JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 10, pp. 157-165, October 2021
https://doi.org/10.9708/jksci.2021.26.10.157

Deep Learning-based Target Masking Scheme for Understanding Meaning of Newly Coined Words

Gun-Min Nam*, Namgyu Kim*

*Graduate Student, Graduate School of Business IT, Kookmin University, Seoul, Korea
*Professor, Graduate School of Business IT, Kookmin University, Seoul, Korea

[Abstract]

Recently, studies using deep learning to analyze a large amount of text are being actively conducted. In particular, a pre-trained language model that applies the learning results of a large amount of text to the analysis of a specific domain text is attracting attention. Among various pre-trained language models, BERT(Bidirectional Encoder Representations from Transformers)-based model is the most widely used. Recently, research to improve the performance of analysis is being conducted through further pre-training using BERT's MLM(Masked Language Model). However, the traditional MLM has difficulties in clearly understands the meaning of sentences containing new words such as newly coined words. Therefore, in this study, we newly propose NTM(Newly coined words Target Masking), which performs masking only on new words. As a result of analyzing about 700,000 movie reviews of portal 'N' by applying the proposed methodology, it was confirmed that the proposed NTM showed superior performance in terms of accuracy of sensitivity analysis compared to the existing random masking.

▸ **Key words**: Target Masking, Deep Learning, BERT, Newly Coined Words, Sentiment Analysis

[요 약]

최근 대량의 텍스트 분석을 위해 딥 러닝(Deep Learning)을 활용하는 연구들이 활발히 수행되고 있으며, 특히 대량의 텍스트에 대한 학습 결과를 특정 도메인 텍스트의 분석에 적용하는 사전 학습 언어 모델(Pre-trained Language Model)이 주목받고 있다. 다양한 사전 학습 언어 모델 중 BERT(Bidirectional Encoder Representations from Transformers) 기반 모델이 가장 널리 활용되고 있으며, 최근에는 BERT의 MLM(Masked Language Model)을 활용한 추가 사전 학습(Further Pre-training)을 통해 분석 성능을 향상시키기 위한 방안이 모색되고 있다. 하지만 전통적인 MLM 방식은 신조어와 같이 새로운 단어가 포함된 문장의 의미를 충분히 명확하게 파악하기 어렵다는 한계를 갖는다. 이에 본 연구에서는 기존의 MLM을 보완하여 신조어에 대해서만 집중적으로 마스킹을 수행하는 신조어 표적 마스킹(NTM: Newly Coined Words Target Masking)을 새롭게 제안한다. 제안 방법론을 적용하여 포털 'N'사의 영화 리뷰 약 70만 건을 분석한 결과, 제안하는 신조어 표적 마스킹이 기존의 무작위 마스킹에 비해 감성 분석의 정확도 측면에서 우수한 성능을 보였다.

▸ **주제어**: 표적 마스킹, 딥러닝, BERT, 신조어, 감성분석

• First Author: Gun-Min Nam, Corresponding Author: Namgyu Kim
*Gun-Min Nam (nam835@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
*Namgyu Kim (ngkim@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
• Received: 2021. 09. 13, Revised: 2021. 10. 06, Accepted: 2021. 10. 06.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 스마트 기기의 사용 증가로 인하여 페이스북, 트위터, 그리고 인스타그램 등 다양한 SNS가 활발하게 사용되고 있다. 이로 인해 이들 SNS를 통해 생산되는 방대한 텍스트를 분석하고자 하는 수요가 꾸준히 증가하고 있으며, 이에 따라 텍스트 마이닝(Text Mining)[1] 및 자연어 처리(Natural Language Processing) 관련 기술에 대한 관심도 높아지고 있다. 텍스트 마이닝이란 웹페이지나 블로그, 이메일 등 전자문서로 된 텍스트 자료에 대한 분석을 통해 유용한 정보를 추출하는 일련의 과정을 의미하며, 대표적인 연구 분야로는 텍스트 집합 내에서 단어들의 잠재적인 주제를 추출하는 토픽 모델링(Topic Modeling)[2], 단어가 갖는 품사의 특징을 활용한 텍스트 증강(Text Augmentation)[3], 방대한 내용을 짧은 분량으로 축약하여 생성하는 텍스트 요약(Text Summarization)[4], 텍스트에 담긴 평가, 태도 등 주관적인 정보와 긍정, 부정 등 감성의 표현 정도를 파악하는 감성 분석(Sentiment Analysis)[5] 등이 있다. 또한 비정형 구조로 표현된 텍스트 데이터를 정형 구조인 벡터로 변환하기 위한 다양한 임베딩(Embedding) 기법[6]이 고안되고 있으며, 특히 최근에는 다양한 딥 러닝(Deep Learning) 기법이 텍스트 임베딩에 적용되어 괄목할 만한 성과를 나타내고 있다.

딥 러닝 기반 텍스트 임베딩은 전술한 바와 같은 다양한 자연어 처리 분야의 성능 향상에 크게 기여하고 있으며, 특히 대량의 데이터를 미리 학습하여 그 결과를 모델로 공개하여 재사용하는 사전 학습 언어 모델(Pre-trained Language Model)을 통해 그 활용성이 향상되었다. 구체적으로는 방대한 데이터에 대한 사전 학습을 통해 획득한 가중치를 하위 문제(Downstream Task) 해결을 위한 분석 과정에서 미세 조정(Fine-tuning)하는 방식, 즉 전이 학습(Transfer Learning)을 통해 사전 학습된 지식을 하위 문제 해결에 사용한다. 전이 학습은 딥 러닝 학습의 한계, 즉 학습을 위해 방대한 데이터와 오랜 시간 및 자원이 필요하다는 한계를 극복하였으며, 이로 인해 텍스트 분석 분야에서도 ELMO(Embeddings from Language Model)[7], BERT(Bidirectional Encoder Representations from Transformer)[8] 등 다양한 사전 학습 언어 모델을 통한 전이 학습이 많이 활용되고 있다.

