ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2022 December: 28(4): 309~328 http://dx.doi.org/10.13088/jiis.2022.28.4.309

한국어 자연어생성에 적합한 사전훈련 언어모델 특성 연구*

송민채
농협중앙회 디지털혁신실
(nicenara@nonghyup.com)

신경식
이화여자대학교 경영대학
(ksshin@ewha.ac.kr)

본 연구는 자연어처리의 분석목적과 추론데이터 성격에 적합한 한국어 사전훈련 언어모델의 특성을 실증분석했다. 이를 위해 자연어생성이 가능한 대표적 사전훈련 언어모델인 BART와 GPT 모델을 실험에 사용했다. 구체적으로 한국어 텍스트를 BART와 GPT 모델에 학습한 사전훈련 언어모델을 사용해 문서요약 생성 성능을 비교했다. 다음으로 추론데이터의 특성에 따라 언어모델의 성능이 어떻게 달라지는지 확인하기 위해 6가지 정보전달성과 4가지 창작물 유형의 한국어 텍스트 문서에 적용했다. 그 결과, 모든 문서유형에서 인코더와 디코더가 모두 있는 BART의 구조가 디코더만 있는 GPT 모델보다 더 높은 성능을 보였다. 추론데이터의 특성이 사전훈련 언어모델의 성능에 미치는 영향을 살펴본 결과, KoGPT는 데이터의 길이에 성능이 비례한 것으로 나타났다. 그러나 길이가 가장 긴 문서에 대해서도 KoGPT보다 KoBART의 성능이 높아 다운스트림 태스크 목적에 맞는 사전훈련 모델의 구조가 자연어생성 성능에 가장 크게 영향을 미치는 요소인 것으로 나타났다. 추가적으로 본 연구에서는 정보전달성과 창작물로 문서의 특징을 구분한 것 외에 품사의 비중으로 문서의 특징을 파악해 사전훈련 언어모델의 성능을 비교했다. 그 결과, KoBART는 어미와 형용사/부사, 동사의 비중이 높을수록 성능이 떨어진 반면 명사의 비중이 클수록 성능이 좋았다. 반면 KoGPT는 KoBART에 비해 품사의 비중과 상관도가 낮았다. 이는 동일한 사전훈련 언어모델이라도 추론데이터의 특성에 따라 자연어생성 성능이 달라지기 때문에 다운스트림 태스크에 사전훈련 언어모델 적용 시 미세조정 외에 추론데이터의 특성에 대한 고려가 중요함을 의미한다. 향후 어순 등 분석을 통해 추론데이터의 특성을 파악하고, 이것이 한국어 생성에 미치는 영향을 분석한다면 한국어 특성에 적합한 언어모델이나 자연어생성 성능 지표 개발이 가능할 것이다.

주제어 : 사전훈련 언어모델, 트랜스포머, 문서 생성요약, BART, GPT

논문접수일 : 2022년 11월 14일 논문수정일 : 2022년 12월 11일 게재확정일 : 2022년 12월 14일
원고유형 : 학술대회 Fast Track 교신저자 : 신경식

1. 서론

자연어란 기계어와 같은 인공언어에 대응되는 개념으로, 일상에서 자연히 발생하여 쓰이는 언어 그 자체를 의미한다. 자연어처리는 이러한 자연언어를 기계가 이해하도록 분석하는 기술로서, 크게 컴퓨터가 인간이 만들어 놓은 대량의 문서에서 정보를 얻는 자연어이해와 자연어이해를 통해 얻어진 정보를 사람이 이해할 수 있게 사람의 언어로 다시 표현하는 자연어생성으로 구분할 수 있다(Chowdhary, 2020).

한편 사전훈련 언어모델(Pre-trained Language Model)이 자연어처리에 적용되면서 자연어처리와 관련된 다양한 분야의 성능이 획기적으로 개

* 본 연구는 한국연구재단의 지원을 받아 수행되었음 (NRF-2021R1A2C1012036)

선되었다(Alyafeai et al., 2020 ; Ruderetal., 2019). 사전훈련 언어모델이란 언어모델의 결과를 전이학습(Transfer learning)을 통해 자연어이해와 자연어생성에 적용한 것이다.

언어모델은 자연어처리를 위해 사용하는 모델로서 문장 자체의 출현 확률을 예측하거나 이전 단어들이 주어졌을 때 다음 단어를 예측하는 모델이다. 이때 언어모델을 훈련하는 과정에서 문장 내 등장하는 단어의 순서를 통해 자연어의 의미를 이해하기 때문에 라벨이 있는 학습데이터를 필요로 하지 않는다. 따라서 라벨링에 소요되는 시간과 비용에 대한 부담이 적어 언어모델의 학습에 대규모 텍스트 데이터 활용이 가능하다(Brown et al., 2020 ; Lewis et al., 2019 ; Huang et al., 2020). 전이학습은 사전에 대량의 데이터를 통해 학습된 언어모델의 결과를 다른 태스크에 재사용하는 방법이다(Kryscinski et al., 2019 ; El-Kassas et al., 2021).

최근 자연어처리 연구는 BERT(Bidirectional Encoder Representations from Transformers) 등 트랜스포머(Transformer) 기반 사전훈련 언어모델이 자연어처리의 여러 분야에 적용되어 우수한 성능을 보임에 따라 기존에 하나의 모델로 하나의 과제를 해결하는 방식에서 하나의 모델로 다양한 과제를 해결하는, 전이학습 활용 방식으로 패러다임이 전환되었다(Vaswani et al., 2017).

또 다른 추세는 사전훈련 모델의 규모 확대이다. 선행연구에 따르면 사전학습한 언어모델은 데이터의 크기와 모델의 매개변수 수가 커질수록 성능이 높아지는 것으로 나타났다(Devlin et al., 2018 ; Wu et al., 2021 ; Zhang et al., 2022). 특히 2020년 openAI가 공개한 GPT(Generative Pre-Training)-3는 사전훈련 언어모델의 획기적 스케일 업을 통해 그 범용적 우수성을 입증하며 전 세계의 이목을 끌었다. GPT-3는 당시 기준으로 가장 모델 사이즈가 컸던 Microsoft의 Turing NLG보다 10배 이상 큰, 1,750억 개의 매개변수를 가진 초대규모 모델이다. GPT-3는 사전훈련 모델의 스케일 업만으로 퓨샷 러닝(Few-shot Learning)이 기존의 미세조정(Fine-tuning)을 통해얻어진 기록을 경신하는 등 놀라운 성능을 보였다(Brown et al., 2020)[1].

한편 어떠한 사전훈련 언어모델 구조가 자연어생성 목적에 적합한지, 추론데이터의 어떤 특징이 사전훈련 언어모델 성능에 영향을 주는지는 분석목적에 맞는 실험 설계와 실증분석을 통해서만 확인 가능하다. 자연어생성이 가능한 BART와 T5, GPT 등 다양한 사전훈련 언어모델을 사용해 다운스트림 태스크에 적용하고, 미세조정을 통해 성능을 높이려는 시도들을 한 연구들은 많지만 어떠한 사전훈련 언어모델 구조가 자연어생성에 더 적합한지를 분석하는 선행연구는 찾아보기 어렵다. 미세조정이나 사전훈련 언어모델의 알고리즘 수정과 같이 성능 개선을 목적으로 연구가 설계된 경우 BART보다 GPT 성능이 낮거나 유사한 연구(Guan et al., 2021 ; Li et al, 2022 ; Lin et al., 2020 ; Liu et al., 2021 ; Lu et al., 2021 ; Luo et al., 2022 ; Tang et al., 2022)도 있는 반면, GPT보다 BART 성능이 높게 나타난 연구도 있었다(Deng et al., 2021 ; Sridhar and Yang, 2022).

1) 미세조정이란 다운스트림 태스크에 해당하는 데이터를 재학습하여 타겟 태스크의 목적에 맞게 다운스트림 태스크의 모델 파라미터를 업데이트하는 것을 의미한다. 이 과정에 여전히 다운스트림 태스크에 대한 다수의 학습데이터 셋이 필요하다. 제로샷 러닝(Zero-shot Learning)은 극단적으로 다운스트림 태스크의 데이터를 전혀 사용하지 않고, 사전훈련 모델의 결과를 바로 다운스트림 태스크에 적용해 추론한다. 퓨샷 러닝은 소수의 다운스트림 태스크 데이터를 학습에 사용하고, 모델의 파라미터를 이 데이터에 맞게 업데이트 후 결과를 추론하는 방식이다.

