도메인 적응 사전 훈련 (Domain-Adaptive Pre-training, DAPT) 한국어 문서 요약

장형국[1], 장현철[2]
[1]고려대학교 컴퓨터정보통신대학원 빅데이터융합학과 석사
[2]고려대학교 컴퓨터정보통신대학원 빅데이터융합학과 석사과정

gagagiga@korea.ac.kr, justinjang87@korea.ac.kr

Domain-Adaptive Pre-training for Korean Document Summarization

Hyungkuk Jang[1], Hyuncheol, Jang[2]
[1]Dept. of Big Data Convergence, Korea University
[2]Dept. of Big Data Convergence, Korea University

요 약

도메인 적응 사전 훈련(Domain-Adaptive Pre-training, DAPT)을 활용한 한국어 문서 요약 연구에서는 특정 도메인의 문서에 대한 이해도와 요약 성능을 향상시키기 위해 DAPT 기법을 적용했다. 이 연구는 사전 훈련된 언어 모델이 일반적인 언어 이해 능력을 넘어 특정 도메인에 최적화된 성능을 발휘할 수 있도록 도메인 특화 데이터셋을 사용하여 추가적인 사전 훈련을 진행한다. 구체적으로, 의료, 법률, 기술 등 다양한 도메인에서 수집한 한국어 텍스트 데이터를 이용하여 모델을 미세 조정하며, 이를 통해 얻은 모델은 도메인에 특화된 용어와 문맥을 효과적으로 처리할 수 있음을 보여준다. 성능 평가에서는 기존 사전 훈련 모델과 DAPT를 적용한 모델을 비교하여 DAPT의 효과를 검증했다. 연구 결과, DAPT를 적용한 모델은 도메인 특화 문서 요약 작업에서 성능 향상을 보였으며, 이는 실제 도메인별 활용에서도 유용할 것으로 기대된다.

1. 서론

도메인 적응 사전 훈련(Domain-Adaptive Pre-training, DAPT)은 특정 도메인의 데이터를 활용하여 기존 사전 훈련된 언어 모델의 성능을 개선하는 기법이다.[1] 최근 자연어 처리(NLP) 분야에서 사전 훈련된 언어 모델들이 다양한 작업에서 좋은 성과를 보이고 있지만, 특정 도메인에서 요구되는 세밀한 이해에는 한계가 있다. 이에 본 연구는 한국어 문서 요약을 위해 DAPT를 적용하여, 모델이 도메인 특화 용어와 문맥을 더 잘 파악하도록 한다.

한국어는 그 구조와 문법이 복잡하여, 전문적인 용어 사용과 문장 구성이 도메인에 따라 크게 달라질 수 있다. 이러한 특성은 표준 언어 모델을 사용할 때 정확한 문맥 이해와 정보 처리에 어려움을 줄 수 있다. 본 연구는 한국어 도메인 특화 데이터셋을 사용해 모델을 추가적으로 사전 훈련시킴으로써, 이러한 문제를 완화하고자 한다.

본 연구의 목적은 DAPT 방법과 2단계 요약을 적용하여 한국어 문서 요약의 성능을 개선하는 것이다. 연구 결과는 도메인 특화 훈련이 요약 작업에 미치는 영향을 평가하고, 한국어 문서 처리 능력에 있어서의 성과를 조명할 예정이다. 또한, 이 연구가 한국어 처리 기술뿐만 아니라, 다양한 언어와 도메인에 적용될 수 있는 방법론적 기여를 할 수 있는지를 검토할 것이다.

2. 활용하고자 하는 연구모델 소개

본 연구에서는 한국어 문서 요약의 성능을 향상시키기 위하여 다양한 사전 훈련된 언어 모델들을 활용하였다. 이 모델들은 각각의 특성을 바탕으로 특정 도메인의 데이터에 대한 적응력을 강화하고, 요약 작업에 있어 더욱 효과적인 결과를 도출할 수 있도록 설계되었다.

KoBERT는 한국어에 최적화된 BERT 마스크 언어 모델로, 한국어의 특성을 잘 반영하여 다양한 NLP 작업에 활용된다.[2] KoBART는 자연어 이해와 생성 능력이 뛰어난 BART 모델의 한국어 버전으로, 특히 요약 및 번역 작업에 유용하다. KoRoBERTa는 RoBERTa 모델의 한국어