이 중 텍스트 분야에서 가장 대표적인 모델 중의 하나인 BERT는 레이블이 없는 대용량의 텍스트 데이터를 학습 데이터로 사용하며, 기존 모델과 달리 양방향 학습을 통해 단어 표현이 문맥을 더욱 정확하게 반영할 수 있게 하였다. BERT는 분석 대상 언어에 맞춰 다양한 형태로 제공되고 있으며, 한국어 분석의 경우 KOBERT[9]가 다양한 연구에서 가장 널리 활용되고 있다. KOBERT는 SKT에서 BERT 기반 한국어 분석을 위해 개발하였으며, 위키피디아 및 뉴스 등에서 수집한 한국어 문장 수백만 개로 구성된 대규모 말뭉치를 학습한 한국어 언어 모델이다.

BERT의 주요 학습 방식인 MLM(Masked Language Model)은 입력 문장 내의 단어를 무작위(Random)로 선정 후 마스킹(Masking)하고, 이후 주변 단어들로부터 마스킹 된 단어를 추론하는 방식으로 학습을 수행한다. 이러한 과정을 통해 문장 내에서 각 단어가 갖는 의미를 정밀하게 파악할 수 있다. 특히, 최근에는 각 단어가 하위 과제와 무관하게 특정 도메인에서 갖는 특수한 의미를 학습하기 위해, 사전 학습과 미세 조정 사이에 MLM을 활용한 추가 사전 학습(Further Pre-training)을 진행하는 방안이 모색되고 있다. 본 연구의 관점에서 바라본 추가 사전 학습의 목적 및 특징은 <Table 1>과 같다.

Table 1. Characteristics of Further Pre-training

	Purpose	Domain Dependency	Task Dependency
Pre-training	Understanding General Vocabulary	X	X
Further Pre-training	Understanding Vocabulary in Specific Domain	O	X
Fine-tuning	Improving Performance of Downstream Tasks	O	O

추가 사전 학습에는 BERT와 동일하게 MLM이 사용되며, 사전 학습된 BERT 모델의 최종 가중치를 가중치 초깃값으로 사용한다. 이 때, MLM은 마스크 되지 않은 단어들의 의미를 통해 마스크가 적용된 단어의 의미를 추론하기 때문에, 사전 학습을 통해 의미가 잘 파악된 단어들이 마스크 되어 가려지고 신조어와 같이 의미가 아직 파악되지 않은 단어들이 마스크가 적용된 단어의 추론에 사용된다면 문맥에 따른 단어의 의미가 정확하게 파악되기 어렵다. 이러한 현상은 '꿀잼'이라는 신조어를 포함한 문장에 MLM을 적용한 예를 나타낸 <Fig. 1>을 통해 살펴볼 수 있다.

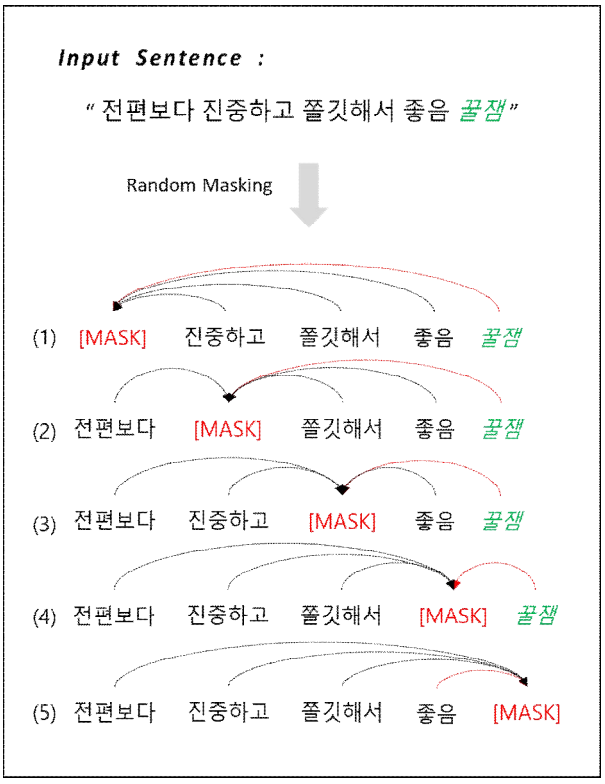

Fig. 1. MASK Candidates of Traditional MLM

<Fig. 1>의 입력 문장은 영화 도메인 댓글 중의 하나로, 긍정의 의미로 분류된 것으로 가정한다. 이 때, <Fig. 1>의 전통적 MLM에 근거하여 무작위 마스킹을 적용할 경우 (1) ~ (5) 중 한 가지 시나리오가 발생하며, BERT는 무작위 방식으로 마스크 단어를 선정하므로 각 시나리오는 동일한 비율로 발생하게 된다. 본 연구에서는 이 중 (5)를 바람직한 경우로 간주하며, 나머지 (1) ~ (4)를 바람직하지 않은 경우로 간주한다. 왜냐하면 (1) ~ (4)의 경우, 사전 학습을 통해 의미가 잘 파악된 단어들이 마스크 되어 가려지고 의미가 전혀 파악되지 않은 단어인 '꿀잼'이 마스크가 적용된 단어의 추론에 사용되었기 때문에 추가 사전 학습이 충분히 효율적으로 이루어지기 어렵기 때문이다. 특히, 사전 학습은 방대한 데이터에 대해 충분한 학습이 이루어지므로 학습 과정에서 단어 의미의 모호성이 해소될 여지가 있지만, 추가 사전 학습은 이와 달리 소량의 데이터에 대해 학습이 이루어지는 경우가 일반적이기 때문에 이러한 한계는 큰 부작용을 야기할 수 있다.

