언어 모델의 뉴스 도메인 요약 성능 비교 분석

류상원°, 김윤수°◇, 이근배°◇
°인공지능대학원, 포항공과대학교
◇컴퓨터공학과, 포항공과대학교
{ryusangwon, yunsu.kim, gblee}@postech.ac.kr

Comparative Analysis of Language Model Performance in News Domain Summarization

Sangwon Ryu°, Yunsu Kim°◇, Gary Geunbae Lee°◇
°Graduate School of AI, POSTECH
◇Department of Computer Science and Engineering, POSTECH

요 약

본 논문에서는 기존의 요약 태스크에서 주로 사용하는 인코더-디코더 모델과 디코더 기반의 언어 모델의 성능을 비교한다. 요약 태스크를 평가하는 주요한 평가 지표인 ROUGE 점수의 경우, 정답 요약문과 모델이 생성한 요약문 간의 겹치는 단어를 기준으로 평가한다. 따라서, 추상적인 요약문을 생성하는 언어 모델의 경우 인코더-디코더 모델에 비해 낮은 ROUGE 점수가 측정되는 경향이 있다. 또한, 최근 연구에서 정답 요약문 자체의 낮은 품질에 대한 문제가 되었고, 이는 곧 ROUGE 점수로 모델이 생성하는 요약문을 평가하는 것에 대한 신뢰도 저하로 이어진다. 따라서, 본 논문에서는 언어 모델의 요약 성능을 보다 다양한 관점에서 평가하여 언어 모델이 기존의 인코더-디코더 모델보다 좋은 요약문을 생성한다는 것을 보인다.

주제어: 자연어 처리, 요약, 언어 모델

1. 서론

최근 대규모 언어 모델이 큰 관심을 끌면서 다양한 태스크에서 대규모 언어 모델을 적용하는 연구들이 늘어나고 있다. 기존에는 요약 태스크에서 언어 모델과 같은 자기 회귀 모델은 인코더-디코더 모델에 비해 좋지 않은 성능을 보인다고 알려져 있었지만 [1], 대규모 언어 모델인 GPT-3 [2]가 등장함에 따라 요약 태스크에도 대규모 언어 모델을 사용하는 연구들이 많이 등장하였다. 요약 태스크를 평가하는 대표적인 방법인 ROUGE 점수로 인코더-디코더 모델과 언어 모델의 요약 성능을 평가하였을 때, 대규모 언어 모델의 제로 샷 성능이 파인 튜닝된 인코더-디코더 모델의 성능에 한참 미치지 못하였다.

하지만 최근 연구 [3] 에 따르면 ROUGE 점수와 다르게 인간 평가 관점에서는 오히려 대규모 언어 모델의 제로 샷 성능이 파인 튜닝된 인코더-디코더 모델의 성능보다 좋은 것으로 드러났다. 구체적으로, GPT-3.5 [4]가 생성한 요약문의 경우 정답 요약문보다 고품질의 요약문을 생성하였고 [5, 6] 심지어 전문가가 만든 요약문보다도 좋은 요약문을 생성할 수 있다고 한다 [3]. ROUGE 점수는 본문과 요약문의 겹치는 단어들의 수를 계산하는 방식이기에, 본문에는 없는 동의어를 주로 생성하는 대규모 언어 모델에서는 ROUGE 점수가 대체로 낮게 측정된다.

또한 최근 제기된 요약 태스크의 정답 요약문의 품질이 좋지 않다는 주장[5, 6, 7, 8]은 이러한 저품질의 정답 요약문과 요약 모델이 생성한 요약문 사이에서 측정되는 ROUGE 점수의 신뢰도 저하를 야기한다.