그러나 이는 사전훈련 언어모델의 학습데이터와 매개변수가 다르고, 다운스트림 태스크에 적용하는 과정에 미세조정을 수행했기 때문에 이러한 결과가 어떤 요인에 기인한 것인지 설명하는 데 한계가 있다.

GPT-3와 같은 초거대 모델은 학습을 위해 슈퍼컴퓨팅 수준의 GPU 성능이 필수적이기 때문에 연구자나 학생, 중소·스타트업과 같은 일반기업이 직접 개발하기에는 부담이 크다. 또한 미세조정 과정에 대량의 연산을 위한 고성능 자원을 필요로 하기 때문에 사전훈련 모델을 활용하는 데에도 제약사항이 많다.

이와 같은 배경에서 본 연구는 사전학습 언어모델의 학습데이터와 매개변수 크기가 아닌, 사전훈련 언어모델의 특성과 다운스트림 태스크의 목적에 주목했다. 현재 많은 연구들이 최고 수준의 성능 경신을 위해 경쟁하며 데이터의 양, 모델의 크기 확대에 집중하고 있으나 충분히 검증된 사전훈련 모델의 결과가 공개되어 이용할 수 있다면 어떠한 기준에 따라 이를 선택하는 것이 적합한지에 대한 실증연구가 제시된다면 활용도 측면에서 그 가치가 클 것이다.

본 연구는 사전훈련 언어모델의 구조가 자연어생성 성능에 미치는 영향을 실증분석하기 위해 성능에 영향을 줄 수 있는 요인들을 실험과정에 통제했다. 즉, (1) 학습에 사용된 데이터는 동일하나 사전훈련 언어모델만 다른 경우, (2) 동일한 사전훈련 언어모델이라도 학습데이터와 매개변수의 크기가 다른 경우를 직접적으로 비교했다. 다음으로 추론데이터의 특성이 사전훈련 언어모델의 자연어생성 성능에 미치는 영향을 분석하기 위해 10가지 유형의 문서를 분석에 사용했다. 사전훈련 언어모델을 생성요약에 적용하는 과정에 미세조정은 수행하지 않았다.

2. 선행연구

2.1 전이학습과 사전훈련 언어모델

일반적인 머신러닝은 하나의 모델로 하나의 과제를 해결하는 방식이다. 반면 전이학습은 특정 태스크에서 수행한 모델의 학습결과를 다른 태스크 수행에 재사용한다. 앞서 수행한 태스크를 업스트림(Upstream) 태스크 또는 사전훈련(Pretraining)이라 하고, 사전훈련한 결과를 적용한 태스크를 다운스트림(Downstream) 태스크라 한다.

자연어처리에서 사전훈련 단계는 일반적인 언어의 의미 이해를 목적으로 모델을 훈련하며, 다운스트림 단계에서 개체인식, 번역, 문서요약, 질의응답 등 구체적 태스크를 수행한다. 전이학습은 2010년대 초반 알렉스넷 딥러닝 알고리즘 등이 등장하며 이미지 처리의 성능 개선에 많은 기여를 했다. 반면, 자연어처리에의 적용은 2017년 구글에서 트랜스포머 알고리즘을 활용한 BERT가 공개되며 본격적으로 시작되었다. 자연어처리에 사전훈련 언어모델의 적용을 통해 다양한 태스크의 성능 개선 가능성을 보임에 따라 자연어처리에 전이학습을 활용하는 것이 최근 연구 동향이다(Alyafeai et al., 2020 ; Ruder et al., 2019).

2.2 트랜스포머 기반 사전훈련 언어모델

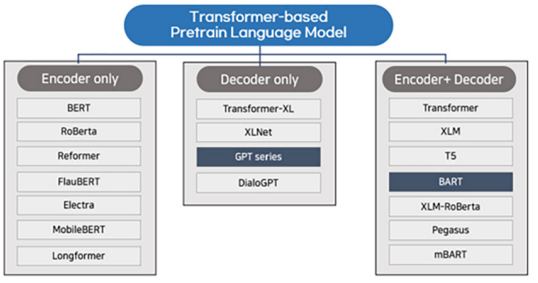
BERT는 컨텍스트를 반영한 언어 표현(Peters et al., 2018), 양방향(Bidirectional) 트랜스포머 구조(Vaswani et al., 2017), 업스트림에서 다운스트림 태스크까지 언어모델의 사전훈련부터 미세조정까지 연속적인 엔드-투-엔드(End-to-End) (Radford et al., 2018 ; Howard and Ruder, 2018) 등 선행연구의 축적으로 등장했다(박호연, 김경재, 2021 ; 박현정, 신경식, 2020). 특히 BERT를 통해 개발된

트랜스포머 공개 이후 최근 개발된 사전훈련 언어모델의 대부분은 트랜스포머 구조에 기반하고 있다(Zaken et al., 2022).

자연어처리에 트랜스포머가 효과적으로 동작하게 된 것은 크게 다음 2가지 요인에 기인한다. 먼저, 그 이전에는 단어의 순서에 따라 데이터를 순차적으로 처리하는 순환신경망 구조의 인코더와 디코더로 구성된 Seq2Seq(Sequence-to- Sequence) 언어모델이 주로 사용되었다. 인코더는 입력 문장을 하나의 단일 벡터로 표현하고, 디코더는 단일 벡터로부터 출력 문장을 생성한다(Cho et al., 2014). 문제는 순환신경망 구조에 기반한 Seq2Seq 모델이 구조적 특성상 병렬 처리를 할 수 없으므로 문장이 길어질수록 계산 속도가 느려진다는 점이다. 또한, 입력 길이가 길어질 경우 순환신경망의 고질적인 문제인 기울기 소실 문제(Vanishing Gradient Problem)로 인해 모델 성능이 하락하는 문제가 있다. 반면 트랜스포머는 언어 처리의 병렬화를 통해 대규모 데이터셋에 대한 학습이 가능함에 따라 기존의 계산 속도 저하 문제를 극복했다(Egonmwan and Chali, 2019 ; Weng et al.,2020).

다음은 셀프 어텐션(Self-attention)이다. 트랜스포머는 인코더의 출력을 디코더의 입력으로 사용하면서 인코더와 디코더만으로 전통적인 Seq2Seq 모델을 구현 가능함에 따라 기울기 소실 문제를 해결하였다(Vaswani et al., 2017).

<Figure 1>에서 보듯 인코더와 디코더 중 어떤 것을 사용하였는지에 따라 인코더만 차용한 구조, 디코더만 차용한 구조, 인코더와 디코더 모두 있는 구조로 구분할 수 있다(Howard and Ruder, 2018 ; Ruder et al., 2019). 인코더만 있는 모델에서는 자연어이해만 가능하기 때문에 자연어이해와 관련된 태스크인 질의응답, 개체명인식, 추출요약 등에만 적용할 수 있다. 반면 디코더만 있거나 인코더와 디코더 모두 있는 모델은 디코더에서 문장 생성 기능도 수행하기 때문에 기계번역이나 생성요약과 같은 자연어생성과 관련된 태스크도 적용이 가능하다. 따라서 본 연구에서는 자연어생성에 관한 태스크 수행이 가능한 대표적 트랜스포머에 기반한 모델인 GPT와 BART를 실험에 사용했다.



〈Figure 1〉 Type of Pre-trained Language Models

2.2.1 GPT(Generative Pre-Training)

GPT-1은 이전 단어들을 통해 다음 단어를 예측하도록 학습하는 일방향 구조(Unidirectional) 언어모델이다(Radford et al., 2018). 이후 공개된 GPT-2는 사전훈련 언어모델만으로 다운스트림 태스크 적용 시 미세조정이 필요 없는 범용성을 가지는 언어모델을 구축하고자 했다. 이를 위해 GPT-2는 GPT-1의 구조를 거의 대부분 따르지만 학습데이터의 크기와 매개변수의 개수 등 모델의 크기를 117배 확장하였다(Radford et al., 2019). 또한 데이터셋의 품질과 크기, 다양성을 동시에 고려해 40GB의 말뭉치로 구성된 고품질 WebText 데이터셋을 직접 구축하였다. Radford 외(2019)에 따르면 GPT-1와 비교해 모델의 크기가 커짐에 따라 다양한 벤치마크 데이터 셋에서 성능 개선을 보였다. 마지막에 공개된 GPT-3는 1,750억 개의 매개변수를 가진 초대규모 모델이다. GPT-3의 학습을 위해 크롤링(4,100억 개), 웹텍스트(190억 개),

위키피디아(30억 개) 등으로 초대규모 데이터셋을 활용했다(Brown et al., 2020). 이러한 GPT 모델들은 공통되게 이전에 등장한 단어들을 통해 다음 단어를 예측하기 위해 자기회귀 디코더(Autoregressive Decoder)를 가진다. 따라서 일방향으로 학습하는 GPT는 자연어생성에 강점을 갖는다고 알려져 있다.