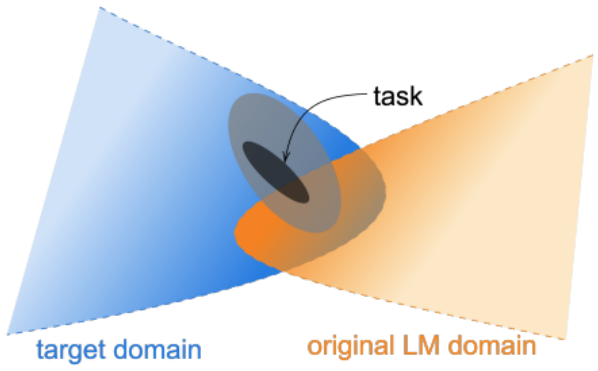
버전으로, 더 많은 데이터와 더 긴 학습을 통해 성능을 개선한 모델이다. polyglot-ko는 다국어 처리에 특화된 모델로 한국어를 포함하여 여러 언어의 처리가 가능하며, 광범위한 언어 데이터에 대한 이해도가 높다. KULLM3는 고려대학교에서 개발한 한국어 대화 모델로, 일반적인 대화뿐만 아니라 지시사항에 따른 반응 생성에 뛰어난 성능을 보여준다. 이 모델은 SOLAR-10.7B-Instruct-v1.0 모델을 기반으로 하여 개발되었으며, 한국어 대화 처리에 최적화되어 있다.[3]

각 모델은 사전 훈련된 기본 구조를 바탕으로 추가적인 도메인 적응 사전 훈련(DAPT)을 통해 특정 도메인의 데이터에 더욱 잘 적응할 수 있도록 한다. 이를 통해 모델은 도메인 특화 용어와 문맥을 더욱 정확하게 파악하고, 이를 요약에 효과적으로 활용할 수 있게 된다. 이러한 접근은 한국어 문서 요약의 정확도와 효율성을 일부 개선했다.

3. 향상 방안 1 (DAPT)

도메인 적응 사전 훈련(Domain-Adaptive Pre-training, DAPT)은 언어 모델이 특정 도메인의 언어 패턴과 용어에 더 잘 적응할 수 있도록 설계된 방법론으로, 기존의 사전 훈련된 언어 모델이 가지고 있는 일반적인 언어 이해 능력을 특정 도메인에 맞게 최적화하는 데 중점을 둔다.[4] 이러한 최적화 과정은 모델이 도메인에 특화된 미묘한 언어적 차이와 전문 용어를 이해하고, 결과적으로 문서 요약과 같은 고차원적인 자연어 처리 작업에서 보다 정확한 성능을 발휘하게 한다. 예를 들어, 법률이나 금융 및 의료와 같은 전문 분야의 문서에서는 일반적인 사전 훈련된 모델로는 잡아내기 어려운 전문 용어와 개념들이 빈번히 등장한다. DAPT를 적용한 모델은 이러한 전문 용어에 대한 이해도가 향상되므로, 문서 요약의 정확도와 품질이 향상된다.[5] 이를 위해 KoBERT, KoBART, KoRoBERTa, polyglot-ko, Kullm3 등의 다양한 한국어 모델들을 대상으로 DAPT를 적용하고 그 효과를 측정했다.

AI Hub의 한국어 데이터셋과 영어를 한국어로 번역한 데이터셋을 활용했다. AI Hub 데이터셋은 다양한 도메인의 텍스트를 포함하고 있으며, 번역된 데이터는 모델이 다양한 언어적 표현과 문맥을 학습하도록 한다.



(그림 1) DAPT 데이터 분포에 대한 설명

4. 향상 방안 2 (DAPT+2단계 요약)

문서 요약에서 DAPT와 더불어 2단계 접근법인 추출 후 생성 요약 방식을 적용하는 것은 복합적인 전략을 통해 요약의 질을 높이려는 시도다.[6] 이 방법은 먼저 원본 문서에서 중요한 정보를 추출하고, 이를 바탕으로 새로운 요약문을 생성하는 과정으로 이루어진다. 추출 단계에서는 문서 내의 핵심 문장이나 구를 식별하여 요약의 '골격'을 마련하며, 생성 단계에서는 이 골격을 사용하여 자연스럽고 읽기 쉬운 요약문을 작성한다. 이러한 복합 접근법은 단순히 중요한 문장만을 나열하는 것을 넘어서, 문맥적으로 일관성 있고 완성도 높은 요약을 도출해내는 데 도움을 준다.

한국어 문서 요약에 있어서 이러한 접근법의 효과를 분석하기 위해, KoBERT, KoBART, KoRoBERTa, Kullm3, polyglot-ko 등의 모델을 사용하여 실험을 진행했다. 모델마다 DAPT를 적용하여 도메인 특화 사전 훈련을 수행한 후, 추출 및 생성 요약 단계를 거치도록 설계했다.

