 정확도가 평가 단계에서 보인 83% 이상의 정확도를 갖는 것이 입증되었다.

```
logits = test_sentences(['남조선 당국의 은폐된 적대시 정책과 무맥무능한 처사로 하여 완전히 풍비박산 나고 최악의 긴장 상태가 조성된 것이
print(logits)
print(np.argmax(logits))
[[ 3.788601  -4.0795274]]
0
```

① 남조선 당국의 은폐된 적대시 정책과 무맥무능한 처사로 하여 완전히 풍비박산 나고 최악의 긴장 상태가 조성된 것이 오늘의 북남관계 ⇒ "0"

```
logits = test_sentences(['북한은 대남 비난에 욕설을 퍼붓고 있다.'])
print(logits)
print(np.argmax(logits))
[[ 2.209704 -2.261375]]
0
```

② 북한은 대남 비난에 욕설을 퍼붓고 있다. ⇒ "0"

```
logits = test_sentences(['이 거세찬 분노를 반영하여 세운 보복 계획들은 우리의 국론으로 확고히 굳어졌다'])
print(logits)
print(np.argmax(logits))
[[ 3.6179557 -3.8615432]]
0
```

③ 이 거세찬 분노를 반영하여 세운 보복 계획들은 우리의 국론으로 확고히 굳어졌다. ⇒ "0"

```
logits = test_sentences([' 2018년 4월 20일 개통된 남북 정상 간 핫라인도 한 번도 가동되지 않은 채 781일 만에 끊기게 됐다.'])
print(logits)
print(np.argmax(logits))
[[ 3.61523  -3.8144197]]
0
```

④ 2018년 4월 20일 개통된 남북 정상 간 핫라인도 한 번도 가동되지 않은 채 781일 만에 끊기게 됐다. ⇒ "0"

```
logits = test_sentences(['북한은 올해 들어 대중 수출입 규모가 전년 대비 90%나 급감해 북한 경제의 핏줄인 장마당이 크게 위축된 가운데 외
print(logits)
print(np.argmax(logits))
[[ 3.0408542 -3.2215393]]
0
```

⑤ 북한은 올해 들어 대중 수출입 규모가 전년 대비 90%나 급감해 북한 경제의 핏줄인 장마당이 크게 위축된 가운데 외화보유액도 고갈 조짐이다. ⇒ "0"

```
logits = test_sentences(['무엇보다 북한이 도발하면 국제사회의 대북 제재망은 더욱 촘촘하고 두터워질 뿐이다.'])
print(logits)
print(np.argmax(logits))
[[ 3.5246727 -3.7914362]]
0
```

⑥ 무엇보다 북한이 도발하면 국제사회의 대북 제재망은 더욱 촘촘하고 두터워질 뿐이다. ⇒ "0"

```
logits = test_sentences(['무적의 혁명강군은 격앙될 대로 격앙된 우리 인민의 원한을 풀어줄 단호한 행동을 개시할 것'])
print(logits)
print(np.argmax(logits))
[[ 3.6608324 -3.8922133]]
0
```

⑦ 무적의 혁명강군은 격앙될 대로 격앙된 우리 인민의 원한을 풀어줄 단호한 행동을 개시할 것' ⇒ "0"

```
logits = test_sentences(['특히 과거에도 북한이 대북 전단을 문제 삼은 적은 있지만 이번엔 대내적으로까지 공론화하고 있다는 점을 주목할 만
print(logits)
print(np.argmax(logits))
[[-3.2287717  3.3786168]]
1
```

⑧ *특히 과거에도 북한이 대북 전단을 문제 삼은 적은 있지만 이번엔 대내적으로까지 공론화하고 있다는 점을 주목할 만하다.* ⇒ *"1"*

```
logits = test_sentences(['지난해 919 남북정상회담 당시 문 대통령의 평양 연설까지 거론하며 남측 정부를 비판한 것이다.'])
print(logits)
print(np.argmax(logits))
[[ 2.8884406 -3.0924592]]
0
```