SKT가 2020년 공개한 KoGPT(1.0)는 openAI의 GPT-2 모델을 대량의 한국어 말뭉치 데이터에 학습한 모델이다. 2021년에 다시 공개된 KoGPT(2.0)은 1.0에 비해 2배 이상 큰 데이터 셋을 학습에 사용했다. SKT 발표에 따르면 1.0에서는 단일 문장 생성에 최적화되었지만 2.0에서는 문맥을 유지하며 다중 문장 생성에 최적화함으로써 의미적으로 연관된 문단 생성도 가능하다고 알려져 있다. 언어모델의 평가지표 중 하나인 Perplexity에서 1.0은 45.4이었지만 2.0에서는 24.6로 성능 개선을 보였다. SKT의 KoGPT는 토크나이저(Tokenizer)로 문자 BPE(Character Byte Piece Encoding)를 사용하였고, 대화에 자주 쓰이는 이모티콘, 이모지(Emogi) 등을 추가해 토큰 인식을 향상시켰다.

한편 카카오도 GPT-3 모델을 한국어 데이터에 학습시킨 KoGPT를 2021년 말 공개했다. 동 모델은 60억개의 매개변수와 2,000억 개 토큰으로 구성되었으며, SKT와 비교해 모델의 매개변수 크기가 6배 정도 크다. 본 연구에서는 SKT와 카카오의 KoGPT 모델을 사용해 추론데이터의 특성과 사전학습 언어모델의 매개변수의 크기 간 관계를 분석해 이것이 자연어생성 성능에 미치는 영향을 확인했다.

2.2.2 BART(Bidirectional Encoder Representation from Transformer)

GPT는 앞서 언급한 것처럼 일방향으로 정보를 학습함에 따라 자연어생성에 강점이 있다. 반면 BERT는 양방향으로 작동하기 때문에 자연어 이해와 관련한 태스크에 강점을 갖는다고 알려져 있다. BERT의 핵심적인 아이디어는 매스크 언어 모델(MLM, Masked Language Model)과 다음 문장 예측(Next-sentence Prediction)이다. 매스크 언어 모델이란 모델이 랜덤하게 매스크 처리된 토큰들을 컨텍스트를 고려하여 다음 단어를 예측하는 방법이다. 2020년 Facebook이 발표한 BART는 이러한 BERT와 GPT의 장점을 결합한 모델이다(Lewis et al., 2019). BART는 Seq2Seq 구조로 구성되어 있어 자연어이해뿐만 아니라 자연어생성도 가능하다. BART는 노이즈 함수(Noise Function)을 이용해 잡음(Noise)이 추가된 데이터에서 잡음을 제거해 단어 의미를 추출하는 방식으로 정보를 학습한다.

SKT에서 2020년 공개한 KoBART는 BART 논문에서 사용된 기법을 동일하게 적용해 40GB 이상의 한국어 텍스트에 학습한 모델이다. SKT의 KoGPT와 동일하게 BART는 토크나이저로 문자열 BPE 방식을 사용하였고, 대화에 자주 쓰이는 이모티콘, 이모지 등을 추가해 토큰 인식을 개선시켰다.

2.3 자연어생성

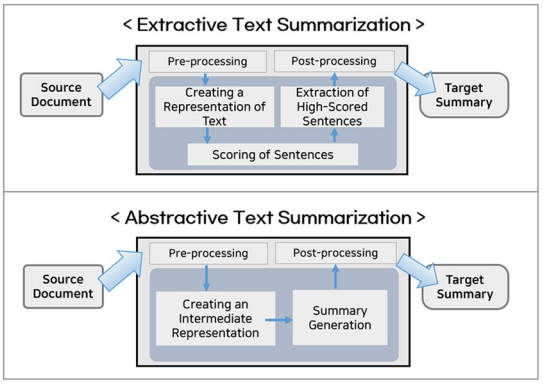
자연어생성은 원본 텍스트의 의미를 이해해야 할 뿐 아니라 자연스러운 문장 생성까지 고려해야 하므로 자연어처리에 관한 보다 고도의 기술을 필요로 한다(Alomari et al., 2022 ; Ermakova et al., 2019). 이러한 이유로 딥러닝 모델의 자연어처리 도입 이후에도 자연어이해에 비해 상대적으로 자연어생성 분야의 성능 개선 속도가 더디었으나 전이학습이 자연어처리 분야에 적용되고 관련 태스크 성능이 크게 개선되면서 최근 활발한

연구가 이뤄지고 있다(Gupta and Gupta, 2019).

한편 자연어생성의 활용 분야는 기계번역, 챗봇, 음성 인식, 이미지 캡셔닝 등에 이르기까지 다양한 영역으로 발전, 확장하고 있다. 그 중에서도 문서 요약 기술은 디지털 매체를 통해 갈수록 더 많은 텍스트 데이터가 생산됨에 따라 사용자에게 사전정보 제공과 필요한 정보를 선별하는 기능을 하기 때문에 더욱 중요해지고 있다(Chowdhary, 2020).

2.4 문서요약

문서요약이란 긴 문서 내에서 핵심적인 내용만을 요약하는 자연어처리 기술이다. 문서요약은 <Figure 2>에서 보듯 요약문의 생성방식에 따라 추출(Extractive) 요약과 생성(Abstractive) 요약으로 구분된다(윤여일 외, 2019).



〈Figure 2〉 Extractive Text Summarization and Abstractive Text Summarization

추출요약은 원문에 있는 문장들 중에서 문서의 주제를 담고 있는 주요 문장 몇 개를 선택해서 요약문을 구성하는 방법이다. 이때 문장을 선택하는 방법으로 단어의 출현 빈도 기반, 문장들 간 연결 그래프 생성하고, 각 그래프들을 그룹화한 뒤 그룹별 주요 문장을 추출하는 방법, 문장 간 유사도를 계산하여 선택할 문장을 결정하는 방법 등이 있다. 이러한 추출요약은 비교적 간단하게 구현 가능하다는 장점이 있지만 요약문의 응집도나 가독성이 다소 떨어진다는 단점이 있다.

생성요약은 문서 내용을 압축해 원문에 있는 문장이 아닌, 새로운 문장으로 요약문을 생성하는 방식이다. 따라서 추출요약과 달리 생성요약은 핵심문장으로 추출할 수 있는 문장이 존재하지 않을 때 문서의 의미를 이해하고, 새로운 문장을 생성할 수 있다는 장점이 있지만 불완전하고 부자연스러운 문장을 만들 수 있다는 단점을 가진다(El-Kassas et al., 2021). 생성요약은 요약문을 생성하는 방식에 따라 정보조합 기법, 압축기법, 재구축 기법으로 구분된다. 생성요약은 추출요약 기법보다 요약문이 효과적으로 압축되지만, 문법에 맞지 않거나 부자연스러운 문장이 생성되는 경우가 많아 기존에는 추출요약 기법에 대한 연구가 많았다(허지욱, 2019). 그러나 Seq2Seq 모델과 사전훈련 언어모델이 도입된 이후 생성요약의 성능이 크게 향상되면서 딥러닝 기반의 생성요약에 대한 연구가 활발해지고 있다(Gupta and Gupta, 2019).

한편 생성요약 분야의 주요 이슈로 사실 불일치(Factual Inconsistency) 문제와 문서의 길이가 긴 경우의 요약 품질이 저하되는 문제가 있다. 사실 불일치 문제란 모델이 생성한 요약문이 원문의 사실과 일치하지 않는 문제를 의미한다. 최근 연구에 따르면 생성 요약문의 약 30%가 문서 원문의 사실과 다른 것으로 나타났다(Cao et al, 2018 ; Falke et al., 2019 ; Gupta and Gupta, 2019). 생성요약 방식으로 생성된 요약문에 틀린 정보가 포함될 경우 경우에 따라 심각한 정보 전달

오류가 발생할 수 있어 생성요약 정보 활용의 큰 제약사항이 될 수 있다. 따라서 사실 불일치 문제를 해결하기 위한 연구들이 활발히 진행되고 있다(신정완 외, 2022 ; Dong et al., 2020 ; Kryscinski et al., 2019 ; Kryscinski et al., 2020 ; Zhu et al., 2021).