이러한 이유로 언어 모델의 요약 성능이 과소 평가되고 있다고 판단되었고, 이에 언어 모델의 요약 성능을 인코더-디코더 모델과 다양한 평가 지표를 통해 비교한다. 본 논문에서는 대규모 언어 모델인 GPT-4 [9]와 작은 언어 모델인 Llama2-7B [10] 그리고 인코더-디코더 모델인 BART [11]의 성능을 다양한 각도에서 비교하고 각 모델이 생성하는 요약문의 특징을 분석한다. 따라서, 본 논문에서는 다양한 평가 방법을 통해 언어 모델이 인코더-디코더 모델에 비해 더 좋은 요약문을 생성한다는 것을 보인다.

2. 본문

기존의 연구들에서 요약 태스크를 진행할 때는 대부분 인코더-디코더 모델을 파인 튜닝해서 사용해왔다 [11, 12]. 이는, 해당 구조의 모델이 인코더를 통해 본문의 정보를 디코딩 과정에서 사용할 수 있는 장점으로 인해 높은 ROUGE 점수로 이어지기 때문이다.

하지만 최근 들어 GPT-3를 필두로 대규모 언어 모델이 큰 관심을 받게 되면서, 요약 태스크에서도 대규모 언어 모델을 사용하는 시도가 늘어나고 있다 [7]. 인코더를 사용하지 않는 대규모 언어 모델의 경우 본문 내용을 디코딩 과정에서 사용하지 않아 추상적인 요약문을 생성하여 인코더-디코더 모델에 비해 낮은 ROUGE 점수를 보인다. 구체적으로, Llama-70B [13], GPT-4와 같이 대규모 언어 모델의 크기가 점점 커지면서 ROUGE 점수가 향상되었음에도 기존 인코더-디코더 모델에 못 미치는 점수를 보인다 [3].

하지만, 인코더-디코더 모델 중 가장 높은 ROUGE 점수를

보이는 BRIO [12]가 인간 평가에서 대규모 언어 모델보다 성능이 떨어진다는 연구 결과 [5]가 제기되었다. 이는, ROUGE 점수가 인코더-디코더 모델에서 더 높더라도 인간 평가 관점에서는 오히려 언어 모델의 성능이 더 좋을 수 있다는 것을 암시하며, 이는 곧 인코더의 필요성에 대한 논의로 이어진다.

본 논문에서는 기존 인코더-디코더 모델과 디코더 기반의 언어 모델이 생성하는 요약문에 대해 분석하고 각 모델이 가지는 장단점에 대해 비교하며 언어 모델이 인코더-디코더 모델보다 효과적으로 뉴스 데이터 셋을 요약한다는 것을 보인다. 먼저, 인코더를 통해서 본문의 정보를 파악하는 것이 ROUGE 점수를 향상하는 데에는 도움을 줄 수 있지만 실제 요약을 할 때 요약문의 품질 향상에 도움을 주지 못할 수 있다. 본문의 단어나 구절을 그대로 요약문으로 사용하면 ROUGE 점수는 높아지지만, 요약문에 적합하지 않은 단어들을 사용하기 때문에 요약문의 흐름이 자연스럽지 않다. 3.3에서 인간 평가를 통해 ROUGE 점수와 사람이 평가하는 요약 점수를 비교하며 자세하게 알아본다.

또한 뉴스 데이터처럼 본문의 내용이 길 때, 본문의 중요한 내용만을 요약문에 포함해야 하지만 인코더-디코더 모델의 경우 중요하지 않은 내용도 요약문에 포함하는 경우가 빈번하게 발생하기 때문에 요약문의 흐름이 어색하다. 반면, 언어 모델의 경우 각 본문의 내용을 그대로 사용하는 것이 아니라 추상적 요약을 하는데, 인코더가 없어 본문에서 사용하는 단어 혹은 구절을 그대로 사용하지 않고 요약에 더 적합한 동의어들을 사용한다. 따라서, 본문의 내용을 그대로 사용하는 경향이 있는 인코더-디코더 모델의 경우 생성한 요약문과 본문 사이에 겹치는 단어의 수가 많아서 ROUGE 점수는 높지만, 오히려 언어 모델이 만드는 요약문이 사람이 읽기에는 더 적합할 수 있다.