추출 후 생성 요약 방식은 문서 요약 작업에서 널리 사용되는 두 가지 접근법을 통합한 방식으로, 정확성 향상, 일관성과 읽기 쉬움, 유연성을 제공한다. 이 방법은 추출 단계에서 중요한 정보를 먼저 식별하고, 이를 바탕으로 생성 단계에서 요약을 생성하기 때문에 요약의 정확성과 관련성이 향상되며, 추출된 정보를 기반으로 요약을 생성함으로써 결과적으로 자연스러우면서도 문맥적으로 일관된 텍스트를 생성할 수 있고, 다양한 도메인과 긴 문서에 대해서도 효과적으로 적용할 수 있는 유연성을 제공한다. 단점으로는 계산 비용, 복잡성, 과적합의 위험이 존재한다. 두 단계의 요약 과정을 거치기 때문에 시간과 계산 비용이 더 많이 소요되며, 시스템의 구현과 유지 관리가 복잡해지고, 특히 훈련 데이터가 제한적인 경우 모델이 훈련 데이터에 과적합되어 새로운 데이터에 대한 일반화 성능이 떨어질 수 있다.

5. 연구 결과

실험 결과에서 ROUGE-L, ROUGE-1, ROUGE-2 평가 지표들을 통해 모델의 성능을 측정했으며, 이를 통해 DAPT 적용 전과 후의 성능을 비교할 수 있었다. 아래의 표는 이러한 비교를 정리한 것으로, DAPT를 적용하지 않은 모델들과 적용한 모델들의 성능 차이를 보여준다.

모델명	ROUGE-L	ROUGE-1	ROUGE-2
Kullm3	44.23	45.12	24.14
KoRoBERTa	43.44	44.68	23.30
KoBART	41.76	43.23	21.94
KoBERT	39.10	42.30	21.17
polyglot-ko	38.59	41.79	20.82

DAPT 적용 전

모델명	ROUGE-L	ROUGE-1	ROUGE-2
Kullm3	49.22	49.53	28.97
KoRoBERTa	48.74	48.66	27.47
KoBART	47.88	48.18	25.82
KoBERT	45.45	45.97	23.76
polyglot-ko	41.14	43.12	21.45

DAPT 적용 후

DAPT와 2단계 요약 접근법을 적용함으로써 각 모델의 문서 요약 능력이 전반적으로 향상되었다.

모델명	ROUGE-L	ROUGE-1	ROUGE-2
Kullm3	50.65	51.84	29.44
KoRoBERTa	49.12	51.25	27.92
KoBART	48.71	49.98	26.35
KoBERT	46.50	47.63	24.27
polyglot-ko	43.28	44.10	20.91

DAPT + 2단계 요약

6. 결론 및 향후 연구

본 연구 결과에서는 DAPT와 2단계(추출 후 생성) 요약 접근법을 적용함으로써 한국어 문서 요약 작업에 있어서 성능 향상을 이루었다. 특히 DAPT가 모델이 도메인 특화 언어 패턴을 학습하는데 있어서 중요한 역할을 하며, 2단계 요약 접근법이 이러한 훈련된 모델이 추출한 정보를 바탕으로 보다 일관성 있는 요약문을 생성하는데 기여한다는 점을 보여준다.

향후 연구에서는 GPT-3.5와 GPT-4와 같은 대규모 언어 모델을 평가자로 도입하여, 평가 방법을 발전시킬 예정이다. 이러한 언어 모델들은 다양한 문맥에서 의미론적 일관성과 표현의 다양성을 평가하는 데 있어 인간 평가자와 유사한 수준의 민감도를 가질 수 있다. 따라서, 문서 요약의 질을 평가하는 기존 지표 외에도 모델이 생성한 요약의 문맥적 적합성, 자연스러움을 평가하는 새로운 지시문을 추가하는 것이 필요하다. 이를 위해 향후 GPT-3.5 및 GPT-4를 사용하여 요약문에 대한 자세한 평가를 실시하고, 이를 바탕으로 모델의 요약 성능을 더욱 세밀하게 분석할 계획이다. 또한, 언어 모델이 생성한 평가 결과를 인간 평가자와 비교 분석하여 언어 모델을 활용한 평가의 타당성과 신뢰성을 추가적으로 검증할 예정이다.

참고문헌

[1] Xu et al., "Domain-Adaptive Pretraining Methods for Dialogue Understanding," arXiv, 2021.

[2] Qiu et al., "Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey," arXiv, 2021.

[3] 구름(KULLM): 한국어 지시어에 특화된 거대 언어 모델, 제35회 한글 및 한국어 정보처리 학술대회 논문집 2023.

[4] Gururangan et al., "Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks," arXiv, 2020.

[5] KF-DeBERTa: 금융 도메인 특화 사전학습 언어모델, 제35회 한글 및 한국어 정보처리 학술대회 논문집, 2023

[6]Encoder-Decoder 및 LLM을 활용한 2단계 한국어 문서 요약, 한국정보학회 인공지능 학술대회 논문집, 2023