다음으로 문서의 길이가 긴 경우이다. 요약문은 원문의 내용을 잘 반영하는 동시에 중요한 정보만을 간추려야 하는데, 원문이 길어질수록 분석해야 하는 데이터의 양도 늘어나므로 계산의 복잡도가 증가한다. 또한 원문 내 핵심이 아닌 노이즈 데이터가 많이 포함된 경우 어떤 내용이 중요한지 판별하기 쉽지 않고, 원문의 주제를 벗어나 원문의 핵심 정보를 포함하지 않는 요약문을 생성하는 문제점이 있다. 따라서 이를 해결하기 위해 LDA(Latent Dirichlet Allocation)을 사용해 문서의 내용에서 중요한 부분에 더 집중해 요약문을 생성하는 등 다양한 연구방법이 제안되고 있다(Fu et al., 2020 ; Narayan et al., 2018 ; Wang et al., 2019).

3. 연구설계 및 분석방법

3.1 분석데이터

생성요약에 필요한 분석데이터는 입력문서와 이에 대응하는 요약문의 집합이다. 본 연구에서는 추론데이터의 특성에 따라 사전훈련 언어모델의 성능이 어떻게 달라지는지 확인하기 위해 한국지능정보사회진흥원(NIA)의 인공지능 학습용 데이터 구축사업(AI-Hub)을 통해 수집된 '요약문 및 레포트 생성' 데이터 셋을 사용했다. 해당 데이터 셋은 문서유형에 따라 크게 정보전달성과 창작물로 분류된다. 각 유형에 속한 문서의 종류와 종류별 건수 및 비중은 <Table 1>과 같다. 실험을 위해 본 연구는 각 문서에서 1,000개를 랜덤하게 추출해 총 1만 건의 분석데이터를 구축했다.

〈Table 1〉 Construction of Analysis Dataset

Type	Document	Number	Share
Information	News article	27,000	15%
	Briefing	20,000	11%
	History & Culture	10,000	5%
	Paper	10,000	5%
	Minute	34,000	19%
	Edit	10,000	5%
Creation	Publisher	10,000	5%
	Speech	40,000	22%
	Literature	12,000	7%
	Narration	10,000	5%
Total		**183,000**	**100%**

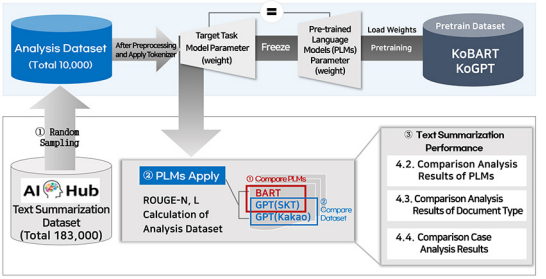
3.2 분석방법

본 연구에서 사용한 각 사전학습 언어모델에 사용된 데이터와 모델 구조는 <Table 2>와 같다. <Figure 3>은 본 연구의 연구설계 과정을 도식화한 것이다. 본 연구에서는 사전훈련 언어모델의 구조가 자연어생성 성능에 미치는 영향을 파악하기 위해 언어모델에 사용된 학습데이터와 매개변수 수가 동일한 SKT의 KoBART와 KoGPT를 비교 분석했다. 이 관점에서 (1) 사전훈련 언어모델 간 비교 (2) 문서유형 간 비교 (3) 케이스별 비교로 구분해 확인했다. 다음으로 사전훈련 언어모델 구조는 동일하나 사전훈련 모델의 매개변수 크기가 자연어생성에 미치는 영향을 파악하기 위해서 카카오와 SKT가 발표한 KoGPT를 비교 분석했다.

〈Table 2〉 Description of Korean Pre-trained Language models

Model	Source	Pretrain Dataset Sources	Number of Parameters	Nubmer of Layers	Number of Heads	Feed-forward Network Dimension
BART	SKT	Korean Wikipedia, News articles, Everyone's corpus v1.0, Blue House national petition, etc.	124M	6	16	3,072
GPT			125M	12	12	3,072
GPT	Kakao	Open Source Public Encyclopedia, News Articles, Korean Wikipedia, Internal data held by Kakao Enterprise etc.	735M	28	16	4,096

<Table 2>에서 확인할 수 있듯이 SKT와 비교해 카카오의 GPT 모델은 매개변수의 수가 6배 정도 큰 모델이다.



〈Figure 3〉 Analysis Process Flow Chart

공통적으로 추론데이터가 가진 고유의 특성에 따라 사전훈련 언어모델의 성능이 어떻게 달라지는지 확인하기 위해 타겟 태스크 적용 시 별도의 미세조정은 하지 않았다.

생성요약을 수행하기 위해 개별 문서의 원문 내 섹션 제목, 부호 등을 제외하고 완전한 문장 형태를 구성하는 부분만 추출한 뒤 토큰화를 적용했다. 토큰화에는 각 언어모델의 학습에 사용된 HuggingFace의 Transformers에서 제공하는 BartTokenizer와 AutoTokenizer를 각각 사용했다.

3.3 문서생성요약 성능지표

본 논문에서는 각 사전훈련 언어모델을 통해 생성된 추론 요약문(Inference)을 사람이 미리 작성해 놓은 참조 요약문(Reference)과 비교하여 모델의 요약 생성 성능을 평가하였다. 구체적으로 자연어생성 성능 평가에 많이 사용되는 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)을 적용했다(Kryscinski et al., 2019; Martschat and Markert, 2017). 본 연구에서는 F1 스코어 기준의 ROUGE를 적용해 평가했다.[2)]

ROUGE는 N-gram의 단어 수에 따라 ROUGE-1, ROUGE-2 등이 있다. ROUGE-1은 추론 요약문과 참조 요약문 간 겹치는 Unigram의 수를 계산하는 지표이며, ROUGE-2는 두 요약문 간 겹치는 Bigram의 수를 계산해주는 지표이다. ROUGE-L은 요약문 내 가장 긴 시퀀스(Sequence)의 Recall을 구하는 방식이다. 이때 요약문 내 단어들의

2) ROUGE는 크게 Recall, Precision, F1 스코어로 계산할 수 있다. Recall은 모델에서 생성한 추론 요약문이 길어질 경우 참조 요약문과 크게 관련이 없을지라도 참조 요약문에 단어의 대부분을 포함할 가능성이 커지기 때문에 지표가 커질 수 있다. 반면 Precision은 추론 요약문의 길이가 짧으면 상대적으로 값이 커질 수 있고, 추론 요약문에 불필요한 단어가 많아질수록 추론 요약문과 참조 요약문이 유사할지라도 지표가 낮아진다. 이러한 이유로 두 가지 경우를 모두 반영할 수 있는 F1 스코어 기준의 ROUGE가 보편적으로 사용된다.

〈Table 3〉 GPT and BART's ROUGE(F1) Results

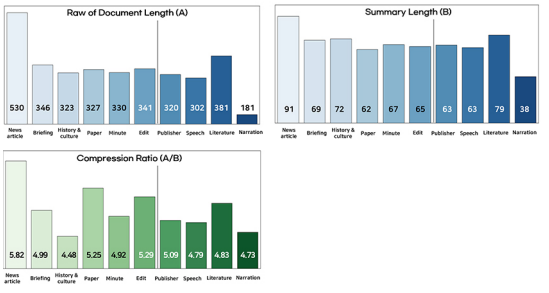
		News article	Briefing	History & culture	Paper	Minute	Edit	Publisher	Speech	Literature	Narration
ROUGE-1	GPT	0.267	0.206	0.190	0.177	0.206	0.180	0.194	0.190	0.198	0.124
	BART	0.467	0.57	0.521	0.437	0.378	0.428	0.554	0.452	0.424	0.411
ROUGE-2	GPT	0.264	0.203	0.187	0.174	0.201	0.177	0.191	0.188	0.195	0.121
	BART	0.363	0.478	0.413	0.314	0.252	0.298	0.446	0.325	0.277	0.301
ROUGE-L	GPT	0.253	0.202	0.184	0.172	0.198	0.175	0.191	0.187	0.191	0.122
	BART	0.398	0.510	0.458	0.373	0.308	0.363	0.501	0.383	0.339	0.352

연속적 매칭을 요구하지 않기 때문에 추론 요약문과 참조 요약문 내 등장하는 단어의 순서가 완전히 일치하지 않더라도 이것이 값을 낮추지 않기 때문에 보다 유연한 성능 계산이 가능하다.

4. 분석결과

4.1 기초통계량

4.1장은 추론데이터의 특성에 따라 사전훈련 요약문 생성에 앞서 문서유형별 특징을 파악하기 위해 탐색적 데이터 분석(Exploratory Data Analysis, EDA)을 수행한 결과이다.