이러한 점에 착안하여 본 연구에서는 의미에 대한 학습이 전혀 이루어지지 않은 신조어에만 마스킹을 적용하는 신조어 표적 마스킹(NTM: Newly Coined Words Target Masking) 기반 추가 사전 학습 방법론을 제안한다. 즉, 제안 방법론은 신조어 중심의 마스킹을 수행하여, 사전 학습을 통해 의미가 파악된 단어로부터 의미가 알려지지 않은 신조어에 의미를 추론하는 학습에 집중한다. 또한, 제안 방법론의 성능 평가를 위해 일반 사전 학습 모델만 적용한 경우, 기존의 무작위 MLM 기반 추가 사전 학습을 수행한 경우, 그리고 제안 방법론인 NTM을 적용한 경우의 하위 과제 분석 성능을 실험을 통해 비교하여 보이고자 한다.

본 논문의 이후 구성은 다음과 같다. 먼저, 다음 장인 2장에서는 텍스트 분석과 관련된 연구, 특히 딥 러닝을 활용한 연구 및 사전 학습 언어 모델에 대한 선행 연구를 살펴본다. 또한 3장에서는 간략한 예시를 통해 본 연구의 전체 방법론을 소개한다. 다음으로 4장에서는 실제 데이터에 대해 제안 방법론 및 기존 방법론을 적용한 실험의 결과를 소개하고, 마지막 5장에서는 본 연구의 기여 및 한계, 그리고 추후 연구 방향을 제시한다.

II. Preliminaries

1. Deep Learning for Text Data

머신러닝(Machine Learning)의 한 분야인 딥 러닝은 방대한 데이터를 활용하여 여러 은닉층으로 구성된 심층 신경망 구조의 각 은닉층에 연결된 가중치를 학습하는 알고리즘으로, 기존의 일반적인 머신러닝에 비해 우수한 성능을 보이고 있다. 구체적으로 딥 러닝을 적용한 자연어 처리는 텍스트 내의 단어 의미 파악을 위해 단어들이 은닉층을 통과하는 과정을 거친다. 이를 위해, 단어를 벡터로 표현하는 임베딩 기법을 사용하며, 대표적인 기법으로는 Word2Vec[10], FastText[11] 등이 존재한다. 하지만, 이들은 단어의 의미 추론 과정에서 한정된 범위의 문맥만을 참조하므로, 텍스트의 전반적인 의미와 문맥을 충분히 다루지 못한다는 한계점이 있다.

이를 극복하고자 딥 러닝을 통계 기반의 언어 모델에 적용하여 은닉층 노드의 출력값이 다음 노드의 입력값으로 들어가는 시퀀스 임베딩(Sequence Embedding) 방법들이 등장하였으며, 대표적인 모델로는 학습 과정에서 시퀀스를 연속적으로 적용하는 순환 신경망 모델인 RNN(Recurrent Neural Network)[12]을 들 수 있다. 하지만 RNN은 시퀀스의 길이가 증가할수록 과거에 등장한 단어들의 의미가 충분한 전달되지 않는다는 한계를 갖는다. 이러한 한계를 해결하기 위해 길이가 증가하여도 단어들의 의미를 충분히 반영할 수 있는 LSTM(Long Short-Term Memory)[13], GRU(Gated Recurrent Unit)[14] 등이 등장하였지만 이들 또한 학습되지 않은 단어가 등장하였을 경우 처리할 수 없는 문제(Out Of Vocabulary)와 장기 의존성을 해결하지 못하는 문제를 가지고 있다.

이러한 한계점을 해결하기 위해 어텐션 메커니즘(Attention Mechanism)[15]이 등장하였다. 어텐션 메커니즘은 학습 과정에서 문장 내의 중요한 단어에 집중하여 학습하는 기법으로, 예측할 단어와 연관이 있는 입력 단어 부분을 중점으로 살펴보며 처리해야 할 정보의 양을 감소시킴으로써 딥 러닝 기반 자연어 이해의 성능을 한 단계 더 발전시킨 것으로 평가받고 있다. 최근에는 셀프 어텐션(Self-Attention)을 기반으로 하는 트랜스포머(Transformer)[16] 모델이 등장하였으며, 이는 토큰 간의 관계를 학습하는 과정에서 광범위한 문맥을 고려할 수 있다는 장점을 갖는다.

