홍 기 형 교수 지도

석사학위 청구논문

대규모 언어 모델과 QLoRA 미세 조정 기반 보완대체의사소통 상징 시퀀스의 한국어 문장 변환

2025

성신여자대학교 대학원

미래융합기술공학과

서 지 우

대규모 언어 모델과 QLoRA 미세 조정 기반 보완대체의사소통 상징 시퀀스의 한국어 문장 변환

홍기형 교수 지도

이 논문을 석사학위논문으로 제출함

2024년 11월

성신여자대학교 대학원
미래융합기술공학과
서 지 우

인 준 서

서지우의 석사학위 논문으로 인준함

2025년 1월

심사위원장 오 장 민 

심 사 위 원 변 혜 원 

심 사 위 원 홍 기 형 

성신여자대학교 일반대학원

논문개요

현대 사회에서의 의사소통의 중요성은 갈수록 높아지고 있으며, 대면 의사소통뿐만 아니라 비대면 의사소통 또한 매우 중요해졌다. 그러나 지적장애, 자폐스펙트럼장애 등 언어 표현과 이해에 어려움을 겪는 사람들의 대면 및 비대면 의사소통은 아직 보장받지 못하고 있다. 보완대체의사소통(Augmentative and Alternative Communication, AAC)은 구어를 통한 의사소통에 어려움이 있는 사람들의 의사소통을 지원하기 위한 도구나 방식으로 그들의 소통 창구로서 언어 장애인 간, 혹은 비장애인과의 의사소통과 사회적 상호 작용을 지원한다. 보완대체의사소통 상징은 표현하고자 하는 개념이나 의미를 직관적으로 표현하는 상징 이미지와, 상징 식별 및 관리를 위한 상징 이름(identifier), 전달하고자 하는 개념이나 의미를 나타내는 표현 어휘(expression)으로 구성되어 있다. AAC 사용자는 AAC 상징을 나열하여 의사나 감정을 상대방에게 전달할 수 있다. 그러나 비장애인의 AAC 상징에 대한 이해도가 낮아 AAC 상징만을 활용한 의사소통에는 제한이 있다. 따라서 AAC 사용자의 원활한 의사소통을 위해서는 AAC 상징 시퀀스를 자연스러운 한국어 문장으로 변환하는 것이 필요하다.

본 연구는 AAC 사용자의 원활한 의사소통을 지원하기 위하여 대규모 언어 모델을 기반으로 AAC 상징 시퀀스를 자연스러운 한국어 문장으로 변환하도록 하며, 제한된 연산 자원 환경에서 AAC 사용자로부터 수집된 추가 데이터에 대한 효율적인 학습을 목적으로 한다. 대규모 언어 모델(Large Language Model, LLM)의 사전 학습 모델을 활용하여 적은 데이

터만으로도 준수한 변환 성능을 기대할 수 있으며, 학습 데이터에 존재하지 않는 시퀀스 데이터 및 숫자 데이터 또한 유연하게 처리할 수 있다. 대규모 언어 모델의 성능 향상과 함께 파라미터 수가 기하급수적으로 증가하고 있어 특정 도메인에 대한 학습을 위해서는 GPU와 같은 물리적 연산 자원에 의존할 수밖에 없다. 제한된 연산 자원 환경에서도 충분한 학습이 가능하도록 지원하는 PEFT(Parameter Efficient Fine-Tuning) 방식 중 하나인 QLoRA(Quantized Low-Rank Adaptation) 미세 조정 방식을 활용하여 물리적 연산 자원의 사용량을 줄이면서도 준수한 성능을 갖추도록 하였다.

또한, 주기적인 모델 업데이트는 모델의 성능을 정기적으로 평가하고, 개선하는 과정으로서 서비스 품질과 사용자 경험 개선에 중요한 역할을 한다. 사용자로부터 수집되는 데이터는 모델 업데이트에 있어 모델이 사용자의 요구에 더욱 잘 적응하도록 한다. 따라서, AAC 사용자로부터 수집되는 추가 학습 데이터셋을 모델에 반영할 수 있는 효율적인 학습 시스템이 필요하다. 초기 데이터셋을 학습한 초기 모델만을 고려한 기존 연구에서 더 나아가 AAC 사용자가 모델을 활용하면서 축적되는 데이터에 대한 추가 학습 모델을 고려하였다. 따라서 본 연구에서 수행한 실험 내용은 다음과 같다.

첫째, 대규모 언어 모델인 GPT-2와 T5의 한국어 모델을 활용한 AAC 상징 시퀀스의 한국어 문장 변환 모델을 설계하고 성능을 비교 실험하였다. 첫 번째 실험을 위하여 한국어 대규모 언어 모델의 사전 학습 토크나이저와 사전 학습 모델을 활용하였다. 성능 지표 산출 결과, GPT-2와 T5 모델 모두 준수한 성능을 보였으며 GPT-2 모델이 BLEU 0.6519,

ROUGE-1 0.7930, ROUGE-2 0.6976, 코사인 유사도 0.7403으로 T5 모델에 비하여 우수한 성능을 보였다.

둘째, 추가 학습 데이터에 대한 미세 조정 방식에 따른 성능을 비교 실험하였다. 1) 기존 학습 데이터와 추가 학습 데이터를 합쳐 전체 미세 조정한 모델, 2) 초기 모델을 추가 학습 데이터만으로 전체 미세 조정한 모델, 3) 초기 모델을 추가 학습 데이터만으로 QLoRA 미세 조정한 모델로 구분하여 성능 및 GPU 메모리 사용량을 비교하였다. 그 결과, GPT-2의 경우 기존 학습 데이터와 추가 학습 데이터를 합친 데이터셋으로 전체 미세 조정한 모델의 성능이 가장 높았으며, 추가 훈련 데이터셋으로만 QLoRA 미세 조정한 모델은 50% 적은 GPU 사용량만으로도 준수한 성능을 보였다. T5 모델에서는 추가 훈련 데이터셋으로만 전체 미세 조정한 모델의 성능이 가장 우수하였으며, QLoRA 미세 조정을 수행한 모델은 30% 미만의 GPU 메모리를 사용하면서도 준수한 성능을 보였다.

셋째, 기존 학습 데이터와 추가 학습 데이터의 비율별 미세 조정의 성능을 비교 실험하였다. 실험 결과, 모든 비율에 대하여 GPT-2와 T5 기반으로 전체 미세 조정 및 QLoRA 미세 조정한 모델이 모두 준수한 성능을 보였으며, 특히 GPT-2 모델 기반 실험에서는 추가 데이터셋의 비율이 전체 데이터셋의 30%, 40%, 50%일 때, 전체 미세 조정한 모델보다 QLoRA 미세 조정한 모델이 좋은 성능을 보임을 확인하였다.

목 차

표 목차

그림 목차

I. 서 론

현대 사회에서 의사소통의 중요성이 갈수록 강조되고 있다. 그러나 지적장애, 자폐스펙트럼장애 등 언어 표현과 이해에 어려움을 겪는 사람들의 의사소통은 아직도 보장받지 못하고 있어 의사소통에 대한 그들의 요구는 계속 커지고 있다. 보완대체의사소통(Augmentative and Alternative Communication, AAC)은 의사소통에 어려움을 가지는 사람들의 소통 창구로서 그들의 언어 표현과 이해를 돕기 위한 도구나 방식이다. AAC 상징은 AAC 사용자의 의사나 감정을 표현하기 위한 도구로, 개념이나 의미를 직관적으로 표현하는 상징 이미지와 상징 식별 및 관리를 위한 상징 이름(Identifier), 전달하고자 하는 개념이나 의미를 나타내는 표현 어휘(Expression)을 가진다. 다양한 AAC 도구는 주로 상징을 조합하여 문장을 표현하는 방식으로 작동하지만, 비장애인과의 의사소통 과정에서 단순한 단어나 구의 나열만으로는 전달하고자 하는 의미를 명확하게 전달하는 데 어려움이 있다. 따라서 순서가 있는 AAC 상징 시퀀스를 자연스럽고 유기적인 한국어 문장으로 변환할 수 있는 효과적인 접근 방식이 필요하다.

최근 딥러닝 기반의 자연어 처리 기술이 발전하면서, 딥러닝을 활용하여 대화형 챗봇을 운영하거나 투자 및 자산 관리 등의 특정 업무를 수행하는 등 금융, 의료뿐만 아니라 다양한 분야에서 딥러닝을 기반으로 하는 서비스를 활용 및 제공하고 있다. 특히 대규모 언어 모델(Large Language Model, LLM)이 등장하면서 영어뿐만 아니라 한국어와 같은 비주류 언어에서도 우수한 문장 이해와 생성이 가능해졌다. 대규모 언어 모델의 사전 학습 모델[24]을 활용하여 학습의 효율을 높이고 적은 데이

터만으로도 일정 수준 이상의 성능을 기대할 수 있으며 학습 데이터에 존재하지 않는 시퀀스 데이터를 더욱 유연하게 처리할 수 있다. 사전 학습 모델을 미세 조정함으로써 대규모 데이터셋의 일반적인 언어 패턴과 문법을 바탕으로 특정 도메인의 특수한 작업에 적합한 모델을 만들 수 있어 AAC 상징 시퀀스를 자연스러운 한국어 문장으로 변환하는 모델을 만들기에 적합하다.

대규모 언어 모델의 성능이 향상됨과 동시에 모델의 파라미터 수 또한 기하급수적으로 늘어나고 있어[35] 이를 학습하기 위한 적합한 물리적 환경을 보장하는 것은 어려움이 있다. 또한, 대규모 언어 모델을 기반으로 학습한 모델을 AAC 사용자가 활용하면서 수집되고 축적되는 데이터를 모델에 주기적으로 반영하여 모델의 성능을 정기적으로 평가하고, 개선하는 과정이 필요하다. 이는 서비스의 품질과 사용자의 경험을 개선하는 데 중요한 역할을 한다. 이러한 과정에서 학습 환경에 따라 물리적 연산 자원과 시간적 자원에는 제약이 따른다. 특히 모델 학습은 물리적 연산 자원에 의존적이라는 점을 들어 PEFT(Parameter Efficient Fine-Tuning)[40] 방식을 통하여 모델을 효율적으로 미세 조정함으로써 적은 연산 자원만으로도 매우 큰 파라미터 수를 가지는 대규모 언어 모델을 특정 도메인과 작업에 최적화된 모델로 학습할 수 있다. 이는 제한된 물리적 자원 환경에서 기존 모델의 성능을 유지하면서도 주기적인 모델 업데이트를 효율적으로 수행하도록 돕는다.

본 연구에서는 LLM인 한국어 GPT-2[56]와 한국어 T5[55] 사전 학습 모델을 기반으로 AAC 상징 시퀀스를 한국어 문장으로 변환하는 모델을 설계하였다. 또한, 제한된 물리적 연산 자원 환경에서의 지속적인 모델

업데이트를 위하여 효율적인 모델 학습을 지원하는 PEFT 방식 중 QLoRA(Quantized Low-Rank Adaptation)[29] 미세 조정 방식을 활용하여 기존 모델을 미세 조정하도록 실험하였다. 이는 기존 AAC 상징 시퀀스의 한국어 문장 변환 연구[2]에서 제안한 접근법이 숫자 데이터를 처리하는 데 취약하다는 한계를 극복하며, 효율적이고 유연한 변환을 제공함으로써 AAC 사용자 간, 혹은 비장애인과의 의사소통의 질을 높이며 AAC 사용자의 자율성과 사회적 상호작용을 보장한다는 점에서 의의가 있다.

본 논문은 다음과 같이 구성되었다. 2장에서는 한국형 보완대체의사소통 상징과 상징 시퀀스를 한국어 문장으로 변환하는 선행 연구를 다루며, 본 연구에서 활용한 대규모 언어 모델과 QLoRA 미세 조정의 이론적 배경을 기술하였다. 3장에서는 대규모 언어 모델을 활용하여 상징 시퀀스를 한국어 문장으로 변환하는 모델의 설계를, 4장에서는 QLoRA 미세 조정 방식을 활용하여 추가 학습 데이터를 기존 모델에 적용하는 방식에 따른 성능과 연산 자원 사용량의 비교 실험을 설계하였다. 5장에서는 설계한 모델 및 실험을 수행하고, 그 결과를 BLEU[30,53], ROUGE-N[27], 코사인 유사도[44] 성능 지표를 통하여 평가하였다. 6장에서는 결론을 통하여 본 연구를 요약하고 향후 연구의 방향을 제시하였다.