다음으로, 최근 요약 태스크의 정답 요약문의 품질이 떨어진다는 연구들이 제기되고 있다. 요약 태스크의 대표적인 데이터 셋인 XSum의 경우 요약문의 품질이 많이 떨어지기 때문에[5] 낮은 품질의 데이터 셋으로 파인 튜닝을 하는 것이 요약 모델의 성능 향상에 도움을 주지 못한다. 마찬가지로 ROUGE 점수의 경우 정답 요약문과 모델이 생성한 요약문의 겹치는 단어를 기준으로 점수를 채점하기 때문에, 낮은 품질의 정답 요약문과 단어를 비교하는 것이 실제 모델의 성능을 제대로 평가하지 못할 수 있다.

마지막으로, 인코더-디코더 모델이 요약하는 방식이 사람이 요약하는 방식과 다르다. 사람이 뉴스를 요약할 때는 요약문에 본문의 문구를 그대로 이용하지 않는다. 또한 본문에서 중요하지 않은 내용은 요약문에 포함하지 않는다. 하지만 인코더-디코더 모델의 경우 본문에 있는 구절을 그대로 사용하는 경우가 발생하고, 본문에서 중요하지 않은 부분도 요약문으로 출력하는 경우가 발생하기 때문에, 이러한 부분에 있어 언어 모델이 인코더-디코더 모델에 비해 사람이 요약하는 방식과 더 유사하다고 할 수 있다.

2.1 모델 별 요약문 비교

해당 장에서는 각 모델이 생성하는 요약문에 대해서 비교 및 분석한다. 인코더-디코더 모델은 인코더를 사용하기 때문에 본문의 세부적인 내용을 요약문에 포함할 수 있다. 하지만, 중요하지 않은 내용을 요약문에 포함하거나 본문에 있는 내용을 그대로 추출하는 경우가 발생한다. 반면에, 언어 모델은 본문에 있는 내용을 그대로 추출하는 것이 아니라 대부분 문장을 새롭게 생성하기 때문에 더 추상적인 형태를 띤다. 또한 언어 모델은 인코더-디코더 모델이 생성한 요약문에 비해 세부적인 내용들을 포함하지 않는다.

표 1에서 각 모델이 생성한 요약문을 보면, 인코더-디코더 모델인 BART의 경우 요약문이 본문의 핵심 내용을 중심으로 요약문을 생성하지 않았지만 디코더 기반의 언어 모델인 GPT-4가 생성한 요약문은 핵심 내용을 중심으로 요약문에 적절한 동의어들을 사용하며 사람이 요약하는 것과 유사하게 요약하였다. 여러 예시에서도 이러한 현상이 비슷하게 나타났는데, BART의 경우 문장 간의 연결이 매끄럽게 이어지지 않았지만 GPT-4와 Llama2-7B의 경우 각 문장이 매끄럽게 이어지는 경향이 있었다.

2.2 낮은 품질의 뉴스 데이터 셋 정답 요약문

뉴스 데이터 셋에서 주로 사용하는 데이터 셋인 CNN/DM과 XSum의 정답 요약문은 모두 품질이 좋지 않고 그중에서도 특히 XSum의 정답 요약문의 품질이 좋지 않다는 것이 밝혀졌다 [5]. XSum 데이터 셋의 정답 요약문의 경우 BBC 뉴스의 첫 번째 단락에서 굵은 글씨로 표시된 문장을 정답 요약문으로 사용하고 있고 요약문이 한 줄로 구성되었다. 따라서 해당 요약문은 본문의 핵심 내용을 제대로 요약하고 있지 않다.