⑨ 지난해 919 남북정상회담 당시 문 대통령의 평양 연설까지 거론하며 남측 정부를 비판한 것이다. ⇒ "0"

```
logits = test_sentences(['북한 당 통일전선부가 남북 공동연락 사무소를 폐쇄하겠다고 엄포를 놓은 데 이어 달나라 타령을 하고 있다며 문재인
print(logits)
print(np.argmax(logits))
[[ 3.8033776 -4.1189075]]
0
```

⑩ 북한 당 통일전선부가 남북 공동연락 사무소를 폐쇄하겠다고 엄포를 놓은 데 이어 달나라 타령을 하고 있다며 문재인 대통령을 직접 겨냥했다 ⇒ "0"

[그림 4-6] 도발 징후 포착기 데이터 새로운 문장 입력 테스트

```
logits = test_sentences(['남북정상회담을 1주일 앞두고 화해 협력 분위기가 무르익고 있다.'])

print(logits)
print(np.argmax(logits))
```

```
[[-3.2518148  3.4301527]]
1
```

⑪ 남북정상회담을 1주일 앞두고 화해 협력 분위기가 무르익고 있다. ⇒ “1”

```
logits = test_sentences(['문재인 대통령과 김정은 국무위원장의 정상회담에서 핵심 의제는 한반도 비핵화와 항구적 평화정착으로 집약된다.']

print(logits)
print(np.argmax(logits))
```

```
[[-3.3706467  3.5857513]]
1
```

⑫ 문재인 대통령과 김정은 국무위원장의 정상회담에서 핵심 의제는 한반도 비핵화와 항구적 평화정착으로 집약된다. ⇒ “1”

```
logits = test_sentences(['인천시는 올해 고려 건국 1100주년을 맞아 남북 역사학자들이 참여하는 국제 학술회의 개최를 추진할 예정이다.'])

print(logits)
print(np.argmax(logits))
```

```
[[-1.6925135  1.7746643]]
1
```

⑬ 인천시는 올해 고려 건국 1100주년을 맞아 남북 역사학자들이 참여하는 국제 학술회의 개최를 추진할 예정이다. ⇒ “1”

```
logits = test_sentences(['탈북민들도 남북정상회담의 성공적 개최를 응원하면서 이번 회담을 계기로 남북 간에 왕래나 경제적 교류가 성사되기

print(logits)
print(np.argmax(logits))
```

```
[[-2.218479   2.5535994]]
1
```

⑭ 탈북민들도 남북정상회담의 성공적 개최를 응원하면서 이번 회담을 계기로 남북 간에 왕래나 경제적 교류가 성사되기를 기대했다. ⇒ “1”

```
logits = test_sentences(['오는 27일 남북 정상회담에서 평화선언을 발표하는 것을 검토하고 있다'])

print(logits)
print(np.argmax(logits))
```

```
[[-2.7457993  3.0402436]]
1
```

⑮ 오는 27일 남북 정상회담에서 평화선언을 발표하는 것을 검토하고 있다. ⇒ “1”

```
logits = test_sentences(['김정은 북한 국무위원장이 트럼프 대통령과 한반도의 비핵화 문제를 기꺼이 논의할 의향이 있다는 것을 재확인했다'

print(logits)
print(np.argmax(logits))
```

```
[[ 3.820104 -4.127878]]
0
```

⑯ *김정은 북한 국무위원장이 트럼프 대통령과 한반도의 비핵화 문제를 기꺼이 논의할 의향이 있다는 것을 재확인했다. ⇒ “0”*

```
logits = test_sentences(['방북길에 오른 남측 예술단과 태권도시범단이 31일 평양에 도착해 4월 1~3일 열릴 공연 준비에 들어갔다.'])

print(logits)
print(np.argmax(logits))
```

```
[[-0.9273928  1.107531 ]]
1
```

⑰ 방북길에 오른 남측 예술단과 태권도시범단이 31일 평양에 도착해 4월 1~3일 열릴 공연 준비에 들어갔다. ⇒ “1”

```
logits = test_sentences(['한미 양국 군이 역사적인 남북정상회담을 하루 앞둔 26일 대규모 연례 군사훈련인 독수리(FE)연습을 사실상 종료했다

print(logits)
print(np.argmax(logits))
```

```
[[ 2.7832527 -2.945438 ]]
0
```