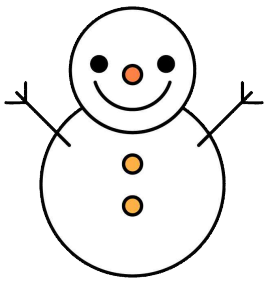
II. 관련 연구 및 이론적 배경

1. 한국형 보완대체의사소통 상징 체계집

보완대체의사소통(Augmentative and Alternative Communication, AAC)은 지적장애, 자폐스펙트럼장애, 뇌병변장애 등 신체적, 언어적, 혹은 발달적 장애를 가지는 등 언어 표현 및 이해에 어려움을 겪는 사람들의 의사소통을 지원하기 위한 다양한 도구나 방식을 말한다[9]. 보완대체의사소통 도구는 의사소통판[5], PECS[14,43], 손담[46]과 같은 로우테크 AAC 도구와 마이토키[45], GeoAAC[9, 12] 등 하이테크 AAC 도구로 구분되며, AAC를 제공하는 매개체로서 언어장애인의 의사소통을 지원한다. 보완대체의사소통의 목적은 단순히 의사소통을 보조하는 것뿐만 아니라 AAC 사용자가 자율적이며 적극적으로 의사소통 및 사회적 상호작용에 참여하고 자신을 표현할 수 있도록 지원하는 데 있다. AAC 사용자는 이러한 AAC 도구를 활용하여 명확한 의사를 전달할 수 있다.

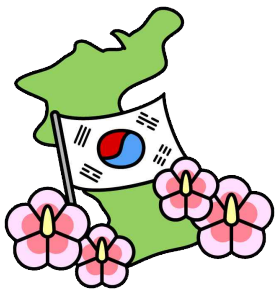
다양한 개념을 2차원적으로 표현한 표상 상징 중 AAC 그림 상징[41]은 상징 이미지와 상징 이름, 상징 표현을 가지며, 직관적인 이미지와 텍스트를 통하여 AAC 사용자의 의사소통을 지원한다. AAC 상징은 주로 AAC 도구를 통하여 전달되며 인사말, 주제, 감정 등 다양한 분야의 단어나 구의 의미를 담고 있다. [표 1]과 같이 '눈사람', '여우' 상징은 상징 이름 및 표현과 그 표현에 대한 직관적인 2차원 이미지를 통하여 의미를 내포한다. AAC 사용자는 이러한 AAC 상징을 학습하고 그들의 일상생활에서 구어적인 표현을 대체하여 활용할 수 있다.

[표 1] 보완대체의사소통 상징 예시

	Example #1	Example #2
Symbol Image		
Symbol Name	눈사람	여우
Symbol Expression	눈사람	여우

그러나 [41]에 따르면 해외에서 개발된 기존 보완대체의사소통 그림 상징의 경우, 언어와 문화가 달라 국내에서 적절하게 활용하기에는 부족함이 있다. 따라서 한국의 정서와 문화가 반영된 새로운 상징 체계를 필요로 하였으며, 이를 반영한 한국형 보완대체의사소통 상징 체계집[41]이 개발되었다. 한국형 보완대체의사소통 상징 체계집은 '추석' 등 한국의 명절, '광복절'과 같은 역사적 기념일 또는 국경일뿐만 아니라 '유관순' 등 한국의 역사적 인물이나 '유재석'과 같은 유명 연예인, '뽀로로' 등의 캐릭터, '카카오톡'과 같은 한국의 주요 메신저 등 한국의 문화가 반영된 다양한 AAC 상징이 있어[표 2] 많은 국내 보완대체의사소통 서비스에서 활용되고 있다.

[표 2] 한국 정서와 문화가 반영된 한국형 보완대체의사소통 상징 예시

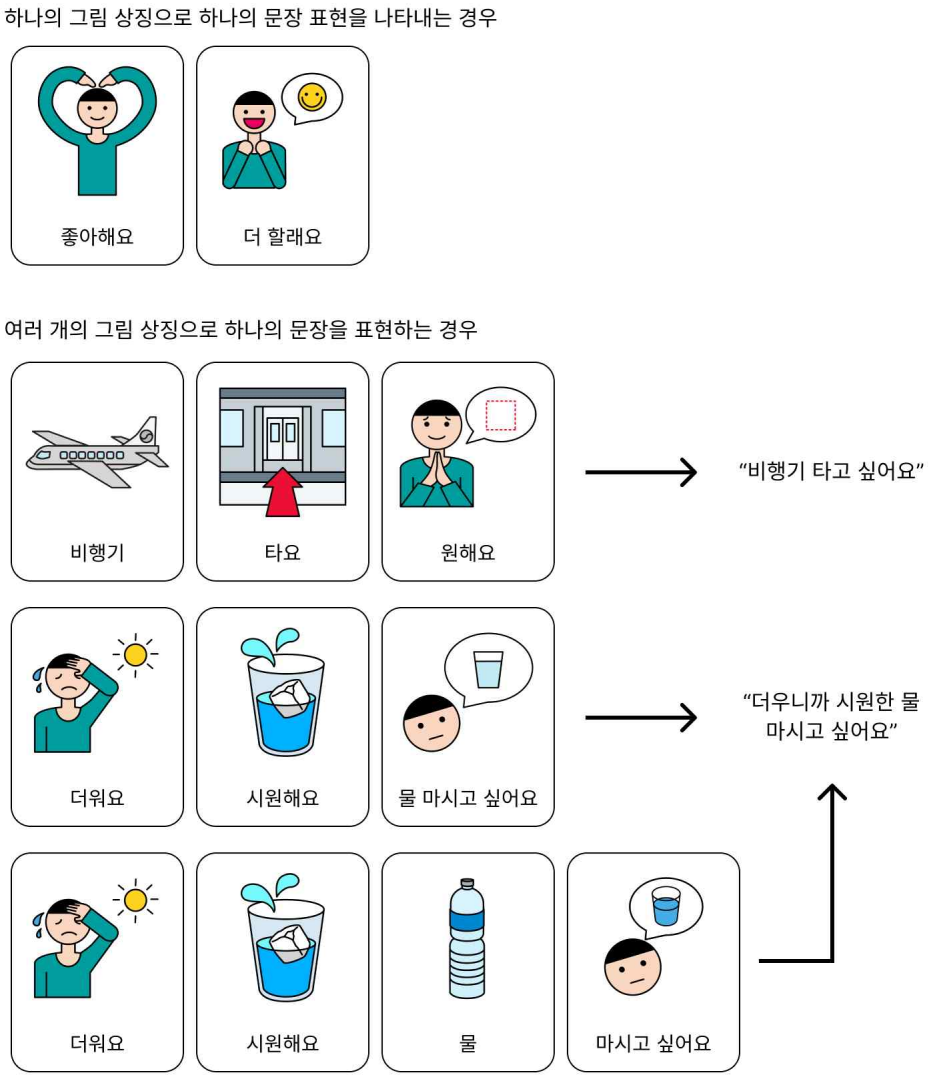
	Example #3	Example #4
Symbol Image		
Symbol Name	광복절	뽀로로
Symbol Expression	광복절	뽀로로

AAC 사용자는 AAC 도구를 활용하여 한국형 보완대체의사소통 상징으로 전달하고자 하는 표현을 구성할 수 있다. AAC 사용자는 AAC 상징을 어순에 맞추어 순차적으로 선택함으로써 상대방이 유추할 수 있는 표현을 구성한다. 순서에 따라 나열된 상징은 주로 시각적으로 전달되거나 TTS(Text-to-Speech)를 통하여 청각적으로 전달된다. 그러나 단순 나열된 상징만으로는 AAC 사용자가 표현하고자 하는 내용을 직관적으로 이해하기에는 한계가 있다.

2. 보완대체의사소통 상징 시퀀스의 한국어 문장 변환

AAC 사용자는 AAC 상징을 활용하여 의사를 표현하고 이해할 수 있다. AAC 상징에는 단어와 구, 문장이 존재한다. 하나의 상징만으로 하나의 문장을 만들 수 있고, 여러 개의 상징을 조합하여 하나의 의미를 담은 문장을 만들 수도 있다. [그림 1]과 같이 '더 할래요.', '좋아해요.'와 같은 하나의 상징이 하나의 문장을 표현할 수 있으며, '비행기', '타요', '원해요' 상징을 이어 '비행기 타고 싶어요'와 같은 문장을 완성하거나 '더워요', '시원해요', '물', '마셔요' 상징을 이어 '더워서 시원한 물 마시고 싶어요.' 처럼 여러 개의 상징을 이어 하나의 문장을 표현할 수 있다.

AAC 상징은 주로 상황과 문맥이 명확한 대면 의사소통에 활용되었다. 그러나 모바일 기기 등 다양한 하드웨어 기기와 채팅 모바일 애플리케이션 등 소프트웨어의 발달로 인하여 비대면 의사소통의 중요성 또한 높아졌으며 다양한 소셜 플랫폼에서 댓글이나 후기의 역할이 확대되어 온라인 상에서 텍스트를 통한 소통의 중요성 또한 확대되었다. 그러나 비장애인들의 보완대체의사소통 상징에 대한 이해도가 높지 않아 AAC 사용자가 일상 생활 속 오프라인, 온라인 상에서 AAC 상징을 활용하여 적극적으로 의사소통에 참여하는 것에는 한계가 있다. 따라서 AAC 상징 시퀀스를 자연스러운 한국어 문장으로 변환하여 상대방에게 전달할 필요가 있다.

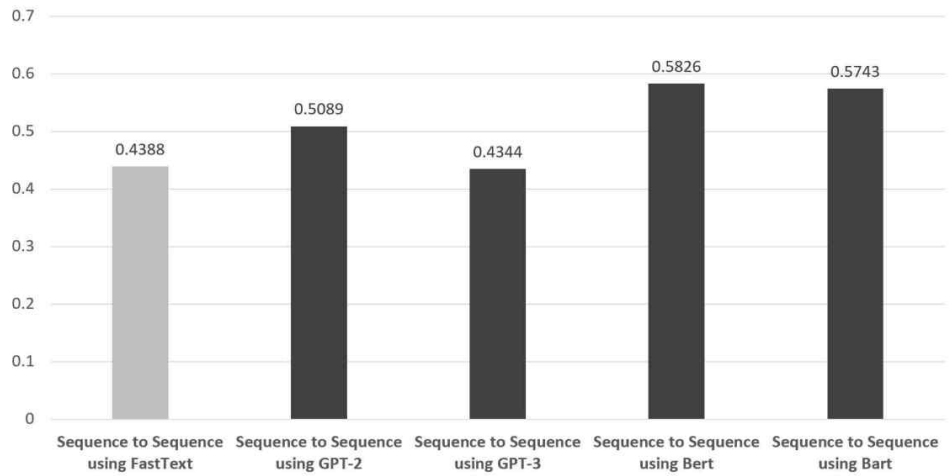


[그림 1] 상징 및 상징 시퀀스의 활용 예시

AAC 상징 시퀀스와 한국어 문장 간 변환을 수행하는 연구는 지속되어 왔다. 최근 연구로 [2]는 AAC 그림 상징이 가지는 다의성을 반영한 AAC 상징 시퀀스의 한국어 문장 변환 모델을 설계하고, 대규모 언어 모델의 임베딩 레이어를 활용한 시퀀스-투-시퀀스

(Sequnce-to-Sequence)[37] 모델을 제안하였다. [2]에서는 문장의 문맥적인 의미를 함축하는 문장 임베딩을 적용하였으며, 문장 수준의 임베딩을 수행하는 대규모 언어 모델인 BART[26], BERT[22], GPT[18,31,32] 계열 모델을 활용하였다.

어텐션 메커니즘(Attention Mechanism)[38] 기반의 게이트 순환 유닛으로 구성된 시퀀스-투-시퀀스 모델을 활용한 실험에서 임베딩 레이어별 실험의 BLEU[30,53] 성능 지표 결과는 [그림 2]와 같다. 특히, BERT[22] 모델을 기반으로 하는 시퀀스-투-시퀀스 모델에서 BLEU 점수가 0.5826으로 가장 높은 성능을 보여 문장 수준의 임베딩을 활용한 모델의 성능이 매우 우수함을 확인하였다.