〈Figure 4〉 Length of Analysis Dataset

<Figure 4>는 1만 건의 분석데이터에 대해 토크나이저 적용 후 문서 길이다. 문서 원문의 길이를 보면 정보전달성에서는 뉴스기사가, 창작물에서는 문학이 길었으며 나래이션은 다른 유형에 비해 매우 길이가 짧았다. 문서 원문과 요약본 길이가 비례하였지만 상대적으로 유형별로 요약본의 길이 차이는 원문보다 크지 않았다. 원문 대비 요약문 길이인 Compression ratio에서는 뉴스기사가 가장 길었고, 나래이션보다 역사문화재가 가장 짧았다. 종합적으로 정보전달성과 창작물 문서 유형 간 차이가 없어 데이터의 길이와 문서의 성격은 관련도가 낮은 것으로 나타났다.

4.2 사전훈련 언어모델간 비교분석 결과

동일 문서 내에서 사전훈련 언어모델 간 성능 비교 결과이다. <Table 3>에서 확인할 수 있듯이 모든 문서유형에서 SKT의 GPT보다 SKT의 BART 성능이 높아 사전훈련 언어모델의 학습에 동일한 데이터를 사용하더라도 인코더와 디코더를 함께 사용한 구조가 자연어생성에 보다 적합한 것으로 확인되었다. 다음으로 GPT 내에서는 모든 문서유형에서 카카오의 GPT가 SKT의 GPT보다 성능이 좋은 것으로 나타났다. 그러나 모델의 규모가 6배 차이가 있었던 것을 감안하면 상대

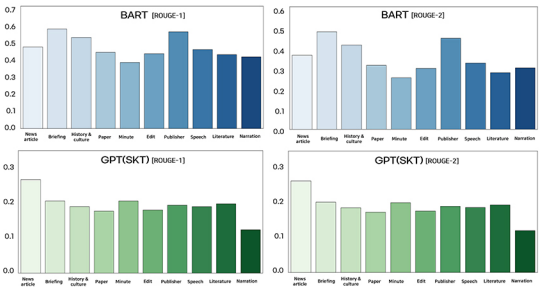
〈Table 4〉 GPT (SKT vs. Kakao) 's ROUGE(F1) Results

		News article	Briefing	History & culture	Paper	Minute	Edit	Publisher	Speech	Literature	Narration
ROUGE-1	SKT	0.267	0.206	0.190	0.177	0.206	0.180	0.194	0.190	0.198	0.124
	Kakao	0.300	0.271	0.257	0.232	0.253	0.235	0.226	0.225	0.287	0.158
ROUGE-2	SKT	0.264	0.203	0.187	0.174	0.201	0.177	0.191	0.188	0.195	0.121
	Kakao	0.295	0.265	0.251	0.228	0.248	0.230	0.222	0.221	0.283	0.154
ROUGE-L	SKT	0.253	0.202	0.184	0.172	0.198	0.175	0.191	0.187	0.191	0.122
	Kakao	0.278	0.261	0.250	0.226	0.238	0.226	0.22	0.216	0.272	0.154

적으로 그 차이는 0.1 정도로 미미했다. Radford 외 (2019) 연구에서는 GPT 모델의 학습에 사용된 데이터가 동일하다는 조건 하에 모델의 매개변수의 크기를 117M, 345M, 762M, 1,542M으로 늘려가며 매개변수의 크기가 성능 개선에 미치는 영향을 분석했다. 다양한 자연어생성 벤치마크 데이터 셋에 적용해 성능을 측정한 결과, 공통적으로 모든 데이터 셋에서 매개변수의 크기가 커질수록 성능이 개선된 것으로 나타났다. 개선 정도는 데이터 셋에 따라 달랐지만 Perplexity 지표 상 비교적 큰 폭의 개선이 있었다. 반면 본 연구는 카카오와 SKT의 GPT 모델 간 비교에서 매개변수의 크기가 자연어생성 성능에 미치는 영향을 확인할 수 있지만 카카오와 SKT의 GPT 모델의 학습에 사용된 데이터의 크기와 품질이 다르기 때문에 <Table 4>의 결과를 단순히 매개변수의 크기 차이로 설명하는 데 한계가 있다.

4.3 문서유형별 비교분석 결과

추론데이터의 특징이 자연어생성 성능에 미치는 영향을 분석하기 위해 4.3장에서는 문서의 특징을 2가지 측면에서 살펴보았다. 먼저 문서 원문과 요약문의 길이, Compression Ratio 등 데이터의 길이에 초점을 두었다.



〈Figure 5〉 Document Type's Results of ROUGE(F1) : GPT vs. BART

다음으로 문서의 특징을 정보전달성인지 창작물인지로 구분해 비교했다. 뉴스기사와 같은 정보전달성 문서는 상대적으로 육하원칙이 잘 지켜지고, 비교적 비슷한 어순을 가진 문장들로 구성되었다면, 창작물은 어순도 다양하는 등 다른 특성을 가질 것으로 판단했다.

먼저 <Table 5>는 추론데이터의 길이와 사전훈련 언어모델 간 상관관계이다. 문서 원문과 요약문 길이, GPT 간 성능지표 간 상관관계는 0.89~0.96 사이로 높은 관계를 보였으며, Compression ratio는 낮았다. 이는 <Figure 5>에서도 확인할 수 있다. <Figure 4>에서 보듯 GPT는 뉴스기사와 같이 데이터 길이가 길었던 문서 유형에서 성능이 높게 나타난 반면, 가장 짧았던 나래이션에서

〈Table 5〉 Correlation of Document Length and ROUGE(F1)

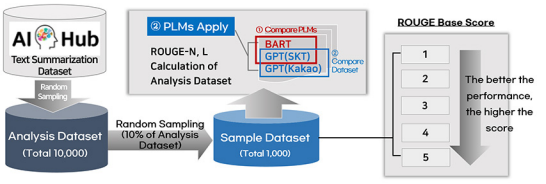
	ROUGE-1(F1)			ROUGE-L(F1)		
	BART	GPT(SKT)	GPT(Kakao)	BART	GPT(SKT)	GPT(Kakao)
Raw of Document Length (A)	0.155	0.946	0.890	0.095	0.935	0.864
Summary Length (B)	0.280	0.958	0.950	0.210	0.955	0.939
Compression Ratio (A/B)	-0.138	0.661	0.566	-0.166	0.640	0.528

성능이 가장 낮음을 확인할 수 있었다. 반면 GPT와 동일한 기준으로 해석했을 때, BART의 경우 데이터 길이와 성능 간 상관관계는 0.01~0.28 정도로 나타나 상대적으로 데이터 길이와 관련이 낮은 것으로 판단된다.

마지막으로 문서 특징을 정보전달성과 창작물로 구분한 경우이다. 정보전달성과 창작물로 문서의 성격을 구분한 경우 BART와 GPT 모두 유의미한 관계를 파악하기 어려웠다. 즉, 정보전달성과 창작물 내에서 개별 문서 간 편차가 커 문서의 특징을 파악할 수 있는 다른 기준의 검토가 필요한 것으로 판단되었다. 따라서 본 연구에서는 사전훈련 언어 모델과 문서 유형을 ROUGE-1 기준으로 구간화해 문서의 특징을 다른 측면에서 검토했다. 이에 대한 상세한 내용은 4.4장의 케이스별 비교 분석결과에 기술하였다.

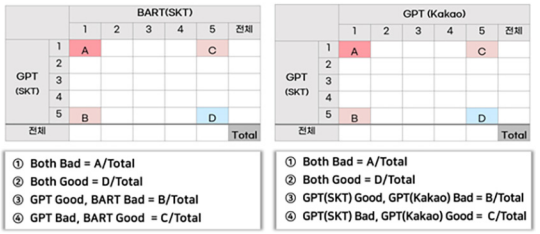
4.4 케이스 비교 분석결과

개별 문서별로 케이스 분석을 수행하기 위해 분석데이터의 10%인 100개를 각 문서유형에서 랜덤하게 추출해 총 1천 건의 샘플데이터를 구축했다. 추출된 개별 샘플데이터에 대해 ROUGE-1 값에 따라 5개로 샘플데이터를 구간화한 뒤 성능이 가장 낮은 구간에 속한 샘플 문서에는 1을, 더 좋은 구간에 속한 문서일수록 더 높은 점수를 부여했다. 따라서 가장 성능이 좋은 구간에 속한 문서는 5가 부여된다.



〈Figure 6〉 Case Analysis Process Flow Chart

이렇게 구간에 따라 점수를 주면 <Figure 7>과 같이 4가지 유형으로 샘플데이터를 분류할 수 있게 된다. 이렇게 유형을 분류한 것은 앞서 언급한 것처럼 정보전달성과 창작물 외에 추론데이터의 특성이 자연어생성 성능에 미치는 영향을 분석하기 위해 문서의 특성을 파악하기 위한 추가 분석이 필요하다고 판단했기 때문이다.