2. Pre-trained Language Model

딥 러닝 알고리즘은 많은 학습 시간 및 대량의 학습 데이터를 필요로 한다는 한계점을 갖는다. 이를 극복하고자, 최근에는 트랜스포머 모델의 개념을 적용하여 사전 학습 언어 모델을 구축하려는 연구가 진행되고 있다. 사전 학습 언어 모델은 위키피디아, 뉴스 기사 등 방대한 양의 텍스트 말뭉치에 대한 학습을 통해 구축된 모델로, 사전 학습 언어 모델을 기반으로 추후 소량의 데이터에 대한 미세 조정을 진행하는 전이 학습을 통해 텍스트 이해 및 표현 분야에서 매우 우수한 성능을 나타내고 있다. 신경망을 기반으로 하는 대표적인 사전 학습 언어 모델로는 BERT, ELMO 등을 들 수 있으며, 그중에서도 BERT는 트랜스포머의 인코더 구조를 기반으로 MLM 및 NSP(Next Sentence Prediction) 방식의 비지도 학습을 수행함으로써 기존 단방향 학습 방식의 한계점을 개선하였다.

BERT의 학습 방식 중, MLM은 문장 내 전체 토큰의 15%에 마스킹을 적용하여 해당 단어를 예측하는 과정을 통해 학습이 이루어지며, 15% 중 80%는 Mask 토큰으로 대체, 10%는 다른 단어로 대체, 그리고 나머지 10%는 변경하지 않고 그대로 유지하는 상태에서 학습이 진행된다. NSP는 입력된 두 문장이 동시에 주어졌을 경우, 첫 문장 이후 다음 문장을 예측하는 방식으로 학습이 진행된다. 최근에는 여러 분석 과제의 성능을 높이고자 사전 학습된 BERT를 활용하여 텍스트와 문서 벡터를 추론하는 연구[17,18]와 금융 정보의 감성분석을 위해 학습한 연구[19]가 이루어졌다.

한편, 최근에는 일반 데이터 말뭉치로 사전 학습된 BERT를 기반으로 특정 도메인의 지식을 BERT에 추가하기 위한 추가 사전 학습을 진행하여 특정 도메인에서 언어 모델의 성능을 더욱 향상시키기 위한 많은 연구들[20]이 수행되고 있으며, 특히 사전 학습에 사용되었던 MLM을 특정 도메인에 적용하여 추가로 학습하는 연구[21,22]들이 다양하게 이루어지고 있다. 하지만 이들 연구 대부분은 무작위로 단어를 선택하여 마스킹을 진행하는 일반적인 MLM 방식을 사용한다. 따라서 사전 학습을 통해 의미가 잘 파악된 단어들이 마스크 되어 가려지고, 이러한 단어의 추론에 의미가 전혀 파악되지 않은 새로운 단어들이 사용될 가능성이 존재한다는 한계를 갖는다. 이에 본 연구에서는 사전 학습에 포함되지 않은 특정 단어에 대해서만 집중적으로 마스킹을 수행하여 보다 효율적인 추가 사전 학습을 수행할 수 있는 방안을 제시한다.

III. The Proposed Scheme

1. Overall Research Process

본 장에서는 본 연구에서 제안하는 신조어 표적 마스킹의 원리에 대한 전체적인 개념을 설명하며, <Fig. 2>는 제안 방법론의 개요를 나타낸다.

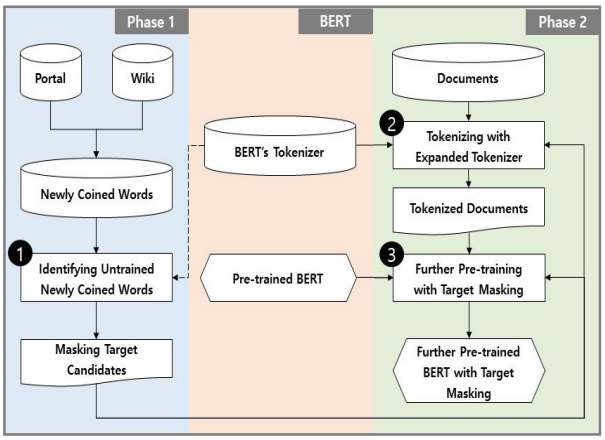

Fig. 2. Overview of the Proposed Methodology

제안 방법론은 다양한 매체를 참고하여 신조어를 추출한 뒤 이를 기반으로 마스킹 타겟을 선정하는 Phase 1, 그리고 선정된 마스킹 타겟을 MLM에 적용하여 추가 사전 학습을 진행하는 Phase 2로 구성된다. Phase 1과 Phase 2의 동작 과정은 본 장의 이후 절에서 가상의 예를 통해 자세히 소개한다.

2. Selection of Mask Candidates

이번 절에서는 <Fig. 2>의 Phase 1을 가상의 예를 통해 소개한다. 먼저, 각종 매체[23,24]를 참고하여 신조어를 수집하는 과정부터 살펴본다.

우선 <Fig. 3>과 같이, 위키피디아(Wikipedia)에서 우리나라의 인터넷 신조어 목록을 정리한 자료[23]를 수집하

였다. 해당 사이트에는 일반 인터넷 신조어, 준말에서 나온 신조어, 온라인 게임에서 유래한 신조어, 방송에서 나온 신조어, 그리고 정치 관련 신조어 등이 정리되어 있으며, 이 가운데 자음, 단일어, 그리고 정치적인 의미를 가진 단어를 제외하여 약 100개의 신조어를 수집하였다.

Fig. 3. Newly Coined Words in Wikipedia

또한 신조어를 정리한 여러 블로그를 통해 약 300개의 신조어를 추가로 수집하였으며, 그중 하나의 예는 <Fig. 4>와 같다. 최종적으로 위키피디아, 블로그에서 수집한 신조어들을 통합하여 중복을 제거한 후 총 338개의 신조어 리스트를 구성하였다.