⑱ *한미 양국 군이 역사적인 남북정상회담을 하루 앞둔 26일 대규모 연례 군사훈련인 독수리(FE)연습을 사실상 종료했다. ⇒ “0”*

```
logits = test_sentences(['정부는 이번 공연을 계기로 남북 문화·체육 교류의 폭과 깊이를 넓혀갈 방침을 세워두고 있다.'])

print(logits)
print(np.argmax(logits))
```

```
[[-2.1382194  2.567173 ]]
1
```

⑲ 정부는 이번 공연을 계기로 남북 문화·체육 교류의 폭과 깊이를 넓혀갈 방침을 세워두고 있다. ⇒ “1”

```
logits = test_sentences(['우리 한번 통일시켜서 전세계적으로 대한민국의 위상을 떨쳐봅시다.'])

print(logits)
print(np.argmax(logits))
```

```
[[-1.9785242  2.2845485]]
1
```

⑳ 우리 한번 통일시켜서 전세계적으로 대한민국의 위상을 떨쳐봅시다 ⇒ “1”

[그림 4-7] 평시 데이터 새로운 문장 입력 테스트

제 5 장 결 론

본 연구에서는 한국어 BERT 모델을 기반으로 북한 도발과 관련된 텍스트데이터를 수집 후 수집한 텍스트데이터의 숨겨진 문맥적 관계와 패턴을 학습시켜 새로운 문장을 대입했을 때 해당 문장이 도발 징후가 있는지 없는지를 판단해보는 실험을 진행해보았다. 이를 위해 Multilingual BERT 모델보다 한국어 자연어 처리에서 성능이 우수한 KoBERT 모델을 사용하였고, epoch과 batch size, learning rate 등의 Hyperparameter를 조정하는 Fine-tuning을 거쳐 최적의 정확도를 갖는 모델을 구현한 이후 최근 도발 사례인 남북 공동연락사무소 폭파와 관련된 텍스트데이터들로 모델을 검증해보았다. 이번 연구는 북한 도발 징후 포착을 기존의 정보작전 분야 감시장비나 군사 전문가의 경험적 판단, 각종 첩보망 외에도 4차 산업혁명 시대에 부합한 최신 과학·공학기술을 통해 예측해보는 실험을 진행하였다는 것에 그 의의가 있다. 군사비밀 수집의 제한으로 북한 관련 언론보도와 정부부처의 공개자료를 사용하여 실험을 하였기 때문에 군사적 근거가 부족한 것은 이번 연구의 한계점이라고 판단된다. 그리고 이 연구를 바로 현행작전에 적용하기에는 무리가 있으며, 향후 철저한 검증 과정을 거쳐야 함은 당연한 사실이다. 이미 미국, 중국 등 AI 선진국에서는 이러한 최신 과학기술을 인간의 삶에 적용하여 사회안전망의 질을 개선하고 있으며, 각종 범죄 예측과 예방에 활용하고 있다. 국방 분야에서도 마찬가지이다. 우리나라는 세계 유일의 분단국가이며, 북한과 정전협정을 맺고 현재 휴전 중인 상태이다. 전쟁은 끝나지 않았다. 언제 다시 북한이 기습적인 도발 행위를 자행할지 모르는 일이다. 이에 대한 철저한 대비가 필요하다. 과학기술이 앞선 나라가 세계 경제와 산업을 선도하듯 국방 과학기술이 앞선 나라가 더욱 강력한 국방력을 갖출 것이다. 추후 연구과제로는 본 논문의 한계점을 보완하기 위해 군사보안에 위반되지 않는 범위 내에서 대중에게 공개된 군사자료

를 추가적으로 수집하여 Fine-tuning 단계에서 활용하여 학습 데이터의 수를 더 확보하고 모델을 학습시키고자 한다. 또한, 수집한 텍스트를 모델을 통해 분석하여 도발 징후 유무에 대한 판단뿐만 아니라 도발 발생 확률은 얼마나 되는지 확률(%)로 제시해주고 어떤 도발유형이 발생할 것인지 까지 예측하는 모델을 개발하고자 한다. 이를 통해, 실제 북한 도발 징후를 포착하는데 더욱 신뢰도 높고 정교한 모델을 얻을 수 있고, 작전수행 간 더 효과적인 상황판단-결심-대응 체계 구축에 기여할 수 있을 것이라 생각한다.