〈Figure 7〉 Classification Criteria for Case Analysis

<Figure 7>의 좌측 그림은 사전훈련 언어모델의 학습에 사용된 데이터는 동일하나 모델의 구

〈Table 6〉 Parts of Speech Share by Document Type

		News article	Briefing	History & culture	Paper	Minute	Edit	Publisher	Speech	Literature	Narration
Josa	Gwanhyeong-gyeog Josa	2%	2%	3%	3%	1%	2%	3%	1%	3%	2%
	Object Josa	4%	4%	4%	4%	3%	4%	4%	4%	4%	4%
	Bojosa	4%	3%	3%	4%	4%	4%	3%	4%	5%	3%
	Busagyeog Josa	5%	5%	6%	5%	4%	5%	5%	6%	5%	5%
	Jugyeog Josa	3%	2%	3%	3%	4%	4%	2%	4%	3%	4%
Josa Total		18%	16%	19%	19%	16%	19%	17%	19%	20%	18%
Eomi Total		9%	6%	9%	11%	11%	13%	8%	11%	12%	10%
Adverb/ Adjective	Gwanhyeongsa	1%	1%	1%	1%	2%	2%	1%	2%	2%	2%
	Adverb	3%	1%	1%	3%	5%	3%	2%	4%	4%	4%
	Adjective	1%	0%	1%	1%	2%	2%	1%	1%	2%	1%
Adverb/Adjective etc. Total		5%	2%	3%	5%	9%	7%	4%	7%	8%	7%
Verb	Verb	4%	2%	3%	5%	4%	5%	3%	4%	6%	4%
	Bojoyongeon	1%	1%	1%	1%	2%	2%	1%	2%	2%	2%
Verb Total		5%	3%	4%	6%	6%	7%	4%	6%	8%	6%
Noun	General Noun	33%	47%	36%	28%	30%	26%	39%	30%	25%	28%
	Dependent Noun	4%	5%	5%	4%	3%	4%	4%	5%	4%	4%
	Proper Noun	4%	2%	4%	4%	3%	3%	5%	1%	1%	3%
	Pronoun	1%	0%	0%	1%	2%	2%	1%	1%	2%	1%
Noun Total		42%	54%	45%	37%	38%	35%	49%	37%	32%	36%

조가 다를 경우를 비교하기 위한 것이다. 구체적으로 유형 ①은 BART와 GPT 모두에서 성능이 좋지 않는 경우, 유형 ②는 둘 다에서 성능이 좋은 경우, 유형 ③은 GPT는 좋으나 BART에서는 성능이 좋지 않은 경우, 유형 ④는 반대로 GPT는 좋지 않으나 BART에서는 좋은 경우이다. 이에 따라 각 유형을 분류한 후 ①~④유형에 속한 문서의 개수와 그 특징을 살펴보았다.

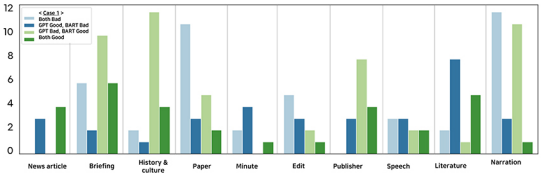
<Figure 7> 우측 그림은 카카오의 GPT와 SKT의 GPT로 유형을 분류한 것이다. 이는 사전훈련 언어모델은 동일하지만 매개변수의 수가 다른 경우의 성능을 비교하기 위한 목적이다. 유형의 분류는 GPT와 BART를 비교한 것과 동일한 기준을 적용했다.

먼저, 모든 문서 유형에서 공통되게 BART와 GPT 중 어느 한쪽만 좋은 경우가 많아 사전훈련 언어모델의 구조가 문서요약 성능에 미치는 영향이 큰 요인 것으로 확인되었다. 두번째로, 보고서와 나래이션은 BART와 GPT 모두 성능이 좋지 않는 문서 수가 다른 유형에 비해 압도적으로 많았다. 다음으로 보도자료와 역사문화재, 간행물은 다른 문서유형에 비해 GPT의 성능은 좋지 않으나 BART의 성능은 좋은 것으로 나타났다. 반대로 문학은 GPT의 성능은 좋으나 BART

〈Table 7〉 Correlation of Document's Parts of Speech Share and ROUGE(F1)

	ROUGE-1(F1)			ROUGE-L(F1)		
	BART	GPT(SKT)	GPT(Kakao)	BART	GPT(SKT)	GPT(Kakao)
Josa	-0.290	-0.181	0.009	-0.331	-0.198	0.023
Eomi	-0.829	-0.247	-0.141	-0.839	-0.274	-0.167
Adverb/Adjective etc.	-0.916	-0.194	-0.191	-0.922	-0.220	-0.250
Verb	-0.843	-0.233	-0.089	-0.870	-0.262	-0.125
Noun	0.904	0.300	0.206	0.917	0.335	0.246

의 요약 생성의 성능은 떨어졌다. SKT의 BART와 GPT에 대한 유형별 문서의 수(비중)를 정리한 것이 <Figure 8>이다. 구체적으로 유형별 분류에 따른 차이가 발생한 원인을 분석하기 위해 샘플데이터에 대한 형태소 분석을 수행했다. 이는 문서를 구성하는 단어의 품사 빈도 수가 문서 성격과 관련이 있을 것이라 판단했기 때문이다.

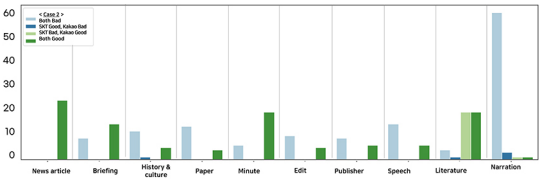


〈Figure 8〉 BART and GPT's Case Analysis ROUGE-1(F1) Results

형태소 분석기마다 각 장단점이 있지만 연산 속도와 형태소 분석결과의 품질 등을 종합적으로 고려해 본 연구에서는 Mecab을 사용했다. Mecab은 총 43개의 품사 태그로 구성되는데, 본 연구에서는 샘플데이터에 대한 형태소 분석 후 빈도수 상위 20개의 품사만 남겼다. 다음으로 관련 품사별로 그룹화한 후 품사에 따라 조사, 어미, 형용사/부사, 동사, 명사로 구분했다. 문서유형별로 상위 빈도 20개 품사에 대해 비중을 정리한 것이 <Table 6>이다.

<Table 7>은 개별 문서에서 차지하는 상위 빈도 품사 비중과 사전훈련 언어모델의 ROUGE-1 간 상관관계를 계산한 결과이다. 분석결과, BART의 경우 어미와 형용사/부사, 동사가 많은 문서일수록 성능이 떨어진 반면 명사가 많을수록 성능이 좋은 것으로 나타났다. 문서의 특징을 품사의 구성에 따라 파악 가능하다고 할 때, 이러한 결과는 BART가 문서의 특징과 관련이 있다는 것을 시사한다. ROUGE-L에서도 ROUGE-1과 유사한 결과를 보였는데, 상관관계 정도는 ROUGE-L이 ROUGE-1 보다 더 크게 나타났다. 이는 ROUGE-L이 문서에서 가장 긴 문장을 기준으로 계산되는 지표이기 때문에 문서 내 품사의 비중이 생성 요약 성능에 미치는 정도가 더 크게 나타난 것으로 보인다.

반면 GPT는 상대적으로 품사의 비중과 생성 요약의 성능 간 관계가 낮았다. BART와 동일한 기준을 적용해 해석하면, 이는 GPT가 문서의 성격보다는 데이터의 길이와 관련이 있는 모델로 볼 수 있다. 이 결과들을 종합하면 BART는 추론 데이터의 성격에, GPT는 데이터의 길이에 영향을 크게 받는 특성을 가진 언어모델로 판단할 수 있다.



〈Figure 9〉 GPT (SKT vs. Kakao) 's Case Analysis ROUGE-1(F1) Results

다음으로 사전훈련 언어모델 구조는 동일하나 사전훈련 모델의 매개변수 크기가 다른 경우이다. 앞서 언급한 것처럼 카카오가 SKT의 GPT에 비해 6배 정도 큰 모델이다. 먼저, <Figure 9>에서 보듯 것처럼 GPT의 경우 SKT나 카카오 중 어느 한쪽만 좋은 문서 유형은 거의 없어서 사전학습 언어모델의 매개변수 수가 자연어생성 성능에 미치는 영향은 미미한 것으로 보인다. 다음으로 나래이션에서는 GPT 2개 모두에서 성능이 좋지 않는 문서 수가 다른 유형에 비해 압도적으로 많았다. 이는 GPT가 데이터 길이에 영향을 많이 받는데, 나래이션의 문서 길이가 다른 유형에 비해 짧았기 때문에 이와 같은 결과가 나온 것으로 해석된다.