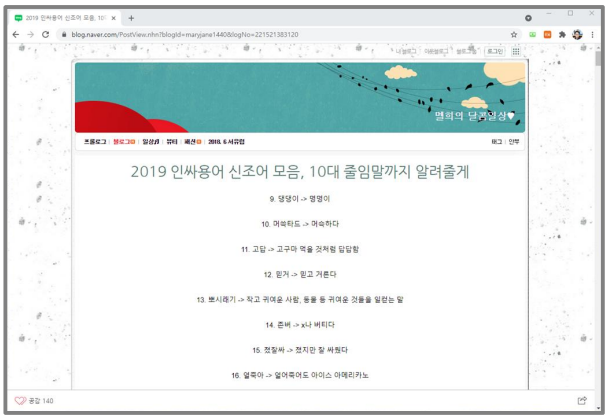

Fig. 4. Newly Coined Words in Blog

제안 방법론은 <Fig. 5>와 같이, 이렇게 식별된 신조어 중 BERT 토크나이저에 이미 포함된 단어를 제외한 나머지를 마스킹 타겟 후보로 사용한다. 만약, 이러한 과정 없이 신조어 리스트 전부가 마스킹 타겟 후보가 된다면, 이미 사전 학습이 완료되어 의미 파악이 이루어진 신조어들도 마스킹에 포함될 수 있기 때문이다.

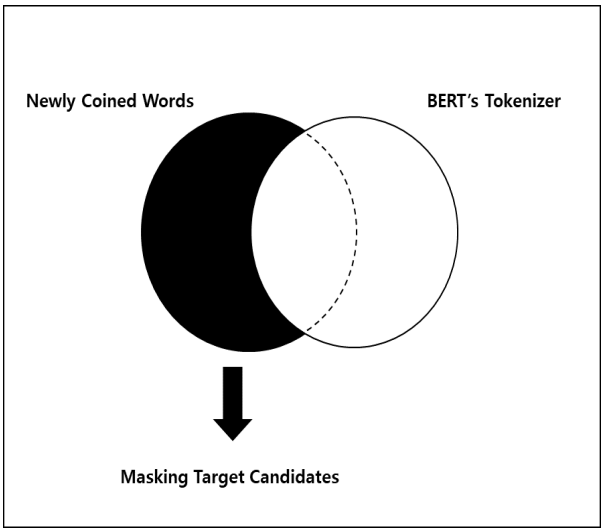

Fig. 5. Masking Target Candidates Selection Method

<Fig. 5>의 정제 과정을 거친 후 최종 추가 신조어 리스트를 확정하며, 이들 단어들이 마스킹 타겟 후보로 사용된다. <Table 2>는 이러한 전체 과정의 결과로 도출된 학습되지 않은 신조어 리스트의 예를 나타낸다.

Table 2. Untrained Newly Coined Words

급식충	극혐	꿀잼	뇌피셜	덕후
컴백	브금	어그로	레알	제곧내
...				
프사	셀카	띵작	광탈	띵작
자만추	극딜	먹방	갠톡	눈팅

<Table 2>에 표시된 단어들은 위키피디아 또는 블로그에서 수집한 신조어이며, 이 중 흐리게 표시된 '컴백', '레알', 그리고 '셀카'는 신조어 리스트에 존재하지만 BERT 토크나이저에도 이미 포함되어 있어서 추가 신조어 후보에서 제외된 단어를 나타낸다. 진하게 표시된 단어들인 '급식충', '극혐', '꿀잼', '뇌피셜', 그리고 '눈팅' 등은 BERT 토크나이저에 포함되지 않은 신조어를 나타내며, 최종적으로 이러한 단어들이 마스킹 타겟 후보(MTC : Masking Target Candidates)로 사용된다.

3. Further Pre-training with Target Masking

이번 절에서는 <Fig. 2>의 Phase 1에서 선정된 마스킹 타겟 후보를 BERT 토크나이저의 사전(Vocabulary)에 추

가하여 확장된 토크나이저(Expanded Tokenizer)를 구성하고, 이를 주어진 문서에 적용하여 신조어의 의미를 최대한 보존하는 형태로 분절을 수행하는 과정을 소개한다.

BERT는 분석 대상 텍스트의 분절화를 위해 BERT 토크나이저 사전을 사용한다. 즉 사전에 포함된 토큰을 의미의 최소 단위로 파악하므로, 사전에 포함되지 않은 단어는 여러 부분으로 분절되어 본래의 의미를 상실하게 된다. 이러한 예는 <Fig. 6>을 통해 확인할 수 있다.

Fig. 6. Tokenizing with Expanded Tokenizer

<Fig. 6>에서 신조어인 '지못미'를 기존의 BERT 토크나이저로 분절할 경우, '지', '못', '미'로 나누어지는 것을 확인할 수 있는데, 이는 BERT 토크나이저는 '지못미' 단어를 하나의 묶음으로 인식하지 못하기 때문이다. '지못미'와 같은 신조어가 분절되지 않고 형태를 보존하기 위해서는 기존의 사전에 마스킹 타겟 후보 단어들을 추가해야 하며, 그 결과 신조어가 분절되지 않고 형태가 잘 유지되는 것을 <Fig. 6>의 하단 예에서 확인할 수 있다.