마찬가지로 [7]에 따르면 CNN/DM 데이터 셋의 정답 요약문의 품질도 좋지 않다. 뉴스에 기본적으로 포함되어야 하는 요소들이 요약문에 포함이 되어있어야 좋은 뉴스 요약문이라 할 수 있는데, CNN/DM 데이터 셋의 정답 요약문에는 [14]에서 주장하는 뉴스의 핵심 요소들이 포함이 되어있지 않다. 표 1에서 정답 요약문은 Floyd Mayweather와 Manny Pacquiao의 경기에 관해 서술하고, 딸인 Iyanna Mayweather에 관한 내용이 간단하게 서술하고 있다. 하지만 본문 내용에는, 오히려 경기에 관한 이야기 보다 Iyanna Mayweather가 그녀의 아버지와 함께 훈련하며 느끼는 감정이 주요 내용이다. 따라서, 이러한 저품질 정답 요약문을 통한 파인 튜닝이나, 저품질 요약문과의 ROUGE 점수는 실제 요약문을 평가하는 데 도움이 되지 않는다.

표 1. 모델별 요약문 비교 분석

	요약문
본문	As the countdown continues to Floyd Mayweather's mega-fight with Manny Pacquiao in Las Vegas on May 2, The Money Man's daughter Iyanna Mayweather has shared her thoughts about life in training with her champion father. Mayweather vs Pacquiao will generate revenue upwards of $300 million in what is being billed as the most lucrative bout in boxing history and, ahead of the May showdown, Iyanna Mayweather offered some insight into her dad's intense training regime. 'When I watch my dad train, it's inspiring to me,' she said. Iyanna Mayweather has been spending time in her father Floyd's training camp . Iyanna watches on as her champion dad gets through another gruelling training session . Iyanna says she is amazed by her dad's work ethic in the gym and is amazed by his jump rump skills . 'To work at hard not only at working out, but to work hard at everything. 'I think my dad fighting Pacquiao... it's just another fight in my opinion.' Floyd Mayweather and\xa0Pacquiao have been keeping boxing fans updated daily on social media with their training schedules and\xa0Iyanna Mayweather explained how impressed she was with her father's work ethic in the gym. 'I like watching my dad jump rope because I've never seen anyone jump rope like that before,' she added. Mayweather posted an update to his Instagram account on Friday as he embarked on another shopping trip . 'It's fun coming to the gym because when dad's not in training camp, the Money team doesn't see each other often so when my dad gets back in training camp, we get back to seeing each other. 'We hang out a lot, we play around, we just have fun outside of the gym. My dad is my best friend.
BART	Floyd Mayweather vs Manny Pacquiao will generate revenue upwards of $300 million.\nIyanna Mayweather has been spending time in her father Floyd's training camp.
GPT-4	Iyanna Mayweather, daughter of champion boxer Floyd Mayweather who is set to fight Manny Pacquiao in Las Vegas in a match predicted to generate $300m, has given insights into her father's training. She said that watching him train and his work ethic in the gym was \"inspiring\", particularly praising his skill at jumping rope. She also said that, outside of the training camp, the Money Team spent a lot of time together joking and having fun.
Llama2-7B	The article talks about the preparation for the biggest fight ever between two great fighters - Floyd Mayweather Jr. and Manny Pacquaio. It also mentions that Iyana Mayweather (Floyds' daughter) watched him train and found it very inspirational. She admires her father's dedication towards fitness and believes that this fight would be no different from any other one.
정답 요약문	Floyd Mayweather will fight Manny Pacquiao in Las Vegas on May 2 .\nThe bout is expected to generate $300 million in revenue .\nIyanna Mayweather has been in training camp with her father Floyd .

표 2. 모델별 ROUGE 점수 비교

	ROUGE-1	ROUGE-2	ROUGE-L
BART	0.41	0.19	0.29
GPT-4	0.34	0.17	0.21
Llama2-7B	0.22	0.06	0.15

2.3 언어 모델에서의 ROUGE 점수 평가

일반적으로 요약 태스크를 평가할 때는 ROUGE 점수를 사용하며 각 모델의 ROUGE 점수는 표 2와 같다. 표 2를 보면 인코더-디코더 모델인 BART가 언어 모델인 GPT-4와 Llama2-7B의 성능을 한참 뛰어넘는다. 앞서 설명했던 것처럼, 인코더-디코더 모델은 인코더를 통해 본문의 단어나 구절을 그대로 요약문에 포함하지만 언어 모델의 경우 디코더만 존재하기 때문에 추상적인 요약을 해서 언어 모델의 ROUGE 점수는 인코더-디코더 모델보다 상대적으로 낮다.