5. 결론 및 시사점

본 연구는 다운스트림 태스크 목적에 따라 어떠한 사전훈련 언어모델구조가 보다 적합한지를 실증분석했다. 이를 위해 BART와 GPT를 선택했으며, 추론데이터의 특징에 따라 사전훈련 언어모델의 성능이 어떻게 달라지는지 분석하기 위해 10가지 유형의 생성요약 문서에 적용했다. 그 결과, 모든 문서유형에서 인코더와 디코더가 모두 있는 BART의 성능이 디코더만 있는 GPT보다 더 좋은 성능을 보였다. 이는 자연스러운 자연어생성을 위해 단순히 디코더가 수행하는 자연어생성 기능뿐 아니라 트랜스포머의 인코더가 수행하는 '자연어이해' 기능 역시 중요하다는 것을 의미한다. 또한 사전훈련에 사용된 매개변수의 크기보다 다운스트림 태스크의 목적에 맞는 언어모델 구조의 선택이 가장 중요하다는 것을 시사한다. 따라서 수행하고자 하는 태스크가 자연어생성일 경우 사전훈련 언어모델의 학습이나 추론에 우선적으로 BART와 같이 인코더와 디코더 모두 있는 언어모델 구조를 추천한다. 한편 카카오의 사전훈련 언어모델인 GPT의 크기가 SKT의 GPT에 비해 6배 큰 것과 비교하면 모든 문서 유형에서 SKT나 카카오의 성능 차이는 매우 미미했다. 그러나 카카오와 SKT의 GPT 모델의 학습에 사용된 데이터의 크기와 품질이 다르기 때문에 이러한 결과를 직접적으로 매개변수의 크기 차이로 해석하는 데 한계가 있다.

다음으로 추론데이터의 특성과 사전훈련 언어모델의 성능 간 관계를 살펴보면 GPT는 데이터의 길이에 자연어생성 성능이 비례한 것으로 나타났다. 그러나 가장 길이가 긴 문서에 대해서도 GPT보다 BART의 성능이 높아 앞서 언급한 것처럼 다운스트림 태스크의 목적에 맞는 언어모델 구조의 선택이 가장 우선적 고려 대상인 것으로 판단된다.

한편 BART의 경우 GPT와 달리 데이터의 길이보다는 문서의 특징과 관련이 있었다. 본 연구에서는 개별 문서의 특징을 문서 내 품사의 비중으로 파악했다. 그 결과, BART의 경우 어미와 형용사/부사, 동사가 많은 문서일수록 성능이 떨어진 반면 명사가 많을수록 성능이 좋은 것으로 나타났다. 이는 형용사와 동사, 조사와 어미 같이 상대적으로 명사에 비해 단어의 원형 변형이 큰 품사가 많은 문서일수록 문장 생성이 어렵다는 것을 시사한다. 또한 영어에 비해 조사가 발

달한 교착어의 특성을 가진 한국어의 경우, BART를 적용해도 상대적으로 성능이 떨어질 수 있기 때문에 이러한 특징을 잘 고려할 경우 성능 개선의 여지가 클 수 있다.

본 연구의 한계점이다. 본 연구는 ROUGE 지표를 적용해 생성요약의 성능을 판단했는데, 한국어 특성을 고려했을 때 ROUGE 지표가 가지는 문제점이 있다. 먼저, ROUGE는 동음이의어에 대한 평가가 어렵다. 따라서 문서 원문과 요약문에서 동일한 의미를 가진 대상을 지칭하나 다른 단어로 표현을 할 경우 성능이 낮게 보이고, 반대로 '사과'와 같이 동일한 단어로 표현되나 전혀 의미가 다를 경우 잘못된 요약문임에도 성능이 높게 나타날 수 있다. 한국어가 다른 언어에 비해 동음이의어가 발달한 것을 감안하면 ROUGE 지표만으로 자연어생성 성능을 판단하는 것에 한계가 있을 수 있다.

다음으로 문장 성분의 배열 순서인 어순에 대한 평가이다. 한국어는 어순이 정해져 있기는 하지만, 주어의 동작, 상태, 성질 등을 설명하는 서술어 이외의 다른 문장 성분들이 문장 안에서 자유롭게 이동할 수 있다. 따라서 조사만 잘 붙어 있으면 한국어의 문장은 어순이 중요하지 않을 수 있다. 이는 한국어가 문장 성분의 관계를 설명하는 조사가 발달된 교착어라는 특징에 기인한다. 반면 ROUGE-2 이상인 경우 단어의 순서까지 고려해 자연어생성 성능을 판단하기 때문에 한국어 적용에 적합하지 않을 수 있다. 반면 영어는 어순이 다를 경우 문장이 어색하므로 어순에 대한 부분도 자연어생성의 성능을 판단하는 데 중요한 고려 요소일 수 있다.

마지막으로 추후 연구과제이다. 본 연구에서는 문서의 특징을 문서 내 품사의 비중으로 파악해 이것과 사전훈련 언어모델의 자연어생성 성능 간 관계를 분석했다. 품사의 비중 외에 단어들의 등장 순서를 통해 문서의 특성을 파악하고, 이것을 한국어의 자연어생성 성능 지표와 연관해 분석한다면 기존에 공개된 BART나 GPT의 구조를 한국어 특성에 적합하게 수정한 모델이나 보완적 자연어생성 지표 개발이 가능할 것이다.

참고문헌(References)

Alomari, A., Norisma, I., Sabri, A. Q. M., and I. Alsmadi, "Deep reinforcement and transfer learning for abstractive text summarization: A review," *Computer Speech & Language*, Vol.71 (2022), 101276.

Alyafeai, Z., AlShaibani, M. S., and I. Ahmad, "A survey on transfer learning in natural language processing," *In ArXiv, vol. abs/2007.04239*, (2020).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and D. Amodei, "Language models are few-shot learners," *In Advances in Neural Information Processing Systems*, Vol.33(2020), 1877～1901.

Cao, Z., Wei, F., Li, W., and S. Li, "Faithful to the original: Fact aware neural abstractive summarization," *In Proceedings of the Thirty Second AAAI Conference on Artificial Intelligence*, (2018), 4784～4791.

Cho, K., Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Y.

Bengio, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” *In Proceedings of the Empirical Methods in Natural Language Processing*, (2014), 1724～1734.

K.R. Chowdhary, Fundamentals of Artificial Intelligence. *Springer*, 2020.

Deng, S., Zhang, N., Yang, J., Ye, H., Tan, C., Chen, M., Huang, S., Huang, F., and H. Chen, “LOGEN: Few-shot Logical Knowledge-Conditioned Text Generation with Self-training,” *In CoRR abs/ 2112.01404. arXiv:2112.01404,* https://arxiv.org/ abs/2112.01404, (2021).

Devlin, J., Chang, M. W., Lee, K., and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *In arXiv:1810.04805,* https://arxiv.org/abs/ 1810.04805, (2018).

Dong, Y., Wang, S., Gan, Z., Cheng, Y., Cheung, J. C. K., and J. Liu, “Multi-fact correction in abstractive text summarization,” *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, (2020), 9320～ 9331.

Egonmwan, E., and Y. Chali, “Transformer and seq2seq model for Paraphrase Generation,” *In Proceedings of the 3rd Workshop on Neural Generation and Translation, China. Association for Computational Linguistics,* (2019), 249～255.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., and H. K. Mohamed, “Automatic Text Summarization: A Comprehensive Survey,” *Expert Systems with Applications*, Vol.165(2021), 113679.

Ermakova, L., Cossu, J.V., and J. Mothe, “A survey on evaluation of summarization methods,” *Information Processing and Management*, Vol.56, No.5(2019), 1794～1814.

Falke, T., Ribeiro, L. F. R., Utama, P. A., Dagan, I., and I. Gurevych, “Ranking generated summaries by correctness: An interesting but challenging application for natural language inference,” *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (2019), 2214～2220.

Fu, X., Wang, J., Zhang, J., Wei, J., and Z. Yang, “Document Summarization with VHTM: Variational Hierarchical Topic-Aware Mechanism,” *In Proceedings of the Association for the Advancement of Artificial Intelligence*, (2020), 7740～7747.