확장된 토크나이저를 사용하여 신조어 표적 마스킹을 수행하는 과정은 <Fig. 7>과 같다. <Fig. 7>의 (A)는 기존의 일반적인 KoBERT 사전 학습 과정을 나타내며, 사전 학습이 완료된 KoBERT의 최종 가중치가 추가 사전 학습 과정인 (B)의 초기 가중치로 사용된다. <Fig. 7>의 (B)는 사전 학습이 완료된 KoBERT를 통해 영화 도메인에 특화된 추가 사전 학습을 진행하는 과정을 나타내며, 입력으로 사용된 문장들은 영화 댓글의 실제 예이다. 전통적인 MLM 기반 추가 사전 학습과 달리 제안 방법론에서는 신조어 표적 마스킹을 수행하며, 이에 따라 <Fig. 7>의 (B)에서는 '띵작', '꿀잼', 그리고 '강추' 등의 신조어에 대해서만 마스킹이 이루어진다.

IV. Experiment

1. Experimental Design

본 장에서는 3장에서 제안한 방법론을 실제 데이터의 분석에 적용한 실험의 과정과 결과를 소개한다. 실험을 위해 'N'사 영화 리뷰 데이터 중 2016년 2월부터 2021년 1월의 60개월 동안 작성된 약 70만 건을 수집하였다. 각 리뷰는 1점부터 10점까지의 평점을 부여하고 있으며, 이 중

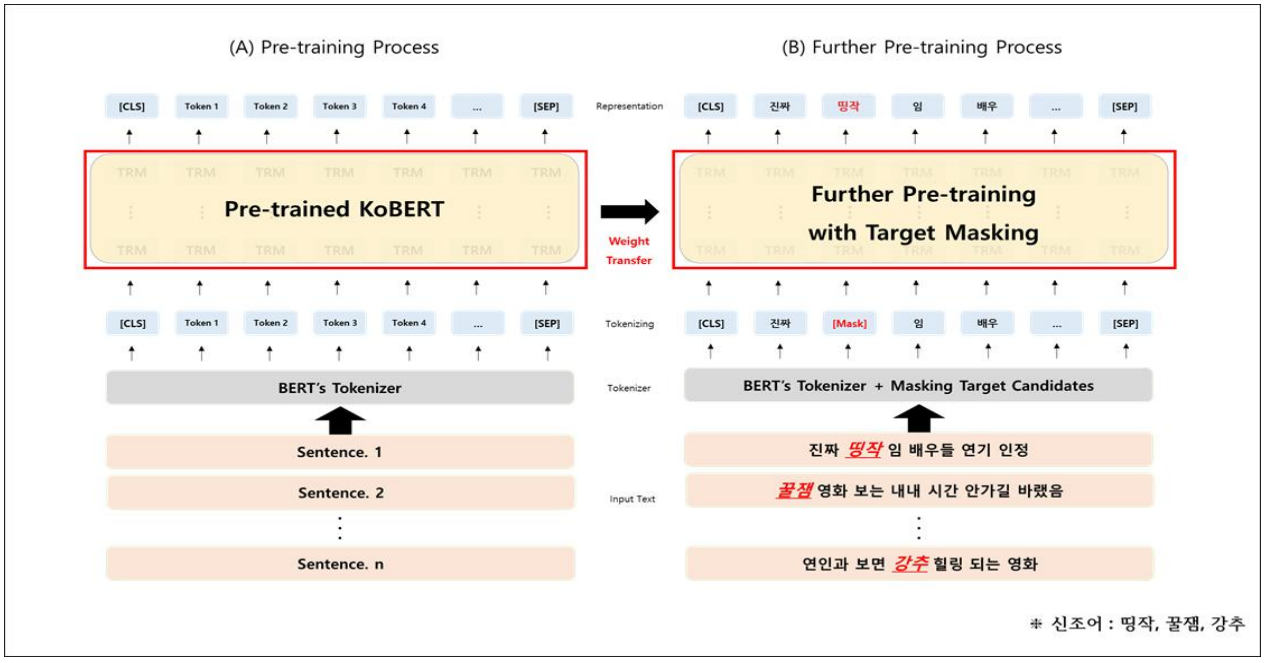

Fig. 7. Further Pre-training with Target Masking Model Process

10점, 1점에 해당하는 리뷰 데이터 약 12만 건을 추출하여 각각 긍정, 부정의 값을 부여한 후 분석에 사용하였다.

추출된 12만 건의 리뷰 데이터를 <Table 3>과 같이 추가 사전 학습용 데이터와 분석 과제인 감성 분석용 데이터로 분할하였다. 추가 사전 학습용 데이터는 긍정, 부정 4만 개씩 총 8만 개로 구성되어 있으며, 모든 문장이 신조어를 포함하고 있다. 한편 감성 분석용 데이터는 긍정, 부정 2만 개씩 총 4만 개로 구성되어 있으며, 이들 중 절반은 신조어를 포함한 문장, 나머지 절반은 신조어를 포함하지 않은 문장이다. 추가 사전 학습은 Pytorch를 기반으로 하는 KoBERT를 사용하였으며, 실험 환경은 Python 3.7을 통해 구축하였다.

Table 3. Composition of Experimental Data

Movie Reviews	Including New Words	Positive	Negative
Further Pre-training (80,000)	YES	40,000	40,000
Sentiment Classification (40,000)	YES	10,000	10,000
	NO	10,000	10,000

본 장에서는 제안 모델의 성능을 비교하기 위해 <Fig. 8>과 같이 총 3가지 모델을 구축하였다. <Fig. 8>의 (A) 모델은 별도의 추가 사전 학습을 진행하지 않고 기본 BERT만을 사용한 모델이며, (B)와 (C) 모델은 사전 학습된 BERT에 추가 사전 학습을 수행한 모델이다. 다만 추가 사전 학습 과정에서 (B)는 전통적인 무작위 방식의 MLM을 사용하였으며, (C)는 본 논문에서 제안하는 방식인 신조어 표적 마스킹을 사용하였다. 즉 본 실험에서 (A) 모델과 (B) 모델의 비교는 추가 사전 학습의 효과를 확인하기 위한 것이며, (B) 모델과 (C) 모델의 비교는 추가 사전 학습 과정에서 무작위 마스킹에 비해 제안 방식인 신조어 표적 마스킹이 나타내는 효과를 확인하기 위한 것이다.