언어 모델은 요약문에 적합한 동의어를 사용하여 요약을 하는 경향이 있다. ROUGE 점수는 단순히 겹치는 단어들의 수를 계산하기 때문에 요약문에 더 적합한 단어라고 하더라도 더 낮은 점수를 받는 경우가 발생한다. 특히, 2.2에서 서술한 것처럼 요약 데이터 셋의 정답 요약문이 제대로 구성되어있지 않기 때문에 정답 요약문과의 ROUGE 점수는 신뢰할 수 없다. 3.3에서 인코더-디코더 모델인 BART가 생성한 요약문이 언어 모델인 GPT-4와 Llama2-7B가 생성한 요약문보다 사람의 선호도가 낮은 것을 보인다.

3. 성능 평가

각 모델의 요약 성능을 측정하기 위해 UniEval [15]과 ChatGPT, 인간 평가를 사용하였다. UniEval은 NLG (Natural Language Generation) 태스크를 Boolean Question Answering으로 평가하는 방식이다. 요약문 간 겹치는 단어들의 수를 계산하는 ROUGE 점수와 다르게 다양한 방법으로 요약문의 성

표 3. 모델별 UniEval 점수 비교

	Coherence	Consistency	Fluency	Relevance
BART	0.95	0.95	0.92	0.78
GPT-4	0.89	0.81	0.94	0.90
Llama2-7B	0.84	0.72	0.91	0.87

표 4. 모델별 ChatGPT 점수 비교

	Coherence	Consistency	Fluency	Relevance
BART	3.78	4.33	4.39	3.61
GPT-4	4.89	4.78	4.94	5.00
Llama2-7B	3.67	3.78	4.00	3.39
정답 요약문	4.11	4.72	4.56	4.28

능을 평가한다. 다음으로, 최근 연구에 따르면 [16] ChatGPT를 통해 채점하는 것이 사람이 평가하는 것과 유사하다 하여 ChatGPT를 통해 각 요약문을 평가하였다. 마지막으로 인간 평가를 사용하였는데, 사람이 요약문을 어떻게 이해하는지가 중요하기 때문에 인간 평가가 요약 태스크에서 가장 중요한 평가 척도이다. 본 논문에서는 다양한 각도에서 각 모델을 비교하기 위해 UniEval, ChatGPT, 인간 평가를 사용하였으며 총 4가지 항목으로 각 요약문을 평가하였다.각 모델들이 생성한 요약문을 평가하는 4가지 항목은 다음과 같다.

- Coherence: 모든 문장이 일관성이 있는지를 평가하는 항목이다. 각 문장이 주장하고 있는 내용이 일관성이 있는지 평가한다.
- Consistency: 요약문과 본문의 내용이 일치하는지를 평가하는 항목이다. 본문에 명시되어 있는 사람, 위치, 날짜 등이 일치하는지 평가한다.
- Fluency: 요약문의 각 문장이 자연스럽게 이루어져 있는지를 평가하는 항목이다.
- Relevance: 요약문이 본문의 중요한 정보만을 포함했는지를 평가하는 항목이다. 요약 태스크에서는 요약문에 본문의 중요한 내용만이 포함되어야 좋은 요약문이기 때문에 요약 태스크에서 가장 중요한 항목이다.