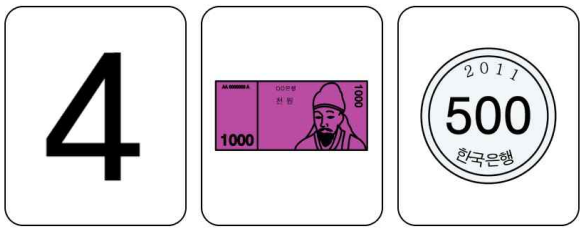
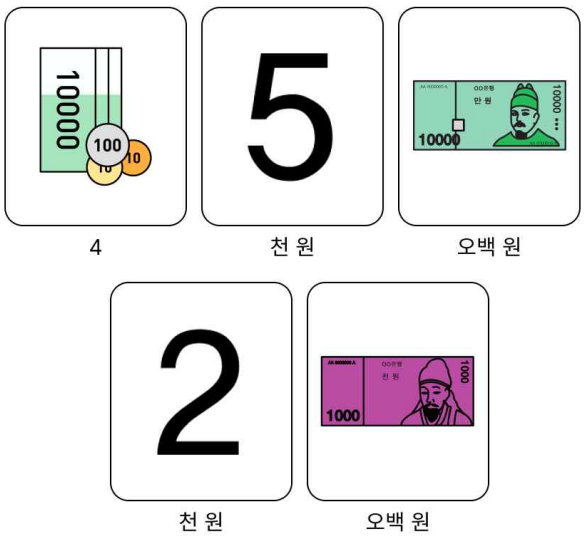


[그림 2] 선행 연구 모델의 BLEU 점수[2]

또한, [2] 연구에서는 시퀀스-투-시퀀스 모델을 학습한 결과를 토대로 숫자 데이터를 처리하는 데 취약함을 확인하였다[표 3]. AAC 상징 '4', '천 원', '오백 원'의 시퀀스를 각 모델에 입력한 결과, BART, GPT3[18]

기반 모델에서는 '사천오백 원'으로 정확하게 변환하였으나, BERT, GPT2[32] 기반 모델의 경우, 다른 숫자로 변환하였다.

[표 3] 숫자 데이터를 포함하는 AAC 상징 시퀀스의 한국어 문장 변환 결과[2]

AAC 상징 시퀀스	모델별 한국어 문장 변환 결과
	**참조 문장 :** 4500원입니다. **BERT :** 4100원입니다. **BART :** 사천오백원입니다 **GPT2 :** 5500원입니다. **GPT3 :** 4500원입니다.
	**참조 문장 :** 현금으로 하면 52000원에 해드릴게요 **BERT :** 현금 하면 5만 2000원입니다. **BART :** 현금으로 하시면 5만 2000원이요. **GPT2 :** 현금하면 7만 3천원이요. **GPT3 :** 현금하시면 십이만원입니다.

3. 대규모 언어 모델

시간의 흐름에 따라서 순서대로 연속적인 값을 가지는 시계열 데이터는 시간이나 순서에 따라 의존성을 가지는 정보의 집합으로 시퀀스 데이터(Sequence Data)[11,13]라 불리며, 자연어 처리, 음성 인식, 금융, 기상 정보 등의 다양한 분야에서 다루어진다. 이러한 시퀀스 데이터의 패턴과 구조를 효과적으로 학습하기 위한 시퀀스 모델이 개발되어 왔다. 시퀀스 모델은 입력 데이터의 순차적 특성을 유지하며 정보를 처리할 수 있는 구조로 이루어져 있으며, 대표적으로 순환 신경망(Recurrent Neural Network, RNN)[7,19,36,51], 장단기 메모리(Long Short-Term Memory, LSTM)[23,36,51], 게이트 순환 유닛(Gated Recurrent Unit, GRU)[19]과 같은 신경망이 있다. 최근에는 시퀀스 모델을 활용하여 복잡한 패턴을 학습하도록 발전하고 있으며, 보다 효율적이고 정확한 예측이 가능해졌다. 특히, 트랜스포머(Transformer)[38] 모델이 등장하면서 기존 시퀀스 모델의 문제점을 해결하였다.

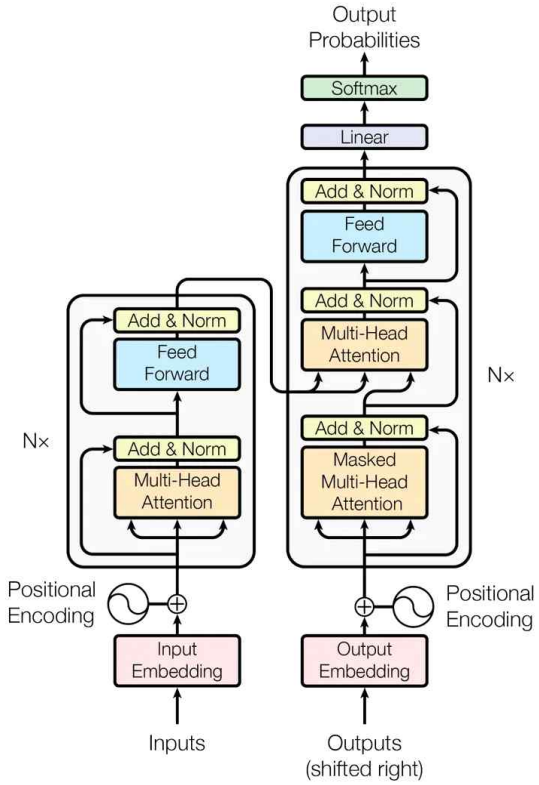
트랜스포머는 자연어 처리 분야에 큰 변화를 가져온 모델 아키텍처로, 기계 번역 등 다양한 자연어 처리 분야 작업의 성능을 크게 향상시켰다. 기존에는 RNN 방식에 기반하여 인코더-디코더[17] 구조의 성능을 높여왔으나, 트랜스포머는 어텐션 아키텍처만을 사용하여 병렬 처리가 가능하며 장기 의존성(Long-Term Dependencies) 문제[34]를 해결하였다. 트랜스포머 아키텍처의 인코더와 디코더는 [그림 3]과 같이 여러 개의 레이어로 구성되며, 포지셔널 인코딩(Positional Encoding), 멀티헤드 어텐션(Multi-Head Attention), 피드 포워드 네트워크(Feed Forward Network)로 이루어져 있다.

트랜스포머의 임베딩 레이어는 텍스트 데이터의 각 단어를 고정된 차원의 벡터 표현으로 변환한다. 트랜스포머의 임베딩 레이어는 의미적으로 유사한 단어를 벡터 공간에서 가까운 위치에 놓이도록 학습하여 변환한다. 트랜스포머는 시퀀스 데이터를 순차적으로 처리하지 않기 때문에 임베딩 벡터에 위치 정보를 더하는 포지셔널 인코딩을 통해 트랜스포머 모델이 입력된 시퀀스 데이터의 위치, 데이터 간의 거리를 파악하도록 한다.

트랜스포머의 인코더는 입력 시퀀스를 받아 이해하고 이를 고차원 벡터로 변환하는 역할을 수행한다. 트랜스포머의 인코더는 다층 구조로 이루어져 있으며 각 층은 [그림 3]의 좌측에 위치한 인코더 구조와 같이 셀프 어텐션(Self-Attention)과 피드 포워드 네트워크(Feed-Forward Network)로 구성된다. 인코더는 입력 시퀀스의 각 위치에 대해 다른 모든 위치와의 관계를 학습하여 입력 시퀀스의 중요한 정보를 추출하고 문맥을 반영한 벡터 표현을 생성한다. 각 인코더 층의 셀프 어텐션 결과는 피드 포워드 신경망에 입력되어 각 토큰의 고차원 벡터 표현을 변환하고 활성화 함수를 통하여 비선형성을 부여받는다.

트랜스포머의 디코더는 입력 시퀀스와 출력 시퀀스 간의 관계를 학습하고 문맥에 맞는 출력을 생성한다. 디코더 또한 다층 구조이며, [그림 3]의 우측에 위치한 디코더 구조와 같이 마스크드 셀프 어텐션(Masked Self-Attention), 멀티 헤드 어텐션(Multi-Head Attention), 피드 포워드 신경망으로 구성된다. 디코더는 마스크드 셀프 어텐션 레이어를 통하여 현재 토큰이 이전 토큰만을 참고하도록 하며, 미래의 토큰에 대한 접근을 제한하여 이전 토큰을 기반으로 다음 토큰을 예측하는 방식으로 작동한

다. 멀티 헤드 어텐션 레이어는 출력 시퀀스를 생성하는 데 필요한 입력 시퀀스와의 관계를 병렬적으로 학습할 수 있도록 한다. 피드 포워드 네트워크는 디코더에서 각 단어의 표현을 더욱 정교하게 하기 위하여 선형 변환과 활성화 함수를 적용한다.



[그림 3] 트랜스포머(Transformer) 구조[4]

트랜스포머 아키텍처의 핵심 요소인 어텐션 메커니즘은 입력 시퀀스 내 토큰 간의 관계를 학습하여 중요한 토큰에 집중하게 함으로써 순환 신경망 기반 시퀀스-투-시퀀스 구조의 성능을 높이고자 고안된 방식이다. 특정 토큰이 시퀀스 내에서 다른 토큰들과 가지는 관련성을 가중치로 표현하며, 이는 특정 토큰의 중요도를 의미한다. 어텐션 메커니즘은 각

입력 요소에 대해 생성한 쿼리(Query) 벡터, 키(Key) 벡터, 값(Value) 벡터로부터 가중치를 구하여 중요한 정보에 집중한 벡터 표현을 생성한다.

셀프 어텐션은 어텐션 메커니즘을 확장한 개념으로, 시퀀스 내의 각 토큰과 다른 토큰 간의 관련성을 계산하는 메커니즘이다. 쿼리는 현재 단어의 정보를 반영하고, 키는 전체 시퀀스 내의 각 단어가 가지고 있는 정보를 나타내며, 값은 실질적인 정보 전달을 담당한다. 쿼리와 키 간 내적을 통해 유사도를 계산하며, 이렇게 계산된 유사도는 소프트맥스(Softmax) 함수를 통해 확률 분포로 변환되어 각 벡터에 가중치를 부여한다. 이를 통해 모델은 현재 단어가 문맥 내에서 얼마나 중요한지 결정하고 문장의 전반적인 관계를 학습한다.

대규모 언어 모델(Large Language Model, LLM)은 수억에서 수천억 개 이상의 파라미터를 방대한 양의 데이터로 학습하여 자연어를 이해하고 생성하는 데 활용되는 인공지능 모델이다. 대규모 언어 모델은 주로 셀프 어텐션 메커니즘을 활용한 트랜스포머 아키텍처 기반 모델이며, 기본적인 언어적 패턴, 의미, 문맥을 파악하고 있어 이를 바탕으로 자연어 생성, 기계 번역, 질의 응답, 대화형 챗봇, 감정 분석, 요약 등 다양한 자연어 처리 작업을 수행할 수 있다. 대표적인 대규모 언어 모델로는 GPT 계열 모델[18,31,32], T5[33], BERT[22] 등이 있으며, 대규모 텍스트 데이터셋을 사용하여 일반적인 언어 패턴을 학습하도록 사전 훈련된 대규모 언어 모델의 사전 학습 모델(Pre-trained Model)을 활용하여 특정 도메인의 작업에 최적화할 수 있다. 본 연구에서는 AAC 상징 시퀀스의 한국어 문장 변환을 위하여 GPT-2와 T5의 한국어 모델을 활용하였다.

1) GPT-2

GPT[31]는 OpenAI에서 개발한 자가 회귀 언어 모델로, 트랜스포머 디코더 부분만을 사용하여 문장의 다음 단어를 예측하는 방식으로 학습된다. GPT-2[32,54]는 8백 만개의 대규모 웹 페이지 텍스트[54]를 활용하여 학습되었으며, 학습 데이터의 특성으로 인해 일상적 문장 생성과 맥락 유지에 탁월한 성능을 보인다. GPT는 주로 언어 생성 및 요약, 대화 생성 등의 자연어 생성 작업에 특화되어 있으며, 질문 응답, 텍스트 요약 등에서도 우수한 성능을 발휘한다. 그러나 GPT의 단방향 학습 방식은 문맥의 양방향성을 활용할 수 없다는 한계를 가진다.