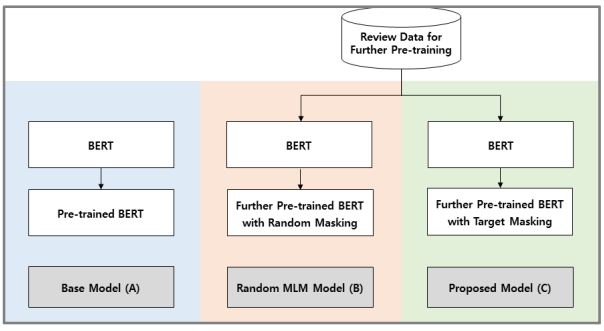

Fig. 8. Three Models for Performance Comparison

2. Experimental Results

본 절에서는 <Fig. 8>에서 구축한 세 가지 언어 모델에 대한 우수성을 평가하기 위해, 각 언어 모델을 사용하여 감성 분석을 수행한 결과를 정확도 측면에서 비교한다. 이를 위해 <Table 3>의 데이터 중 감성 분석용 데이터 4만 건을 60%, 20%, 20%로 나누어 각각 학습(Train), 검증(Validation), 테스트(Test) 용으로 사용하였으며, 감성 분석에는 Kim CNN[25] 모델을 사용하였다.

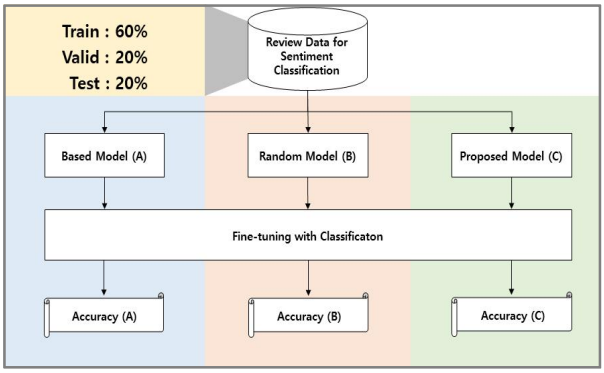

Fig. 9. Architecture of Sentiment Classification

세 가지 언어 모델을 통해 감성 분석을 수행한 결과를 비교하기 위한 평가 지표로는 분류 정확도(Accuracy)를 사용하였다. <Fig. 10>은 세 가지 모델을 적용한 학습, 검증, 그리고 테스트 정확도를 나타낸다. 실험 결과 추가 사전 학습을 적용하지 않은 (A) 모델에 비해, 추가 사전 학습 과정이 더해진 (B),(C) 모델의 분류 정확도가 높게 나타남을 확인하였다. 또한 (B)와 (C) 중에서도 무작위로 MLM을 적용한 (B) 모델에 비해, 신조어 표적 마스킹을 적용한 제안 모델인 (C)가 더욱 높은 분류 정확도를 나타냄을 확인하였다.

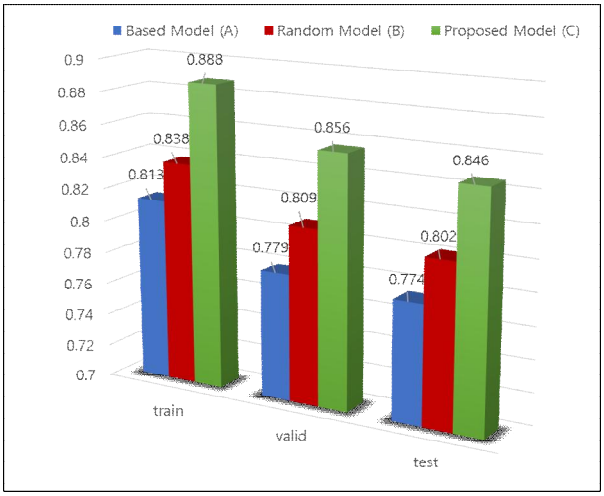

Fig. 10. Performance Evaluation (Accuracy)

V. Conclusions

최근 대량의 텍스트 분석을 위해 대량의 텍스트에 대한 학습 결과를 특정 도메인 텍스트의 분석에 적용하는 사전 학습 언어 모델이 주목받고 있으며, 최근에는 BERT의 MLM을 활용한 추가 사전 학습을 통해 하위 과제의 분석 성능을 향상시키기 위한 방안이 모색되고 있다. 본 연구에서는 사전 학습에 포함되지 않은 단어의 의미를 더욱 정확하게 이해하기 위해 전통적인 MLM을 변형한 신조어 표적 마스킹 방법론을 제안하였으며, 포털 'N'사의 영화 리뷰 약 70만 건을 분석한 실험을 통해 제안하는 신조어 표적 마스킹이 기존의 무작위 마스킹에 비해 감성 분석의 정확도 측면에서 우수한 성능을 보임을 확인하였다.