3.1 UniEval

위 표 3에서 각 모델별 UniEval 점수를 보면, GPT-4의 점수가 전반적으로 높은 것을 볼 수 있다. GPT-4와 Llama2-7B의 경우 전반적으로 GPT-4의 성능이 높다. 두 모델 모두 언어 모델이기 때문에 점수 분포도가 유사하고, 인코더-디코더 모델인 BART의 경우 언어 모델과 다른 분포를 보인다. 특히 BART와 Llama2-7B을 비교하면, BART는 상대적으로 consistency가 높고 Llama2-7B의 경우 상대적으로 relevance가 높다. BART가 생성하는 요약문의 경우 본문에 있는 단어나 구절을 그대로 사용하는 경우가 있어서 본문과 요약문의 일치성을 보는 consistency가 높고 본문의 핵심 내용이 아닌 부분을 요약문에 포함하는 경우가 있어서 relevance가 낮은 것을 볼 수 있다.. Llama2-7B의 경우 추상적인 요약을 하고 종종 할루시네이션 문제가 발생하기 때문에 consistency가 낮고 본문에서 중요하지 않은 내용은 주로 포함하지 않기 때문에 relevance가 높은

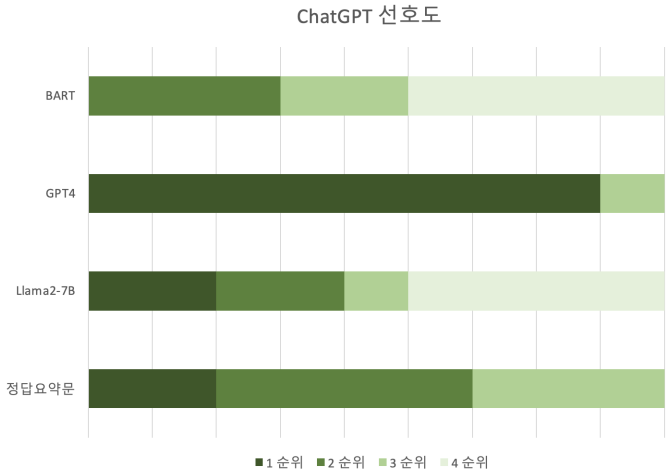

그림 1. 각 요약문 별 ChatGPT 선호도

것을 볼 수 있다.

UniEval 점수는 ROUGE 점수와는 다르게 다양한 관점에서 언어 모델의 요약 성능을 비교할 수 있다. 하지만 UniEval도 자동 측정 방법의 하나기 때문에, 실제 사람의 선호도와는 다를 수 있으며, [3]에 따르면 자동 점수 측정 방법으로는 요약문을 제대로 평가할 수 없으며, 요약문을 평가하기 위해서는 인간 평가를 동반해야 한다고 한다.

3.2 ChatGPT 평가

최근 연구 결과에 따르면 ChatGPT의 평가 성능이 사람이 평가하는 것과 비슷하다고 한다 [16]. 해당 논문에 따르면 text-davinci-003과 chatGPT가 평가하는 것이 인간 평가와 유사하여, 이러한 대규모 언어 모델을 통해 인간 평가를 대체할 수 있다고 한다. 실제로 그림 1과 그림 2을 보면 모델별 ChatGPT가 선호하는 요약문과 사람이 선호하는 요약문이 대체로 비슷하다. 그림 1과 그림 2를 보면 ChatGPT는 대부분 GPT-4가 생성한 요약문을 선호하였고, Llama-7B와 정답 요약문이 그 뒤를 이었다.

그림 1과 2을 보면 BART와 Llama-7B에서 생성하는 요약문의 특징을 발견할 수 있었는데, BART의 경우 Llama2-7B에 비해 대부분 2순위와 3순위에 포진되어 있고 Llama2-7B의 경우 BART에 비해 1순위와 4순위의 비율이 높았다. 이를 통해 알 수 있는 것은 BART의 경우 파인 튜닝이 되어있어서 대체로

표 5. 모델별 인간 평가 점수 비교

	Coherence	Consistency	Fluency	Relevance	Overall
BART	3.56	3.89	3.99	2.81	2.78
GPT-4	4.08	3.92	3.99	3.57	3.63
Llama2-7B	3.61	3.56	4.05	2.96	2.72
정답 요약문	3.40	3.68	3.96	2.84	2.55