2) T5

T5[33]는 구글(Google)에서 개발한 언어 모델로, 트랜스포머의 인코더-디코더 구조를 기반으로 한다. T5 모델은 위키피디아, C4(OColossal Clean Crawled Corpus)[50] 데이터셋 등을 기반으로 대규모 학습을 수행하였다. T5는 입력과 출력을 모두 텍스트로 변환하는 방식으로 요약, 번역, 질문-응답, 문장 완성 등 전반적인 텍스트 작업에서 활용되며, 텍스트-텍스트 포맷으로 인하여 다양한 작업에 유연하게 적용할 수 있다는 장점이 있다. 또한 대규모 학습 데이터를 통하여 성능을 향상시켰지만, 모델 크기에 따라 많은 양의 연산 자원을 필요로 한다는 단점이 있다.

4. QLoRA(Quantized Low-Rank Adaptation) 미세 조정

전이 학습(Transfer Learning)[15,21,42]은 하나의 작업에서 학습한 지식을 다른 작업에 적용하여 모델의 성능을 개선하도록 하는 학습 기법이며, 훈련 데이터가 제한적일 때 특히 효과적이다. 대규모 데이터셋에서 사전 학습(Pre-Training)하여 일반적인 언어의 패턴, 문맥, 구조를 학습한 모델은 다른 작업에 대한 별도의 데이터 학습 없이도 기본적인 이해 능력을 갖춘다. 이는 대규모 데이터가 필요한 자연어 처리 및 컴퓨터 비전 등의 분야에서 광범위하게 활용된다.

미세 조정(Fine-Tuning)은 사전 학습된 모델을 특정 도메인의 특정 작업에 맞추어 성능을 최적화하는 과정이다. 미세 조정은 사전 학습 모델을 새로운 작업에서도 높은 성능을 발휘하도록 사전 학습 모델의 파라미터를 특정 데이터셋에 맞춰 학습한다. 예를 들어, 대규모 언어 모델인 GPT 계열의 모델을 대규모 텍스트 데이터셋으로 사전 학습한 후 금융, 의료 등 특정 도메인의 데이터셋에 맞추어 파라미터를 미세 조정하여 주가 예측 등 특정 작업에 적합한 모델로 학습할 수 있다.

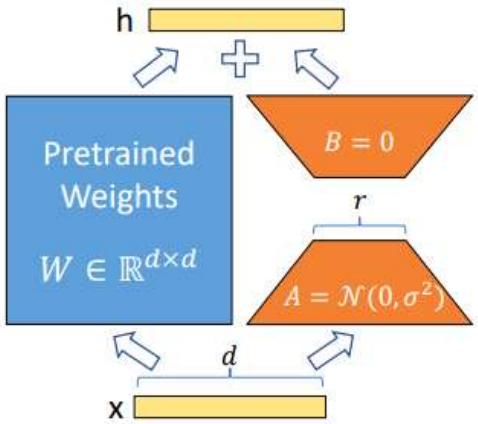
미세 조정 방식은 크게 전체 미세 조정(Full Fine-Tuning)과 PEFT(Parameter Efficient Fine-Tuning)[8,40] 방식으로 구분할 수 있다. 전체 미세 조정은 사전 학습된 모델의 모든 파라미터를 특정 작업에 맞춰 조정하는 방식이다. 일반적으로, 모델이 특정 작업의 데이터셋에 대한 미세 조정을 수행할 때 전체 미세 조정을 통하여 모델이 새로운 작업을 수행하도록 한다. 모든 파라미터를 조정함으로써 특정한 작업에 대한 모델의 성능을 극대화할 수 있으나 모델의 방대한 파라미터 수로 인하여

매우 많은 연산 자원을 요구한다는 단점이 있다. 또한, 모델의 크기가 지속적으로 커지고 있다는 점에서 모든 파라미터를 조정하는 것은 더욱 불가능해지고 있다. 따라서, 미세 조정 과정에서 요구하는 시간적 비용과 물리적 자원으로 인하여 모바일 기기와 같은 제한된 환경, 혹은 일반적인 물리적 연산 자원 환경에서의 개발 혹은 실시간 응용에는 제약이 따른다.

PEFT는 전체 미세 조정 방식과 달리 모델의 특정 파라미터 집합 또는 계층만을 선택적으로 조정하는 미세 조정 방식이다. 학습 과정에서 다운스트림 데이터셋에 따라 일부 파라미터만을 조정하기 때문에 적은 양의 데이터와 연산 자원으로도 모델의 성능을 높일 수 있다. 적은 수의 파라미터를 학습하는 것만으로도 모델의 전체 파라미터를 미세 조정하는 것과 유사한 효과를 가지며, 사전 학습된 모델의 정보를 미세 조정 과정에서 잊어버리는 파괴적 망각(Catastrophic Forgetting)[28] 문제를 방지하는 차원에서 전체 미세 조정보다 뛰어난 성능을 갖추는 경우도 있다.

대표적인 PEFT 방식으로는 LoRA(Low-Rank Adaptation)[25], QLoRA(Quantized Low-Rank Adaptation)[16,20,29] 등이 있으며 각각의 방식에 따라 특정 파라미터를 조정한다. LoRA 방식은 [그림 4]와 같이 기존 모델의 파라미터(Pretrained Weights)를 고정하고, 작은 행렬(A, B)을 추가하여 적은 양의 파라미터만을 학습한다. LoRA는 기존 모델의 파라미터를 고정함으로써 기존의 모델이 학습한 특성을 손상시키지 않고도 다양한 작업에 적합하게 파라미터를 학습할 수 있도록 한다. LoRA는 매우 많은 파라미터 수를 가지는 대규모 언어 모델을 학습하기 위하여 요구되는 물리적 연산 자원의 양을 매우 감소시켜 연산 자원에 제한이 있는 환경에서 대형 모델의 미세 조정을 수행하는 데 적합하다.

QLoRA[20]는 LoRA에서 변형된 방식으로, 저차원 행렬에 양자화 기법을 적용하여 학습 시 사용되는 연산 자원의 양을 더욱 절약하는 것과 동시에 LoRA의 효율성을 극대화한 미세 조정 방식이다. QLoRA는 [그림 4]의 LoRA와 동일하게 기존 모델의 파라미터(Pretrained Weights)를 고정하면서 모델 내 일부 파라미터의 차원을 줄이는 방식으로 동작한다. QLoRA는 LoRA 방식과 함께 4-bit 양자화 기법을 도입함으로써 학습하는 파라미터 수를 더욱 줄여 메모리 효율을 높임과 동시에 성능을 유지할 수 있다. 양자화는 모델의 가중치를 낮은 비트로 표현하여 모델 크기를 줄이되 중요한 정보 손실을 최소화하는 것을 목표로 한다. 이러한 점에서 QLoRA 미세 조정 방식은 특히 자원이 제한된 환경에서 대규모 모델의 미세 조정을 지원한다.



[그림 4] LoRA 미세 조정 학습 방식[25]

III. 대규모 언어 모델 기반 상징 시퀀스의 한국어 문장 변환

3장에서는 대규모 언어 모델을 활용하여 AAC 상징 시퀀스의 한국어 문장 변환을 수행하도록 설계한 모델을 제시한다.

1. 데이터 전처리

본 연구의 실험은 AI Hub에서 제공하는 한국어 대화 데이터[48]와 KETI 공개 데이터[47] 중 일상 데이터, AAC 사용자의 주사용 어휘와 문장 관련 연구[1,3,6,10]에서 수집된 한국어 문장과, 이에 대응되는 한국형 AAC 상징 체계집[41]의 AAC 그림 상징을 통해 구축된 상징 시퀀스 데이터 셋[13]으로 구성된 총 12,844개의 데이터를 사용하였다[표 4]. 학습에 활용하기 위하여 12,844개의 데이터 전처리 작업으로 중복되는 데이터와 이상치를 제거하였다.

[표 4] 학습 데이터 수 (단위: 개)

데이터 출처	데이터 수
AI HUB	11,179
AAC 사용자 연구 논문	1,665
합계	12,844

2. 대규모 언어 모델 기반 상징 시퀀스의 한국어 문장 변환 모델

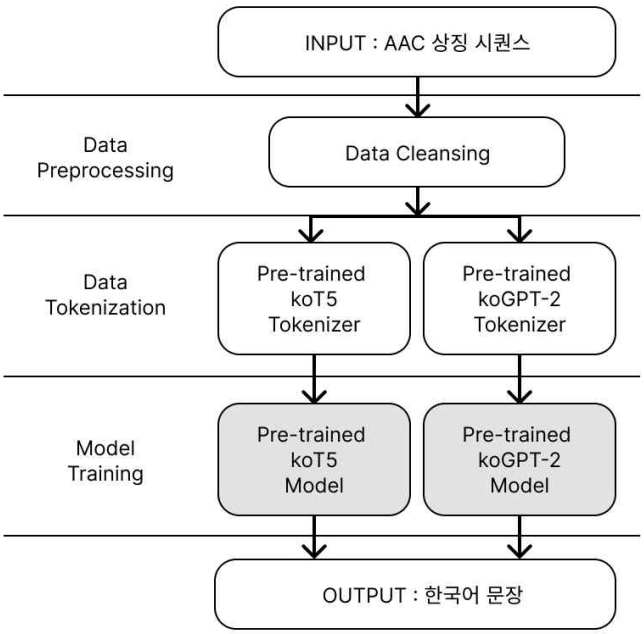
본 연구에서는 AAC 상징 시퀀스의 한국어 문장 변환을 위하여 GPT-2와 T5 대규모 언어 모델의 한국어 모델인 koGPT-2[56]와 koT5[55]를 활용하였다. [표 5]와 같이 SKT-AI 사에서 제공하는 GPT-2 한국어 모델인 kogpt2-base-v2 모델과 PAUST 사에서 제공하는 pko-t5-base 모델을 활용하였다. kogpt2는 1억 2천5백만 개의 파라미터 수를 가지며 한국어 위키 백과, 뉴스, 모두의 말뭉치 등 대규모 데이터셋에 대하여 학습된 모델이다. pko-t5는 2억 5천만 개의 파라미터 수를 가지며 한국어 나무위키, 위키피디아, 모두의 말뭉치 등 대규모 학습 데이터셋을 활용하여 사전 학습된 모델이다.

[표 5] 한국어 대규모 언어 모델

모델	제공	년도	학습 데이터셋	파라미터 수
kogpt2-base-v2	SKT-AI	2020	한국어 위키 백과, 뉴스, 모두의 말뭉치 등	125M
pko-t5-base	PAUST	2022	한국어 나무위키, 위키피디아, 모두의 말뭉치 등	250M

두 가지 모델을 활용한 실험의 데이터 처리 및 학습 흐름은 [그림 5]와 같다. 데이터는 표현하고자 하는 참조 문장과 AAC 상징 시퀀스의 상징 표현이 묶인 문자열로 이루어져 있다. 이를 사전 학습 모델이 입력으로 처리 가능한 토큰의 시퀀스로 변환하기 위하여 각 모델이 제공하는

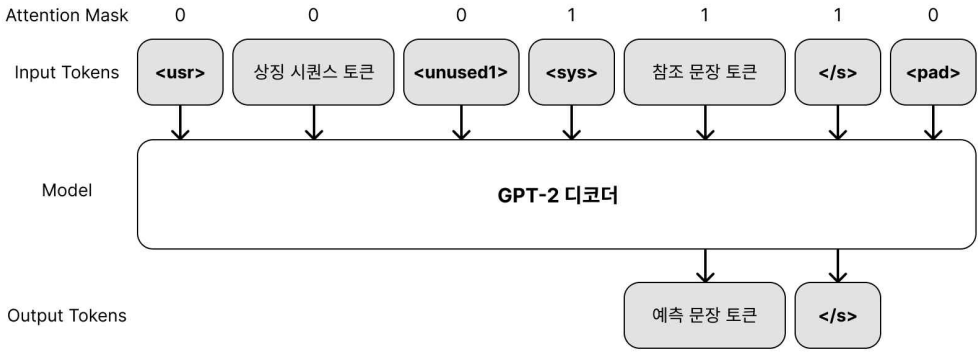
사전 학습 토크나이저(Pre-trained Tokenizer)를 활용하였다. 토큰화된 AAC 상징 시퀀스 데이터는 각 사전 학습 모델에 입력되며, 예측된 한국어 문장을 출력한다.