본 연구에서는 딥 러닝 기반 텍스트 분석 분야에서 많은 연구가 이루어지고 있는 BERT의 무작위 MLM 기반 추가 사전 학습의 개선 방향을 제시했으며, 이는 본 연구의 학술적 기여로 인정받을 수 있다. 또한 추가 사전 학습을 통한 성능 향상 및 신조어 표적 마스킹을 통한 성능 향상 방안을 제시했다는 점에서 본 연구의 성과는 실무적으로도 활용도가 매우 높을 것으로 기대한다.

한편, 추후 연구에서는 다음과 같은 분석이 이루어져야 한다. 우선 본 연구에서는 BERT의 사전 학습에 포함되지 않은 단어로 신조어를 선택하여 분석을 수행하였지만, 추후 각 도메인의 전문어 등 사전 학습에 포함되지 않은 다른 유형의 단어에 대해서도 표적 마스킹을 수행하는 방안을 모색할 필요가 있다. 또한 추가 사전 학습의 성능은 학습에 사용되는 데이터의 양에 의해 크게 좌우되는 바, 신조어 표적 마스킹에 사용되는 학습 데이터의 양에 따른 성능 개선 정도에 대한 엄밀한 분석이 이루어져야 한다. 또한 신조어와 일반어의 혼합 마스킹을 통해 분석 성능을 향상시키는 방안에 대한 고찰이 이루어질 필요가 있다.

REFERENCES

[1] A. Tan, "Text Mining: The State of the Art and the Challenges," Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases, pp. 65-70 Jan, 1999.

[2] B. Gretarsson, J.O. Donovan, S. Bostandjiev, T. Hollerer, A. Asuncion, D. Newman, and P. Smyth, "TopicNets : Visual Analysis of Large Text Corpora with Topic Modeling," ACM Transactions on Intelligent Systems and Technology, Vol. 3, No. 2, pp. 1-26, Feb, 2012.

[3] M. Kim, N. Kim, "Text Augmentation Using Hierarchy-based Word Replacement," Journal of The Korea Society of Computer and Information, Vol. 26, No. 1, pp. 57-67, Jan, 2021.

[4] W. Gao, P. Li, and K. Darwish, "Joint Topic Modeling for Event Summarization Across News and Social Media Streams," Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 1173-1182, Nov, 2012.

[5] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, Vol. 5, No. 1, pp. 1-167, May, 2012.

[6] Q. Le, T. Mikolov, "Distributed Representations of Sentences and Documents," Proceedings of the 31st International Conference on Machine Learning, Vol. 32, pp. 1188-1196, May, 2014.

[7] M. Peter, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep Contextualized Word Representations," Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, Vol. 1, pp. 2227-2237, Mar, 2018.

[8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, May, 2018.

[9] SKTBrain, https://github.com/SKTBrain/KoBERT

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Proceedings of the International Conference on Learning Representations, ICLR, Sep, 2013.

[11] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," Transactions of the Association for Computational Linguistics, Vol. 5, pp. 135-146, Jun, 2017.

[12] T. Mikolov, M. Karafiát, L. Burget, and J. Cernocký, "Recurrent Neural Network Based Language Model," Eleventh Annual Conference of the International Speech Communication Association, Sep, 2010.

[13] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Journal of Neural Computation, Vol. 9, No. 8, pp. 1735-1780, Nov, 1997.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," arXiv:1412.3555, Dec, 2014.

[15] D. Bahdanau, C. Kyunghyun, and B. Yoshua, "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv:1409.0473, May, 2014.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000-6010, 2017.

[17] Y. Yun, N. Kim, "Self-Supervised Document Representation Method," Journal of The Korea Society of Computer and

Information, Vol. 25, No. 5, pp.187-197, May, 2020.
[18] A. Adhikari, A. Ram, R. Tang, and J. Lin, "DocBERT: BERT for Document Classification," arXiv:1904.08398, Apr, 2019.
[19] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," arXiv:1908.10063, Aug, 2019.
[20] V. D. Viellieber, and M.. Aßenmacher, "Pre-trained Language Models as Knowledge Bases for Automotive Complaint Analysis," arXiv:2012.02558, Dec, 2020.
[21] C. Sung, T. Dhamecha, S. Saha, T. Ma, V. Reddy, and R. Arora, "Pre-training BERT on Domain Resources for Short Answer Grading," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, pp. 6071-6075, Nov, 2019.
[22] Y. Gu, Z. Zhang, X. Wang, Z. Liu, and M. Sun, "Train No Evil: Selective Masking for Task-guided Pre-training," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 6966-6974, Nov, 2020.
[23] Wikipedia, https://ko.wikipedia.org/wiki/대한민국의_인터넷_신조어_목록.
[24] Naver Blog, https://blog.naver.com/PostView.nhn?blogId=maryjane1440&logNo=221521383120.
[25] Y. Kim, "Convolutional Neural Networks for Sentence Classification," arXiv:1408.5882, Sep, 2014.

Authors

Gun-Min Nam received the B.S. degree in Mathematics and Informational Statistics from Wonkwang University in 2020 and currently enrolled in Graduate School of Business IT, Kookmin University.

Gun-Min Nam is interested in text mining and deep learning.

Namgyu Kim received the B.S. in Computer Engineering from Seoul National University in 1998, M.S. and Ph.D. degrees in Management Engineering from KAIST, Korea, in 2000 and 2007, respectively.

Dr. Kim joined the faculty of the School of Management Information Systems at Kookmin University, Seoul, Korea, in 2007. He is currently a professor in the Graduate School of Business IT at Kookmin University. He is interested in text mining, deep learning, and data modeling.