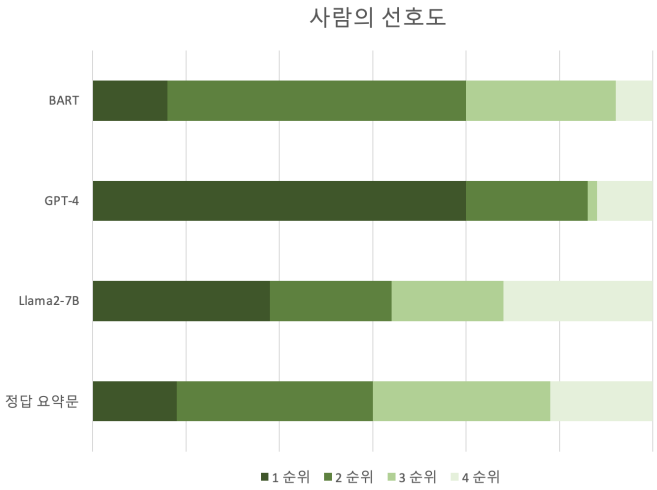

그림 2. 각 요약문 별 사람의 선호도

매끄러운 문장을 생성하며, 요약문이 본문의 핵심 내용을 포함하지 않고 있어 가장 선호하는 요약문은 아니었지만 각 문장이 자연스러우므로 전반적으로 중간 정도의 성능을 보였다는 것이다. 반면 Llama-7B의 경우 대체로 요약문의 품질이 좋았지만, 언어 모델의 크기가 작기 때문에 종종 이상한 문장을 생성하는 경우가 발생하여 요약문의 품질 편차가 컸다.

표 4를 보면 BART의 점수가 Llama-7B보다 높지만, 요약문별 선호도에서는 Llama-7B가 BART보다 높은 것을 볼 수 있다.

3.3 인간 평가(Human Evaluation)

요약 태스크에서 인간 평가는 가장 중요하다. 사람이 실제로 해당 요약문을 어떻게 생각하는지가 가장 중요하기 때문에 요약 태스크에서 인간 평가는 필수적이다. 표 5와 그림 2를 보면 인간 평가에서도 GPT-4가 생성한 요약문이 가장 좋은 평가를 받았다. 또한, UniEval 점수와 비슷하게 BART와 Llama2-7B의 성능이 유사했는데, 앞서 설명한 것과 같이 BART가 Llama2-7B보다 Consistency는 높지만, Relevance는 떨어지는 것을 볼 수 있다. 그림 2의 사람의 선호도 면에서도 ChatGPT의 평가와 유사하게 나타났다. Llama2-7B는 좋은 요약문을 생성할 때도 많지만 엉뚱한 문장을 생성하는 예도 있어서 요약문별로 품질 차이가 크지만, BART의 경우 대부분 평범하거나 좋지 않은 요약문을 생성하기 때문에 2순위, 3순위에 위치한다.

본 연구에서 Llama2-7B를 사용할 때 일반적으로 요약 태스크에서 사용하는 프롬프트를 사용하였고 학습하지 않는 제로샷으로 진행했다. 그럼에도 불구하고 Llama-7B가 파인 튜닝된 BART 모델과 필적할만한 성능을 내었다. 따라서, 언어 모델의 프롬프트만 다르게 주어도 인코더-디코더 모델보다 더 좋은 요약문을 생성할 잠재력이 존재한다. 또한 최근에 PEFT를 통해서 대규모 언어 모델을 파인 튜닝하는 기법들이 연구가 진행되고 있는데 LoRA [17]나 Adapter [18]와 같이 언어 모델을 파인 튜닝해서 사용한다면 Llama2-7B과 같이 작은 크기의 언어 모델이 인코더-디코더 모델보다 뛰어난 성능을 보일 잠재력이 있는 것으로 예상된다.