[그림 5] 대규모 언어 모델 기반 상징 시퀀스의 한국어 문장 변환 모델

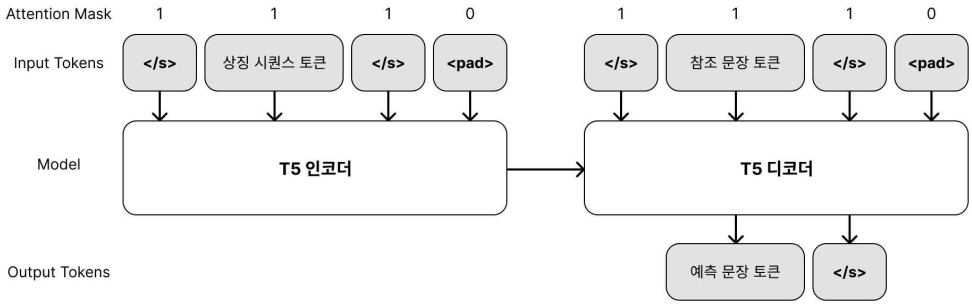
GPT-2와 T5 모델은 다른 모델 구조를 가지고 있다. GPT-2 모델은 디코더만으로 구성되어 있으며, T5 모델은 인코더와 디코더로 이루어져 있다. 상징 시퀀스의 토큰과 참조 문장 토큰 간의 관계를 학습하기 위하여 상이한 모델 구조에 따라 다른 방식으로 학습을 수행하여야 한다. GPT-2 기반 모델의 학습 방식은 [그림 6]과 같다. GPT 계열의 모델은 디코더만으로 구성 되어있기 때문에 상징 시퀀스의 토큰과 참조 문장의 토큰을 함께 디코더 입력으로 하며, 상징 시퀀스 토큰과 참조 문장 토큰

은 <usr>, <unused1>, <sys> 등 특수 토큰을 활용하여 구분하도록 학습하였다. 제안하는 모델의 학습 방식은 질문-응답에 대한 학습과 유사한 방식으로 <usr>는 상징 시퀀스 토큰이 입력됨을, <unused1>은 상징 시퀀스 토큰이 종료되었음을 명시한다. <sys> 특수 토큰은 참조 문장 토큰의 시작을 명시하여 모델이 상징 시퀀스 입력에 따라 예측한 토큰을 출력하도록 한다. 입력 토큰에 대한 어텐션 마스크는 참조 문장 토큰과 관련된 토큰에 대해서만 1로 처리하여 예측 문장 토큰을 출력하도록 하였다.



[그림 6] GPT-2 기반 모델 학습 방식

T5 모델 기반 모델의 학습은 [그림 7]과 같이 수행하였다. T5 모델은 입력을 이해하는 인코더와 그에 대한 예측을 수행하는 디코더로 구성되어 있다. 따라서 인코더의 입력은 상징 시퀀스 토큰을, 디코더의 입력은 참조 문장의 토큰으로 하여 상징 시퀀스와 참조 문장의 관계를 학습하도록 하였다. 또한, 인코더와 디코더 입력에 따라 각 레이어가 집중해야 할 토큰의 어텐션 마스크를 1로 설정하였다.



[그림 7] T5 기반 모델 학습 방식

IV. QLoRA 방식 기반 상징 시퀀스의 한국어 문장 변환 모델 미세 조정

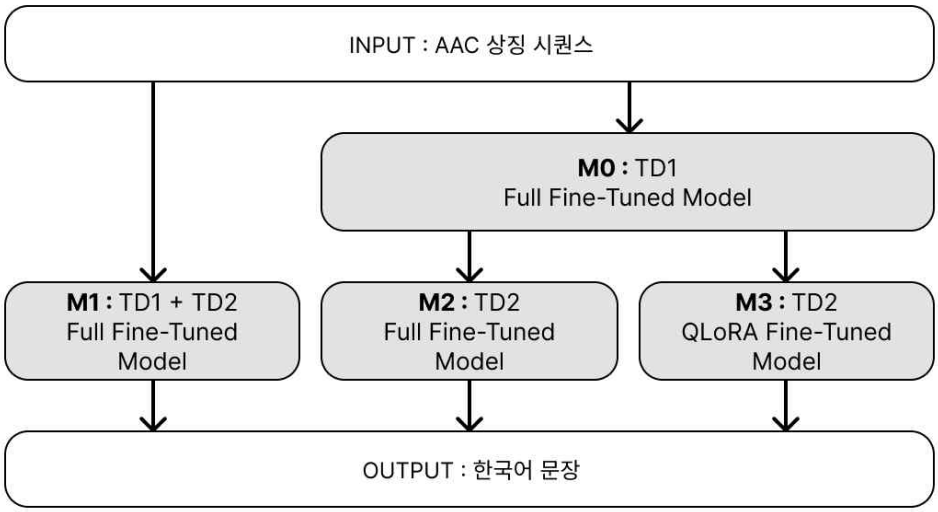
주기적인 모델 업데이트는 단순히 새로운 데이터를 추가 반영하는 것뿐만 아니라 모델의 성능을 정기적으로 평가하고, 개선하는 과정으로서 서비스 품질의 향상과 사용자 경험 개선에 중요한 역할을 한다. 사용자로부터 수집되는 데이터는 실시간 피드백을 제공하며, 모델이 사용자의 개별적인 요구와 다양한 상황에 더 잘 적응하도록 돕는다. 따라서 AAC 사용자의 활용 데이터를 수집하고, 수집한 데이터를 기존 모델에 반영할 수 있는 효율적인 학습 시스템이 구축되어야 한다. 이를 통하여 모델이 지속적으로 변화하는 AAC 사용자의 요구를 반영하여, 정확도를 향상시키며 오류를 줄일 수 있다. AAC 상징 시퀀스를 한국어 문장으로 변환하는 모델이 AAC 사용자에게 유의미한 결과를 제공하고, 자연스러운 문장을 통하여 AAC 사용자의 의사소통을 지원하기 위하여 새롭게 수집된 데이터를 모델에 지속적으로 반영하는 것은 필수적이다. 따라서, AAC 상징 시퀀스의 한국어 문장 변환 모델에 추가된 데이터를 반영하기 위하여 모델의 파라미터를 추가 학습하여야 한다.

본 연구에서는 기존 모델이 추가 수집된 데이터를 반영하도록 추가 학습하는 실험을 수행하기 위하여 데이터를 다음과 같이 설정하였다. 학습 데이터(Train Data), 검증 데이터(Validation Data), 시험 데이터(Test Data)의 비율은 8:1:1로 유지하며, 학습 데이터를 기존 학습 데이터(TD1)와 추가 학습 데이터(TD2)로 분리하였다. 기존 데이터(TD1)와 새로운 데이터(TD2)는 각각 전체 학습 데이터의 80%와 20%로 분리하였다. 새

롭게 수집된 데이터셋에 대한 추가 학습을 위하여 수행하는 미세 조정 방식별 구분은 다음과 같다.

- **M0** : 사전 학습 모델을 기존 학습 데이터(TD1)로 학습한 모델
- **M1** : 기존 학습 데이터(TD1)와 추가 학습 데이터(TD2)를 모두 활용하여 기존 사전 학습 모델을 전체 미세 조정한 모델
- **M2** : M0 모델을 추가 학습 데이터(TD2)로 전체 미세 조정(Full Fine-Tuning)한 모델
- **M3** : M0 모델을 추가 학습 데이터(TD2)로 QLoRA 미세 조정한 모델

[그림 8]과 같이 기존 학습 데이터셋에 대하여 학습한 모델 M0를 초기 모델로 하며, 새롭게 추가된 학습 데이터를 어떤 미세 조정 방식을 활용하여 추가 학습하였는지에 따라 M1부터 M3까지로 분리하였다. M1 모델은 LLM의 사전 학습 모델을 기존 학습 데이터(TD1)와 추가 학습 데이터(TD2)를 합친 데이터셋으로 전체 미세 조정한 모델이다. M2 모델은 M0 모델을 추가 학습 데이터(TD2)만으로 전체 미세 조정한 모델(Full Fine-Tuned Model)이고, M3 모델은 M0 모델을 추가 학습 데이터(TD2)만으로 QLoRA 미세 조정한 모델이다. 축적되는 데이터에 대한 미세 조정 방식에 따른 M1, M2, M3 모델의 성능과 학습에 활용된 연산 자원 사용량을 비교한다.



[그림 8] 추가 학습 데이터 학습을 위한 미세 조정 방식에 따른 상징 시퀀스의 한국어 문장 변환 성능 비교 실험

또한, 기존 학습 데이터(TD1)와 추가 학습 데이터(TD2)의 비율별 미세 조정 성능 비교를 위하여 데이터셋 비율에 따라 R1부터 R5까지 구분하였다. 구분에 따른 각 학습 데이터의 개수는 [표 6]과 같다. 데이터셋 구분에 따라 기존 학습 데이터(TD1)는 10%씩 증가하도록, 추가 학습 데이터(TD2)는 10%씩 감소하도록 설정하였다. 여기서 R4의 경우 앞선 미세 조정 방식에 따른 상징 시퀀스의 한국어 문장 변환 성능 비교 실험 시 기존 학습 데이터(TD1)와 추가 학습 데이터(TD2)의 비율이 동일하다.

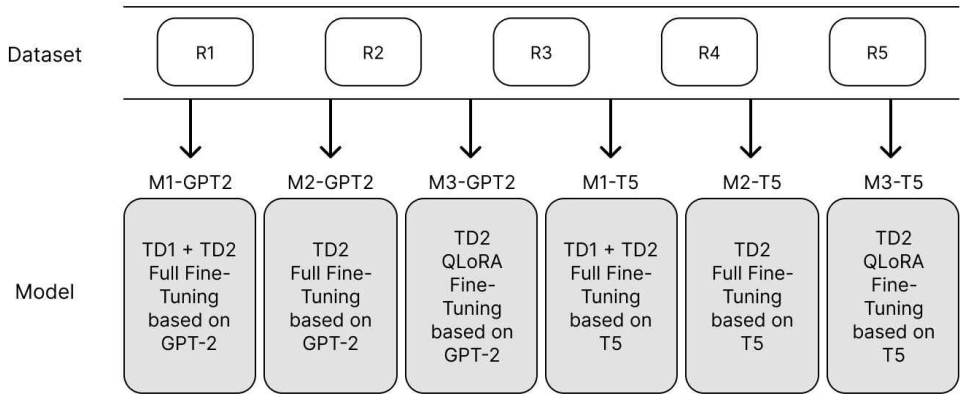
[표 6] 기존 학습 데이터와 추가 학습 데이터의 비율에 따른 구분 및 데이터의 개수

데이터셋 구분	R1	R2	R3	R4	R5
기존 학습 데이터(TD1) 개수 (비율)	5,137 (50%)	6,165 (60%)	7,192 (70%)	8,220 (80%)	9,247 (90%)
추가 학습 데이터(TD2) 개수 (비율)	5,138 (50%)	4,110 (40%)	3,083 (30%)	2,055 (20%)	1,028 (10%)

기존 학습 데이터셋(TD1)과 추가 학습 데이터셋(TD2) 비율에 따라 구분된 R1부터 R5까지의 데이터셋을 [그림 9]와 같이 대규모 언어 모델과 미세 조정 방식에 따라 구분되는 모델에 학습하여 성능을 비교한다. 대규모 언어 모델의 종류와 미세 조정 방식에 따른 모델은 다음과 같이 구분하였다.

- **M0-GPT2** : GPT-2 한국어 사전 학습 모델을 기존 학습 데이터(TD1)으로 전체 미세 조정(Full Fine-Tuning)한 초기 모델이다.
- **M1-GPT2** : GPT-2 한국어 사전 학습 모델을 기존 학습 데이터(TD1)와 추가 학습 데이터(TD2)를 합친 데이터셋으로 전체 미세 조정(Full Fine-Tuning)한 모델이다.
- **M2-GPT2** : M0-GPT2 모델을 추가 학습 데이터(TD2)만으로 전체 미세 조정(Full Fine-Tuning)한 모델이다.
- **M3-GPT2** : M0-GPT2 모델을 추가 학습 데이터(TD2)만으로 QLoRA 미세 조정한 모델이다.