참 고 문 헌

[1] 김민욱, 박수찬, 군·업체 상생하는 방산진흥 정책 시급하다, 국방과 기술, Vol.406, (2012). p.8-11.

[2] Lee, K.J., Kim, Y.T., and Jun, Y.K., “The Status of Smart National Defense Innovation Based on Technology of the 4th Industrial Revolution”, The 2020 spring conference of Korea Society of IT services, (2020), p.353-356.

[3] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton., “Deep learning.”, Nature, Vol.521, (2015), p.436-444.

[4] F. Ronen and I. Dagan, “Knowledge Discovery in Textual Database”, Proceedings of the First International Conference on Knowledge Discovery and Data Mining, Vol.95, (1995), p.112-117.

[5] Devlin, J., Chang, M. W., and Lee, K. T.,“BERT: Pre-training of deep bidirectional transformers for language understanding,”, arXiv preprint arXiv:1810.04805, (2018).

[6] Hwang, S.H., and Kim, D.H., “BERT-based Classification Model for Korean Documents”, The Journal of Society for e-Business Studies, Vol.25, (2020), p.203-214.

[7] Jang, E.A., Choi, H.R., and Lee, H.C., “Stock prediction using combination of BERT sentiment Analysis and Macro economy index”, Journal of The Korea Society of Computer and Information, Vol.25, (2020), p.47-56.

[8] Park, C.E., and Lee, C.K., “Sentimental Analysis of Korean Movie Review using Variational Inference and RNN based on BERT”, KIISE Transactions on Computing Practices, Vol.25,

(2019), p.552-558.

[9] Kao, A., and Poteet, S. R., Natural language processing and text mining, Springer Science & Business Media. London (2007).

[10] Kang, H.S., and Yang, J.H., “Performance Comparison of Word2 vec and fastText Embedding Models”, Journal of Digital Conten ts Society, Vol.21, (2020), p.1335-1343.

[11] Choi, D.J., Yun. Y.S., Yun. S.J., and Oh, Y.H., “An Efficient Ski p N - gram Language Model For Keyword Spotting”, The Kore an Institute of Information Scientists and Engineers, Vol.25, (199 8), p.279-281.

[12] Hwang, S.S., “신경망 기반 언어모델의 연구 동향”, The Journal o f The Korean Institute of Communication Sciences, Vol.35, (201 8), p.22-28.

[13] Jung, Y.M., “정보검색에서의 언어모델 적용에 관한 분석”, Korean Library And Information Science Society, (2005), p.37-57.

[14] Gyu-Hyung Lee, G.H., Lee, Y.D., and Koo, I.S., “An RNN-base d Fault Detection Scheme for Digital Sensor”, The Journal of T he Institute of Internet, Broadcasting and Communication, Vol.1 9, (2019), p.29-35.

[15] Kim, J.M., and Lee, J.H., “Text Document Classification Based o n Recurrent Neural Network Using Word2vec”, Journal of Korea n Institute of Intelligent Systems, Vol.27, (2017), p.560-565.

[16] CHO, H.S., “Visual Question Answering Using Simple Natural L anguage Processing”, Seoul National University, (2018), p.1-52.

[17] A. Vaswani, et al., "Attention Is All You Need", Proc. of 31st Co nference on Neural Information Processing Systems (NIPS), (2017)

[18] Choi, K.H., and Lee, C.K., “Sentiment analysis with GRU-Attent ion”, The Korean Institute of Information Scientists and Enginee rs, (2015), p.557-559.

[19] Choi, Y.S., and Lee, K.J., “Performance Analysis of Korean Mor phological Analyzer based on Transformer and BERT”, The Kor ean Institute of Information Scientists and Engineers, Vol.47, (20 20), p.730-741.