Guan, J., Mao, X., Fan, C., Liu, Z., Ding, W., and M. Huang, “Long Text Generation by Modeling Sentence-Level and Discourse-Level Coherence,” *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol.1(2021), 6379～6393.

Gupta, S., and S. K. Gupta, “Abstractive summarization: An overview of the state of the art,” *Expert Systems with Applications*, Vol.121, No.1(2019), 49～65.

Heu, J.-U, “Analysis and Comparison of Query focused Korean Document Summarization using Word Embedding,” *The Journal of the Institute of Internet, Broadcasting and Communication*, Vol.19, No.6(2019), 161～167.

Kryscinski, W., Keskar, N. S., McCann, B., Xiong, C., and R. Socher, “Neural text summarization: A critical evaluation,” *In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, (2019), 540～551.

Kryscinski, W., McCann, B., Xiong, C., and R. Socher, “Evaluating the factual consistency of abstractive text summarization,” *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, (2020), 9332～9346.

Howard, J., and S. Ruder, “Universal Language Model Fine-tuning for Text Classification,” *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol.1(2018), 328～339.

Huang, D., Cui, L., Yang, S., Bao, G. Wang, K., Xie, J. and Y. Zhang, “What have we achieved on text summarization?,” *In Proceedings of Conference on Empirical Methods in Natural Language Processing,* (2020), 446～469.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and L. Zettlemoyer, “Bart: Denoising sequence-to-sequencepre-training for natural language generation, translation, and comprehension.” *In arXiv preprint arXiv:1910.13461*, (2019).

Li, J., Tang, T., Nie, J.Y., Wen, J. R., and W. X. Zhao, “Learning to Transfer Prompts for Text Generation,” *In arXiv preprint arXiv:2205.01543, http://arxiv.org/abs/2205.01543*, (2022).

Lin, B. Y., Zhou., W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y., and X. Ren, “CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning,” *In Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing*, (2020), 1823～1840.

Liu, Y., Wan, Y., He, L., Peng, H., and P. S. Yu, “KG-BART: Knowledge Graph-Augmented BART for Generative Commonsense Reasoning,” *In Proceedings of the Association for the Advancement of Artificial Intelligence*, Vol.35 (2021), 6418～6425.

Lu, X., West, P., Zellers, R., Bras, R. L., Bhagavatula, C., and Y. Choi, “NEUROLOGIC DECODING: (Un)supervised Neural Text Generation with Predicate Logic Constraints,” *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (2021), 4288～4299.

Luo, Y., Lu, M., Liu, G., and S. Wang, “Few-Shot Table-to-Text Generation with Prefix-Controlled Generator,” *In THE 29TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, arXiv:2208.10709,* (2022).

Martschat, S., and K. Markert, “Improving ROUGE for Timeline Summarization,” *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol.2(2017), 285～290.

Narayan, S., Cohen, S., and M. Lapata, “Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization,” *In Proceedings of the Empirical Methods in Natural Language Processing*, (2018), 1797～1807.

Park, H-y, and K.-j., Kim, “Recommender system using BERT sentiment analysis,” *Journal of Intelligence and Information Systems*, Vol.27, No.2(2021), 1～15.

Park, H-j, and K.-s., Shin, “Aspect-Based Sentiment Analysis Using BERT: Developing Aspect Category Sentiment Classification Models,” *Journal of Intelligence and Information Systems*, Vol.26, No.4(2020), 1～25.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and L. Zettlemoyer, “Deep Contextualized Word Representations,”

In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1(2018), 2227～2237.

Radford, A., Narasimhan, K., Salimans, T., and I. Sutskever, "Improving language understanding by generative pre-training," *In https://s3-us-west-2. amazonaws. com/openaiassets/research-covers/languageunsupervised/language understanding paper*, (2018).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, 1(8):9(2019).

Ruder, S., Peters, M. E., Swayamdipta, S., and T. Wolf, "Transfer learning in natural language processing," *In Proceedings of NAACL-HLT: Tutorials*, (2019), 15～18.

Shin, J., Noh, Y., Song, H.-J, and S., Park, "Solving Factual Inconsistency in Abstractive Summarization using Named Entity Fact Discrimination," *Journal of KIISE*, Vol.49, No.3(2022), 231～240.

Sridhar R., and D. Yang, "Explaining Toxic Text via Knowledge Enhanced Text Generation," *In North American Chapter of the Association for Computational Linguistics*, (2022), 811～826.

Tang, T., Li, J., Zhao, W. X., and J. R. Wen, "Context-Tuning: Learning Contextualized Prompts for Natural Language Generation," *In arXiv preprint arXiv:2201.08670*, (2022).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, Vol.30(2017).

Wang, Y., Li, J., Pong, H., King, I., Lyu, M., and S. Shi, "Topic-Aware Neural Key phrase Generation for Social Media Language," *In Proceedings of the Association for Computational Linguistics*, (2019), 2516～2526

Weng, R., Yu, H., Huang, S., Cheng, S., and W. Luo, "Acquiring Knowledge from Pre-Trained Model to Neural Machine Translation," *In Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, Vol.34(2020), 9266～9273.

Wu, S., Zhao, X., Yu, T., Zhang, R., Shen, C., Liu, H., Li, F., Zhu, H., Luo, J., Xu, L., and X. Zhang, "Yuan 1.0: Large-Scale Pre-trained Language Model in Zero-Shot and Few-Shot Learning," *In ArXiv, abs/2110.04725*, (2021).

Yun. Y. Ko, E, and N. Kim, "Subject-Balanced Intelligent Text Summarization Scheme," *Journal of Intelligence and Information Systems*, Vol.25, No.2(2019), 141～166.

Zaken, E. B., Ravfogel, S., and Y. Goldberg, "BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models," *In Proceedings of the Association for Computational Linguistics*, (2022).

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., and L. Zettlemoyer, "OPT: Open Pre-trained Transformer Language Models," *In arXiv preprint arXiv:2205.01068*, (2022).

Zhu, C., Hinthorn, W., Xu, R., Zeng, Q., Zeng, M., Huang, X., and M. Jiang, "Enhancing factual consistency of abstractive summarization," *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (2021), 718～733.

Abstract

A Study of Pre-trained Language Models for Korean Language Generation

Minchae Song* · Kyung-shik Shin**

This study empirically analyzed a Korean pre-trained language models (PLMs) designed for natural language generation. The performance of two PLMs – BART and GPT – at the task of abstractive text summarization was compared. To investigate how performance depends on the characteristics of the inference data, ten different document types, containing six types of informational content and creation content, were considered. It was found that BART (which can both generate and understand natural language) performed better than GPT (which can only generate). Upon more detailed examination of the effect of inference data characteristics, the performance of GPT was found to be proportional to the length of the input text. However, even for the longest documents (with optimal GPT performance), BART still out-performed GPT, suggesting that the greatest influence on downstream performance is not the size of the training data or PLMs parameters but the structural suitability of the PLMs for the applied downstream task. The performance of different PLMs was also compared through analyzing parts of speech (POS) shares. BART's performance was inversely related to the proportion of prefixes, adjectives, adverbs and verbs but positively related to that of nouns. This result emphasizes the importance of taking the inference data's characteristics into account when fine-tuning a PLMs for its intended downstream task.

Key Words : Pre-train Language Model, Transformer, Abstractive text summarization, BART, GPT

Received : November 14, 2022 Revised : December 11, 2022 Accepted : December 14, 2022
Corresponding Author : Kyung-shik Shin

* Nonghyup, The Department of Big Data Strategy
** Corresponding Author: Kyung-shik Shin
School of Business, Ewha Womans University
52 Ewhayeodae-gil, Seodaemun-gu, Seoul, 120-750, Korea
Tel: +82-02-3277-2799, Fax: +82-2-3277-2766, E-mail: ksshin@ewha.ac.k

저 자 소 개

송민채

이화여자대학교에서 빅데이터분석학 박사학위를 취득하였다. 박사학위 취득 후 NH금융지주 금융연구소를 거쳐 현재 농협중앙회 디지털혁신실에 재직 중이다. 주요 연구분야는 자연어처리와 유통 및 금융 빅데이터의 활용 등이다.

신경식

현재 이화여자대학교 경영대학 경영학부 교수로 재직 중이다. 연세대학교 경영학과를 졸업하고, 미국 George Washington University에서 MBA, KAIST에서 경영공학 Ph.D.를 취득하였다. 주요 연구분야는 데이터 마이닝과 비즈니스 인텔리전스, 빅데이터 분석, 비즈니스 애널리틱스(Business Analytics), 인공지능 응용과 지식공학 등이다