- **M0-T5** : T5 한국어 사전 학습 모델을 기존 학습 데이터(TD1)으로 전체 미세 조정(Full Fine-Tuning)한 초기 모델이다.
- **M1-T5** : T5 한국어 사전 학습 모델을 기존 학습 데이터(TD1)와 추가 학습 데이터(TD2)를 합친 데이터셋으로 전체 미세 조정(Full Fine-Tuning)한 모델이다.
- **M2-T5** : M0-T5 모델을 추가 학습 데이터(TD2)만으로 전체 미세 조정(Full Fine-Tuning)한 모델이다.
- **M3-T5** : M0-T5 모델을 추가 학습 데이터(TD2)만으로 QLoRA 미세 조정한 모델이다.



[그림 9] 기존 학습 데이터와 추가 학습 데이터 비율별 미세 조정 방식에 따른 성능 비교 실험 설계

V. 모델 실험 및 평가

5장에서는 본 연구에서 수행한 실험과 평가 결과를 기술하며, 실험 내용과 설계, 실험 결과와 평가 지표별 성능 수치 등을 산출하고 분석한다. 본 연구에서 수행한 실험은 다음과 같다.

- 대규모 언어 모델(LLM)을 활용한 AAC 상징 시퀀스의 한국어 문장 변환 실험
- 추가 학습 데이터(TD2)에 대한 미세 조정 방식별 성능 및 연산 자원(GPU) 사용량 비교 실험
- 기존 학습 데이터(TD1)와 추가 학습 데이터(TD2)의 비율별 미세 조정 성능 비교 실험

1. 실험 설계

본 연구의 실험 환경은 [표 7]과 같다. CPU는 Intel 프로세서를 사용하였으며, 64GB RAM, GPU는 NVIDIA의 GeForce Titan X를 활용하였다. 소프트웨어 환경으로 운영체제는 Linux Ubuntu Desktop 22.04 LTS를 사용하였고, Python 3.9 버전과 TensorFlow 2.14.0, CUDA 11.8, cuDNN 8.7 환경에서 모델 개발 및 학습을 수행하였다.

[표 7] 실험 환경

분류		환경
H/W	CPU	Intel
	RAM	64.0GB
	GPU	GeForce TITAN X 12GB
S/W	OS	Linux Ubuntu Desktop 22.04 LTS
	Python	3.9
	Tensorflow	2.14.0
	CUDA	11.8
	cuDNN	8.7

학습 단계에서 손실 함수는 예측 문장과 참조 문장 간의 차이를 측정한다. 본 연구에서는 모델이 예측한 문장과 참조 문장 간의 유사성을 확인하기 위한 손실 함수로 CrossEntropyLoss[52]를 설정하였다. CrossEntropyLoss는 단어 단위의 오류를 산출하여 전체 문장에 대한 손실을 계산하기 때문에 AAC 상징 시퀀스의 한국어 문장 변환 과정에서의 문맥적 일관성과 자연스러움을 강화할 수 있다.

최적화 알고리즘으로는 AdamW[49] 최적화 알고리즘을 활용하였다. AdamW는 Adam(Adaptative Moment Estimation)의 원리를 기반으로 하며 오차에 제곱을 활용한 L2 규제를 적용함으로써 과적합을 방지하고, 특정 가중치가 너무 커지는 것을 방지하는 가중치 감소(Weight Decay)가 가능한 최적화 알고리즘이다. 또한, AdamW는 파라미터 업데이트 시, 가속도를 이용하는 모멘텀(Momentum)과 적응적 학습률(Adaptive

Learning Rate)을 활용하여 빠르고 안정적인 학습을 가능하게 한다.

학습의 최적화를 위하여 본 연구에서 설정한 모델별 하이퍼파라미터는 [표 8]과 같다. GPT-2 기반 모델의 학습률은 5e-5, T5 기반 모델의 학습률은 1e-6로 모델별 학습 속도에 따라 다르게 설정하였다. 두 모델 모두 학습 최대 에폭(Epoch) 수는 100으로 설정하였고, 산출된 손실 기준으로 20 에폭 동안 개선이 없을 경우 조기 종료(Early Stopping)하도록 설정하였으며 배치 크기는 32로 동일하게 설정하였다.

[표 8] 모델별 하이퍼파라미터 설정

하이퍼파라미터	GPT-2	T5
학습률	5e-5	1e-6
에폭(Epoch) 수	100	
배치 크기	32	
최적화 알고리즘	AdamW	
손실 함수	CrossEntropyLoss	
조기 종료 에폭 기준	20	

QLoRA 미세 조정을 적용하기 위하여 각 모델별 LoRA 모듈에 대한 하이퍼파라미터를 [표 9]와 같이 설정하였다. r은 LoRA에서 추가된 저랭크 행렬의 랭크를 의미하며, GPT-2와 T5 모델 모두 16을 적용하였다. r 값이 클수록 학습에 활용하는 파라미터 수가 증가하며, 메모리 사용량이 증가한다. lora_alpha는 LoRA 어댑터의 스케일링 값으로 어댑터로부터 나온 출력 값에 곱해지는 값을 의미하며, 두 모델 모두 32로 설정하였다. GPT-2 모델의 경우, 타겟 모듈을 c_attn과 c_proj 모듈로 설정하였다. c_attn 모듈은 GPT 모델의 쿼리(Query), 키(Key), 값(Value)를 처리하는

부분이며, c_proj 모듈은 어텐션 출력 값을 후처리하는 부분이다. T5 모델의 경우, GPT-2와 같이 쿼리와 값을 처리하는 q, v 모듈을 타겟 모듈로 설정하였다. GPT-2 모델의 경우 드롭아웃 비율을 0.1, T5 모델의 경우 드롭아웃 비율을 0.01로 설정하여 학습 시 일부 가중치를 무작위로 비활성화하여 특정한 패턴에 대한 모델의 과적합을 방지하였다. 두 모델의 편향은 업데이트하지 않도록 설정하였으며, GPT-2 모델은 디코더로만 구성되어 있어 CAUSAL_LM으로, T5 모델은 인코더-디코더로 구성되어 있어 SEQ_2_SEQ_LM으로 설정하여 LoRA가 각 유형에 맞추어 작업할 수 있도록 하이퍼파라미터를 설정하였다.

[표 9] 모델별 LoRA 하이퍼파라미터 설정

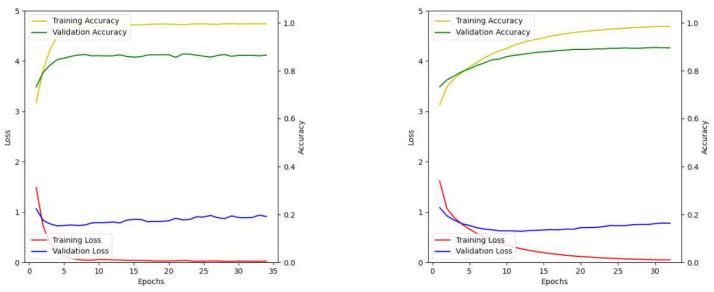
하이퍼파라미터	GPT-2	T5
r	16	
lora_alpha	32	
target_modules	c_attn, c_proj	q, v
lora_dropout	0.1	0.01
bias	none	
task_type	CAUSAL_LM	SEQ_2_SEQ_LM

2. 실험 결과

본 연구에서 수행한 대규모 언어 모델을 활용한 AAC 상징 시퀀스의 한국어 문장 변환 실험, 기존 학습 데이터와 추가 학습 데이터의 비율별 미세 조정 성능 비교 실험에 대한 모델별 학습 그래프 및 실험 결과는 다음과 같다.

- **대규모 언어 모델(LLM)을 활용한 AAC 상징 시퀀스의 한국어 문장 변환 실험**

대규모 언어 모델을 활용한 AAC 상징 시퀀스의 한국어 문장 변환을 실험하기 위하여 학습, 검증과 시험은 전체 데이터셋의 80%, 10%, 10%로 분류하여 수행하였다. [그림 10]은 대규모 언어 모델을 활용한 AAC 상징 시퀀스의 한국어 문장 변환 실험에 대한 학습 그래프이다. GPT-2 기반 모델과 T5 기반 모델이 모두 30-35 에폭 안에서 학습을 종료하였으며, GPT-2 기반 모델의 학습 그래프(a)가 T5 기반 모델의 학습 그래프(b)보다 급격한 경사를 보였다.



(a) GPT-2 기반 모델 학습 그래프 (b) T5 기반 모델 학습 그래프

[그림 10] 모델별 AAC 상징 시퀀스의 한국어 문장 변환 학습 그래프

조기 종료를 적용한 학습 종료 후 모델별 실험 결과는 [표 10]과 같다. GPT-2 기반 모델의 학습 손실은 0.0162, 정확도는 0.9968, 검증 손실은 0.8622, 정확도 0.8692이며, T5 기반 모델의 학습 손실은 0.0470, 정확도는 0.9848, 검증 손실 0.8073, 정확도 0.8948로 학습이 종료되었다. GPT-2 모델의 학습 성능이 높았고, T5 모델의 검증 성능이 높았음을 실험을 통하여 확인하였다.

[표 10] 모델별 실험 결과

모델	Train		Validation	
	Loss	Accuracy	Loss	Accuracy
GPT-2	0.0162	0.9968	0.8622	0.8692
T5	0.0470	0.9848	0.8073	0.8948

GPT-2와 T5 사전 학습 모델을 활용한 AAC 상징 시퀀스의 한국어 문장 변환 결과는 [표 11]과 같다. '아메리카노 한 잔 주문할 수 있을까요?' 문장을 표현하기 위한 '아메리카노', '하나', '컵', '주문할래요', '돼요' 상징의 시퀀스는 입력하였을 때, 두 모델은 모두 올바른 문장을 예측하였다 [표 11-(1)].

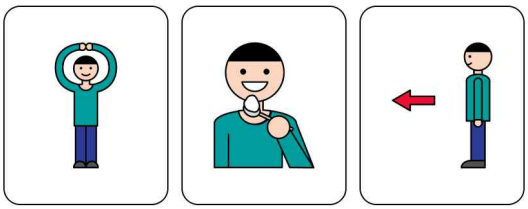
[표 11]의 (2)와 같이 '한우는 6만 원 수입산은 2만 원부터 있습니다' 문장을 표현하기 위해 숫자를 포함하는 9개의 상징 시퀀스를 입력하였을 때, 두 모델 모두 참조 문장과 유사하게 예측하였다. 이를 통하여 긴 상징 시퀀스의 문장 예측에도 유연하며, 숫자 데이터에 대한 처리와 예측이 잘 되는 것을 알 수 있다. 그러나, '한 달하면 학생은 팔만 오천 원 일반은 구만 원입니다' 문장에 대한 상징 시퀀스를 변환하였을 때 예측한 '한

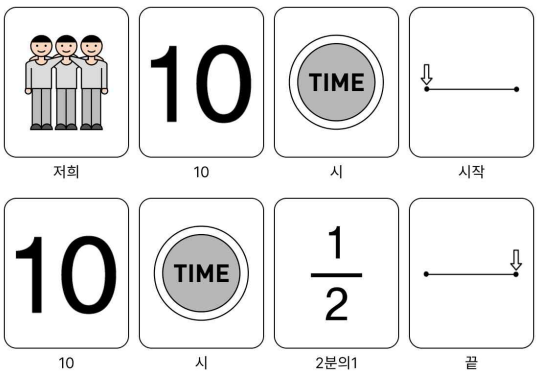
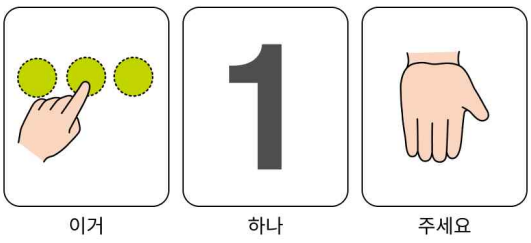
달 학생은 85000원이고요 보통은 91000원인데요' 문장 중 '1000'처럼 포함되지 않은 상징의 표현이 포함되거나[표 11-(5)], '129300원이고 콜라 세 개 서비스입니다' 문장에 대해 '12만 구천 원인데 콜라는 세 개 서비스로 나가요'와 같이 예측하여 '300' 의미를 누락하는 등 포함된 상징의 표현이 누락되는 오류가 존재하였다[표 11-(6)].