[20] https://github.com/SKTBrain/KoBERT.

[21] Shin, J.H., Kim, Y.K., and Lee, J.H., "Transformer-based Auto matic Post-Editing for Machine Translation", KOREA INFORM ATION SCIENCE SOCIETY, Vol.25, (2019), p.64-69.

[22] Park, C.E., Lee, C.K., Lim, J.H., and Kim, H.K., “Korean Depend ency Parsing with BERT”, The Korean Institute of Information Scientists and Engineers, (2019), p.530-532.

[23] Lim, J.H., Kim, H.K., and Kim, Y.K., “Recent R&D Trends for Pretrained Language Model”, Electronics and Telecommunication s Research Institute, Vol.35, (2020), p.9-19.

[24] Yoo, S.Y., and Jeong, O.R., “Korean Contextual Information Extr action System using BERT and Knowledge Graph”, Korean Soci ety for Internet Information, Vol.21, (2020), p.123-131.

[25] Lee, S.A., Jang, H.S., Baik, Y.M., Park, S.J., and Shin, H.P., “A Small-Scale Korean-Specific BERT Language Model”, Journal o

f KIISE, Vol.47, (2020), p.687.

[26] 김영이,“안보100대사건:<47>연평해전”, 국방일보, 2006년 12월 21일자.

[27] 국방부 정훈공보관실, “서해교전과 6·25전쟁의 교훈”, 국방일보, 2003년 6월 20일자.

[28] “북한군의 NLL 침범 및 도발과 우리 군의 대응”, 국방일보, 2002년 7월 2일자, 4면.

[29] Kim, K.N., “The Prospects for North Korean Military Policy Changes after the Sixth Nuclear Test and the ROKs Response”, Korean Association Of Unification Strategy, Vol.9-53, (2017), p.45.

[30] 대한민국 정부, 천안함 피격사건 백서, 인쇄의 창, 서울(2011), p.38-146

[31] 대한민국 국방부, 2010 국방백서, 대한민국 국방부, 서울(2010), p.266-267.

[32] 대한민국 국방부, 2016 국방백서, 대한민국 국방부, 서울(2016), p.213-218

[33] Park, W.S., “Disaster Politics in Korea : Why We Can't Fix the Stable Even After the Horse is Stolen”, The Korea Association for Public Management, Vol.28, (2014), p.25-54.

[34] Kingma, D. and Ba, J., “Adam: A method for stochastic optimization”, Proceedings of the 3rd International Conference on Learning Representations, (2015).

Abstract

BERT-Based Catching Signs of Provocation of North Korea Using Language Modeling

KIM Donghun

Department of Engineering Practice

Graduate School of Engineering Practice

Seoul National University

After the Korean War, North Korea continued to perpetrate in strategic and tactical provocations that threatened the security of the Republic of Korea and international community. As a result, countless soldiers of the Korean Armed Forces were sacrificed, and economic losses such as withdrawal of foreign investment capital and damage recovery costs were also caused. Accordingly, the Republic of Korean armed forces is operating various intelligence surveillance assets to protect the security of the Republic of Korea and the lives and assets of the people from North Korean provocation threats, but it is true that there are some limitations in predicting and responding to all threats. Therefore, in this study, to predict North Korean provocation in line with Digital Transformation, we proposed a model that captures North Korean provocation symptoms from text data related to North Korean provocation by using the latest deep learning-based

natural language processing model, BERT model. First, 11 cases of North Korean provocation were selected. Data for 1 month prior to each provocation occurrence date and peacetime data in which no provocation occurred were compared and learned. Through Fine tuning, the accuracy of learning and test stage was optimized, and the performance of the model was verified by substituting a new random sentence into the model. Ultimately, in addition to the use of intelligence surveillance assets, it is intended to contribute to minimizing damage to allies and contributing to victory in battles by enhancing the surveillance capability of the Republic of Korea armed forces by using the latest science and technology to detect provocation signs.

keywords : BERT, Deep Learning, Text mining, Language model, NLP, Provocation of North Korea
Student Number : 2019-28970