또한, 다양한 평가를 통해 정답 요약문의 품질이 낮은 것을 볼 수 있었다. 모든 평가에서 정답 요약문은 GPT-4에 비해 훨씬 낮은 평가를 받았고, 심지어 인간 평가에서는 모든 요약문 중 가장 낮은 평가를 받았다. 이를 통해, 현재 요약 데이터 셋의 정답 요약문으로 평가를 하는 것이 좋지 않다는 것을 다시 한번 확인할 수 있다.

인간 평가는 총 7명의 전문가가 참여하였으며 평가자의 요약문 선호도에 따라 인간 평가 점수가 큰 편차를 보였다. 이는 평가자마다 선호하는 요약문의 스타일과 요약문에 요구하는 내용이 달랐기 때문이다.

4. 결론

본 논문에서는 뉴스 요약 태스크에서 인코더-디코더 모델과 언어 모델의 성능을 비교 및 분석하였다. 또한 현재 요약 데이터 셋으로 언어 모델의 성능을 충분히 평가할 수 없다는 것을 여러 관점에서 보여주었다. 인코더가 없는 언어 모델의 특성상 인코더-디코더 모델보다 추상적 요약을 하게 되고 이에 따라 단순히 정답 데이터 셋의 겹치는 단어들의 개수로 평가하는 ROUGE 점수는 낮게 측정되는 경향이 있다. 또한 인코더-디코더 모델의 경우, 품질이 낮은 데이터 셋으로 파인 튜닝을 해서, 실제 출력되는 요약문의 품질이 좋지 않았다. 따라서, BART의 경우 ROUGE 점수는 언어 모델인 GPT-4보다 높지만 UniEval, ChatGPT 평가, 인간 평가에서는 훨씬 낮은 점수를 보였으며, 심지어 Llama2-7B와도 유사한 성능을 보였다.

언어 모델의 경우 크기와 성능이 급격하게 향상하면서, 프롬프트 엔지니어링과 사고의 사슬만으로도 모델 성능 향상이 이루어지고 있다. 또한 사람의 피드백을 통한 강화학습과 지시학습을 통해 더 섬세한 학습이 가능해졌기에 앞으로 언어 모델을 활용한 요약이 주를 이룰 것으로 예상된다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2023-00217286)

참고문헌

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, Vol. 1, No. 8, p. 9, 2019.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[3] T. Goyal, J. J. Li, and G. Durrett, "News summarization and evaluation in the era of gpt-3," *arXiv preprint arXiv:2209.12356*, 2022.

[4] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022.

[5] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *arXiv preprint arXiv:2301.13848*, 2023.

[6] Y. Liu, A. R. Fabbri, P. Liu, D. Radev, and A. Cohan, "On learning to summarize with large language models as references," *arXiv preprint arXiv:2305.14239*, 2023.

[7] Y. Wang, Z. Zhang, and R. Wang, "Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method," *arXiv preprint arXiv:2305.13412*, 2023.

[8] Y. Liu, A. R. Fabbri, P. Liu, Y. Zhao, L. Nan, R. Han, S. Han, S. Joty, C.-S. Wu, C. Xiong *et al.*, "Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation," *arXiv preprint arXiv:2212.07981*, 2022.

[9] OpenAI, "Gpt-4 technical report," 2023.

[10] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[11] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[12] Y. Liu, P. Liu, D. Radev, and G. Neubig, "Brio: Bringing order to abstractive summarization," *arXiv preprint arXiv:2203.16804*, 2022.

[13] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[14] H. D. Lasswell, "The structure and function of communication in society," *The communication of ideas*, Vol. 37, No. 1, pp. 136–139, 1948.

[15] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han, "Towards a unified multi-dimensional evaluator for text generation," *arXiv preprint arXiv:2210.07197*, 2022.

[16] C.-H. Chiang and H.-y. Lee, "Can large language models be an alternative to human evaluations?" *arXiv preprint arXiv:2305.01937*, 2023.

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[18] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," *International Conference on Machine Learning*, pp. 2790–2799, 2019.