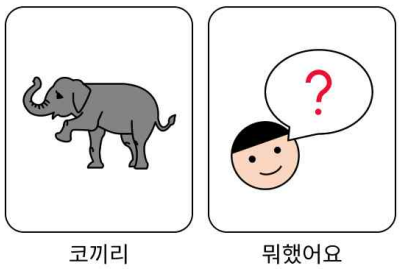
[표 11]의 (3)과 같이 '가격은 어떻게 되나요?' 문장을 표현한 '가격', '어떻게' 상징 시퀀스는 GPT-2의 경우, '가격은 어떻게 되나요?'로 예측하였고, T5의 경우, '가격은 어떻게 해드릴까요?'와 같이 예측하였다. '네 먹고 갈거예요' 문장에 대한 상징 시퀀스를 GPT-2 모델은 '네 드시고 가시나요?'로, T5 모델은 '네 먹고 갈거예요 금방 갑니다'와 같이 예측하였다[표 11-(4)]. 이와 같이 모델에 따라서 문장의 발화 주체를 다르게 설정한 경우도 있어 이에 대한 학습 및 처리가 필요하다. 또한, '이거' 상징 표현을 '고거'로 예측하거나[표 11-(8)], 개수를 세는 단위인 '개' 상징 표현을 '두 판'으로 예측하는[표 11-(9)] 등 오류가 있었고, [표 11]의 (10)과 같이 '코끼리가 뭐하고 있어?' 문장에 대한 상징 시퀀스를 '코끼리가 뭐하고 있어? 코끼리도 뭔가를 준비했는데요?'처럼 하나의 문장이 종료된 후, 다음 문장까지 예측하는 등의 오류가 존재하였다.

[표 11] AAC 상징 시퀀스의 한국어 문장 변환 결과

번호	AAC 상징 시퀀스	모델별 한국어 문장 변환 결과
(1)		**참조 문장 :** 아메리카노 한 잔 주문할 수 있을까요? **GPT-2 :** 아메리카노 한 잔 주문할 수 있을까요? **T5 :** 아메리카노 한 잔 주문할 수 있을까요?
(2)		**참조 문장 :** 한우는 6만 원 수입산은 2만 원부터 있습니다. **GPT-2 :** 한우 6만 원이고 외국산 2만 원도 있어요 **T5 :** 한우 6만 원이고요 외국산 2만 원까지 있습니다
(3)		**참조 문장 :** 가격은 어떻게 되나요? **GPT-2 :** 가격은 어떻게 되나요? **T5 :** 가격은 어떻게 해드릴까요?

(4)		**참조 문장 :** 네 먹고 갈 거예요 **GPT-2 :** 네 드시고 가시나요? **T5 :** 네 먹고 갈거예요 금방 갑니다
(5)		**참조 문장 :** 한 달하면 학생은 팔만 오천 원 일반은 구만 원입니다 **GPT-2 :** 한달 학생들 8만 5천원인데 보통으로 9만 **T5 :** 한달 학생은 85000원이고요 보통은 91000원인데요

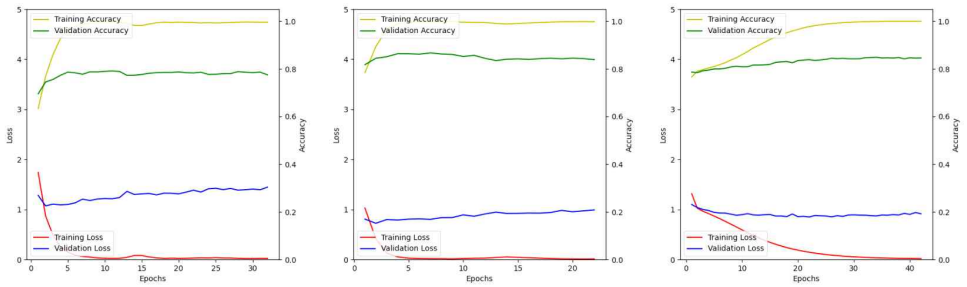
(6)		**참조 문장 :** 129300원이고 콜라 세 개 서비스입니다 **GPT-2 :** 12만 구천 원인데 콜라는 세 개 서비스로 나가요 **T5 :** 12만 9300원이요 콜라 세 개까지 서비스로 드릴게요
(7)		**참조 문장 :** 저희 열시부터 열시 반이요 **GPT-2 :** 저희 10시에 시작해서 10시 반까지 해요 **T5 :** 저희 열시부터 10시반까지 하고 있습니다
(8)		**참조 문장 :** 이걸로 하나 주세요 **GPT-2 :** 고거 하나 주세요 **T5 :** 이거 하나만 주세요

(9)		**참조 문장 :** 개당 가격이 7000원이예요 **GPT-2 :** 개당 가격이 7천 원입니다 **T5 :** 두 판에 가격은 7천원입니다
(10)		**참조 문장 :** 코끼리가 뭐하고 있어? **GPT-2 :** 코끼리가 뭐하고 있어? **T5 :** 코끼리가 뭐하고 있어? 코끼리도 뭔가를 준비했는데요?

- **기존 학습 데이터(TD1)와 추가 학습 데이터(TD2)의 비율별 GPT-2 모델 기반 미세 조정 성능 비교 실험**

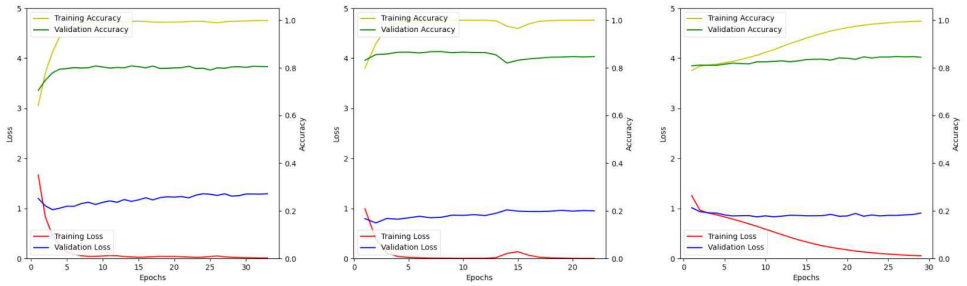
[그림 11]은 GPT-2 모델을 기반으로 기존 학습 데이터셋(TD1)과 추가 학습 데이터셋(TD2)의 비율 및 추가 데이터셋을 미세 조정하는 방식에 따라 성능을 비교하는 실험에 대한 학습 그래프이다. 각 데이터셋 비율별로 (1) R1 데이터셋에 대한 미세 조정, (2) R2 데이터셋에 대한 미세 조정, (3) R3 데이터셋에 대한 미세 조정, (4) R4 데이터셋에 대한 미세 조정, (5) R5 데이터셋에 대한 미세 조정으로 구분하였고, 각 데이터셋에 대하여 초기 모델 및 미세 조정 방식에 따라 (a) 초기 모델, (b) 초기 모델을 추가 학습 데이터셋(TD2)만으로 전체 미세 조정한 모델, (c) 초기 모델을 추가 학습 데이터셋(TD2)만으로 QLoRA 미세 조정한 모델의 학습 그래프를 순서대로 나열하였다. 정확도 그래프의 y축은 0.0부터 1.0까지로 고정하였고, 손실 그래프의 y축은 0부터 5까지로 동일하게 고정하였다.

모든 데이터셋 비율에서 추가 데이터(TD2)만으로 전체 미세 조정한 모델([그림 11]의 (1)-(b), (2)-(b), (3)-(b), (4)-(b), (5)-(b))의 학습 그래프는 급격한 경사를, QLoRA 미세 조정한 모델([그림 11]의 (1)-(c), (2)-(c), (3)-(c), (4)-(c), (5)-(c))의 학습 그래프는 비교적 완만한 경사를 보였다. 에폭 수를 나타내는 그래프의 x축을 보았을 때, 대부분 추가 학습 데이터(TD2)만으로 QLoRA 미세 조정을 수행한 모델(M3)이 추가 학습 데이터(TD2)만으로 전체 미세 조정한 모델(M2)보다 많은 에폭 수로 학습된 것을 볼 수 있다.



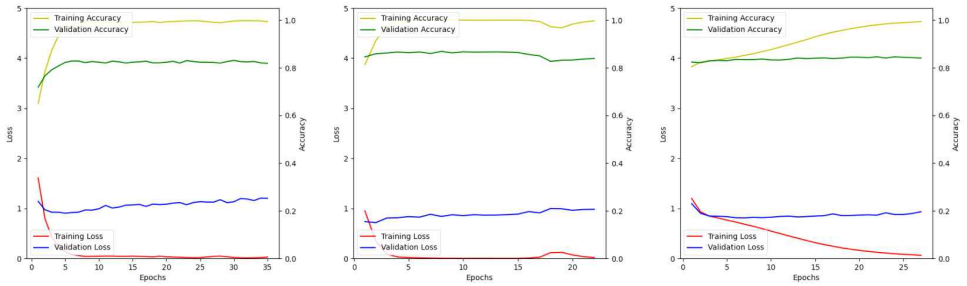
(a) 초기 모델(M0-GPT-2) **(b)** TD2에 대한 추가 전체 미세 조정 모델 (M2-GPT-2) **(c)** TD2에 대한 추가 QLoRA 미세 조정 모델 (M3-GPT-2)

(1) R1 데이터셋에 대한 GPT-2 기반 미세 조정 학습 그래프



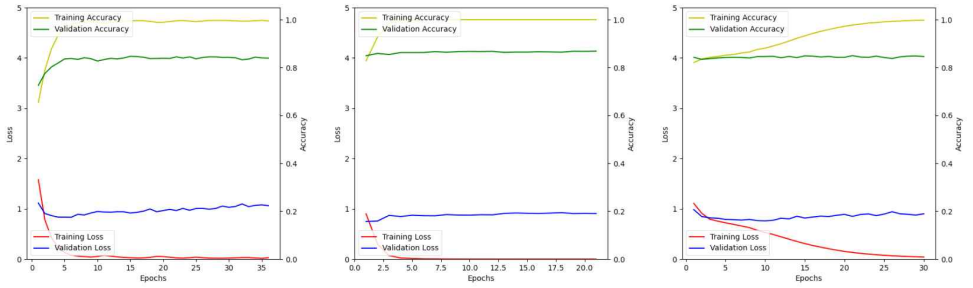
(a) 초기 모델(M0-GPT-2) **(b)** TD2에 대한 추가 전체 미세 조정 모델 (M2-GPT-2) **(c)** TD2에 대한 추가 QLoRA 미세 조정 모델 (M3-GPT-2)

(2) R2 데이터셋에 대한 GPT-2 기반 미세 조정 학습 그래프



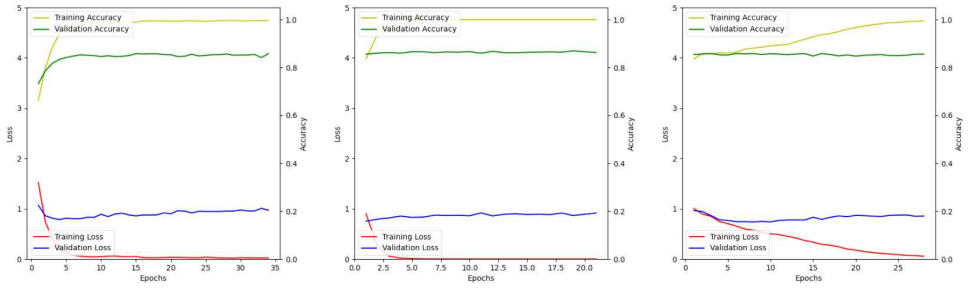
(a) 초기 모델(M0-GPT-2) **(b)** TD2에 대한 추가 전체 미세 조정 모델 (M2-GPT-2) **(c)** TD2에 대한 추가 QLoRA 미세 조정 모델 (M3-GPT-2)

(3) R3 데이터셋에 대한 GPT-2 기반 미세 조정 학습 그래프



(a) 초기 모델(M0-GPT-2) **(b)** TD2에 대한 추가 전체 미세 조정 모델 (M2-GPT-2) **(c)** TD2에 대한 추가 QLoRA 미세 조정 모델 (M3-GPT-2)

(4) R4 데이터셋에 대한 GPT-2 기반 미세 조정 학습 그래프



(a) 초기 모델(M0-GPT-2) **(b)** TD2에 대한 추가 전체 미세 조정 모델 (M2-GPT-2) **(c)** TD2에 대한 추가 QLoRA 미세 조정 모델 (M3-GPT-2)

(5) R5 데이터셋에 대한 GPT-2 기반 미세 조정 학습 그래프

[그림 11] 기존 데이터셋(TD1)과 추가 데이터셋(TD2)의 비율별 GPT-2 기반 미세 조정 학습 그래프

데이터셋 비율 및 미세 조정 방식에 따른 학습 및 검증 손실, 정확도는 [표 12]와 같다. GPT-2 기반 모델의 경우, 대체로 학습 손실은 전체 미세 조정한 모델(Full)이 QLoRA 미세 조정한 모델보다 낮았으며, 검증 손실은 QLoRA 미세 조정한 모델이 전체(Full) 미세 조정한 모델보다 낮았다. 학습 정확도의 경우 데이터셋 비율에 따라 상이한 결과가 나타났으나, 검증 정확도의 경우 추가 데이터셋의 비율이 30%(R3), 40%(R2), 50%(R1)일 때의 각 검증 정확도가 0.8497, 0.8701, 0.8756으로 QLoRA 미세 조정한 모델이 전체(Full) 미세 조정한 모델보다 높은 검증 정확도를 보였다. 또한, 전체(Full) 미세 조정을 수행한 모델의 경우 추가 데이터셋

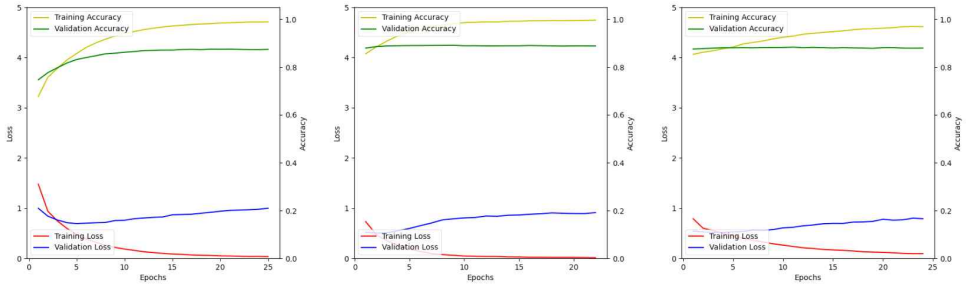
(TD2)의 비율이 높아짐에 따라 정확도가 감소하는 양상을 보이는 반면, QLoRA 미세 조정을 수행한 모델의 경우 추가 데이터셋(TD2)의 비율이 높아짐에 따라 주로 정확도가 향상되었음을 확인할 수 있다.

[표 12] 학습 데이터 비율별 GPT-2 기반 미세 조정 실험 결과

데이터셋	미세 조정	Train		Validation	
		Loss	Accuracy	Loss	Accuracy
R1	Full	**0.0085**	0.9984	1.0292	0.8369
	QLoRA	0.0246	**0.9993**	**0.8460**	**0.8756**
R2	Full	**0.0037**	**0.9995**	0.9436	0.8500
	QLoRA	0.0524	0.9962	**0.8438**	**0.8701**
R3	Full	**0.0455**	0.9894	0.9917	0.8357
	QLoRA	0.0465	**0.9967**	**0.8686**	**0.8497**
R4	Full	**0.0007**	**0.9999**	0.9101	**0.8642**
	QLoRA	0.0454	0.9966	**0.8879**	0.8447
R5	Full	**0.0007**	**0.9999**	0.9400	**0.8648**
	QLoRA	0.0869	0.9881	**0.8701**	0.8495

- **기존 학습 데이터(TD1)와 추가 학습 데이터(TD2)의 비율별 T5 모델 기반 미세 조정 성능 비교 실험**

[그림 12]는 T5 모델을 기반으로 기존 학습 데이터셋(TD1)과 추가 학습 데이터셋(TD2)의 비율 및 미세 조정 방식에 따라 성능을 비교하는 실험에 대한 학습 그래프이다. 초기 모델 및 미세 조정 방식에 따라 각 데이터셋별 (a) 초기 모델, (b) 초기 모델을 추가 학습 데이터셋(TD2)만으로 전체 미세 조정한 모델, (c) 초기 모델을 추가 학습 데이터셋(TD2)만으로 QLoRA 미세 조정한 모델의 학습 그래프를 나열하였다. T5 모델 또한 [그림 12] 중 추가 학습 데이터셋(TD2)에 대하여 QLoRA 미세 조정한 모델의 학습 그래프인 (1)-(c), (2)-(c), (3)-(c), (4)-(c), (5)-(c) 그래프가 추가 학습 데이터셋(TD2)에 대하여 전체 미세 조정한 모델의 학습 그래프보다 학습 시작 단계에서 다소 완만한 경사를 보였고, 학습으로 진행된 에폭 수는 유사함을 알 수 있다.

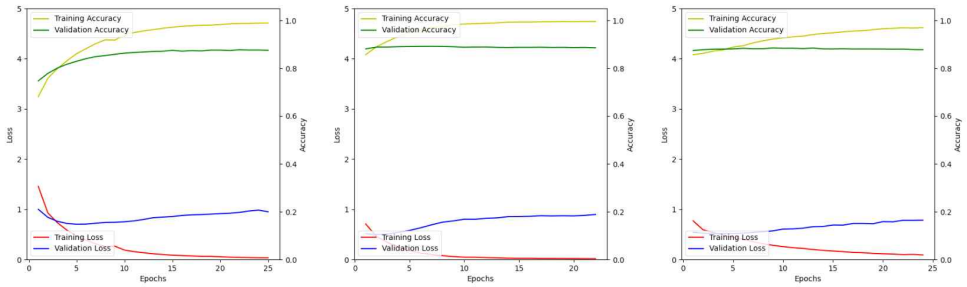


(a) 초기 모델(M0-T5)

(b) TD2에 대한 추가 전체 미세 조정 모델 (M2-T5)

(c) TD2에 대한 추가 QLoRA 미세 조정 모델 (M3-T5)

(1) R1 데이터셋에 대한 T5 기반 미세 조정 학습 그래프

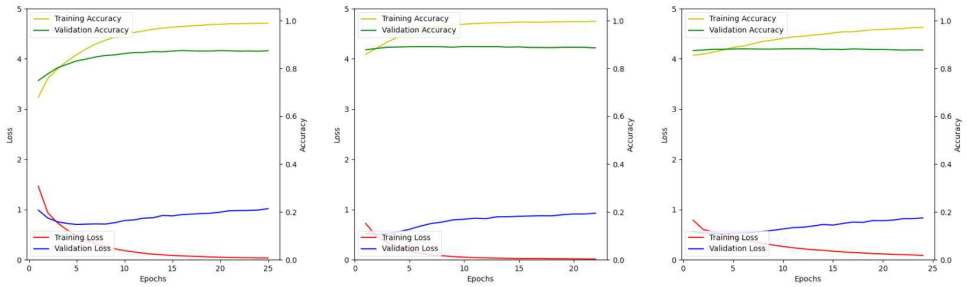


(a) 초기 모델(M0-T5)

(b) TD2에 대한 추가 전체 미세 조정 모델 (M2-T5)

(c) TD2에 대한 추가 QLoRA 미세 조정 모델 (M3-T5)

(2) R2 데이터셋에 대한 T5 기반 미세 조정 학습 그래프

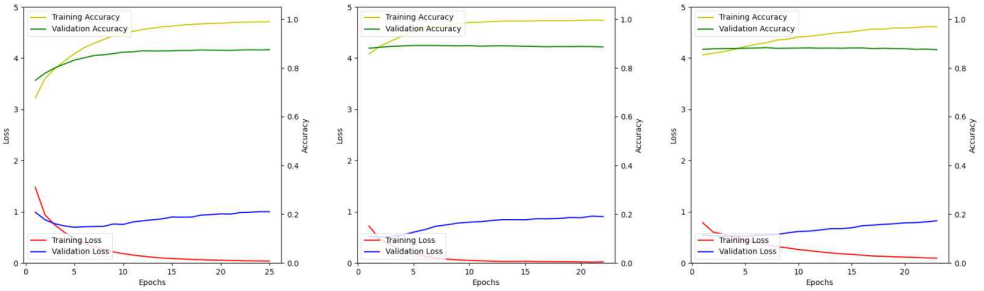


(a) 초기 모델(M0-T5)

(b) TD2에 대한 추가 전체 미세 조정 모델 (M2-T5)

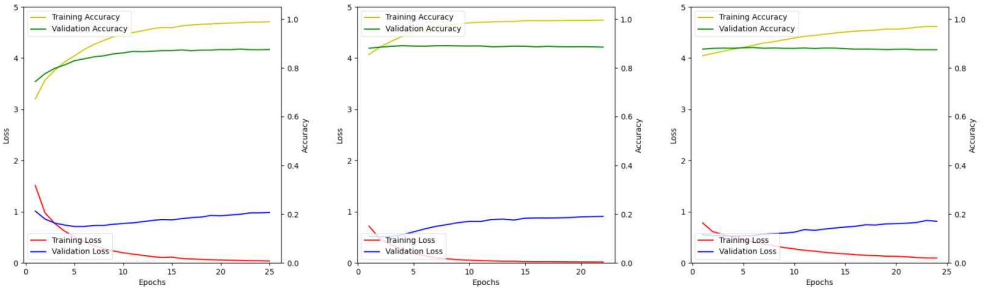
(c) TD2에 대한 추가 QLoRA 미세 조정 모델 (M3-T5)

(3) R3 데이터셋에 대한 T5 기반 미세 조정 학습 그래프



(a) 초기 모델(M0-T5) **(b)** TD2에 대한 추가 전체 미세 조정 모델 (M2-T5) **(c)** TD2에 대한 추가 QLoRA 미세 조정 모델 (M3-T5)

(4) R4 데이터셋에 대한 T5 기반 미세 조정 학습 그래프



(a) 초기 모델(M0-T5) **(b)** TD2에 대한 추가 전체 미세 조정 모델 (M2-T5) **(c)** TD2에 대한 추가 QLoRA 미세 조정 모델 (M3-T5)

(5) R5 데이터셋에 대한 T5 기반 미세 조정 학습 그래프

[그림 12] 기존 데이터셋(TD1)과 추가 데이터셋(TD2)의 비율별 T5 기반 미세 조정 학습 그래프

데이터셋 비율 및 미세 조정 방식에 따른 학습 및 검증 손실, 정확도는 [표 13]과 같다. T5 기반 모델의 경우 모든 데이터셋 비율에 대하여 QLoRA 미세 조정한 모델보다 전체 미세 조정한 모델(Full)의 학습 손실이 낮고 학습 정확도는 높았으며, 검증 손실과 검증 정확도 또한 높았다.

[표 13] 학습 데이터 비율별 T5 기반 미세 조정 실험 결과

데이터셋	미세 조정	Train		Validation	
		Loss	Accuracy	Loss	Accuracy
R1	Full	**0.0146**	**0.9966**	0.9060	**0.8846**
	QLoRA	0.0922	0.9712	**0.8026**	0.8759
R2	Full	**0.0210**	**0.9940**	0.8872	**0.8856**
	QLoRA	0.0894	0.9713	**0.8184**	0.8752
R3	Full	**0.0157**	**0.9960**	0.8980	**0.8867**
	QLoRA	0.0916	0.9708	**0.7970**	0.8768
R4	Full	**0.0183**	**0.9950**	0.8858	**0.8875**
	QLoRA	0.0918	0.9702	**0.7972**	0.8783
R5	Full	**0.0188**	**0.9956**	0.9047	**0.8883**
	QLoRA	0.0807	0.9735	**0.8255**	0.8795

3. 실험 평가

1) BLEU (Bilingual Evaluation Understudy)

BLEU(Bilingual Evaluation Understudy)[30] 성능 지표는 기계 번역 성능을 자동으로 평가하기 위한 지표로, 예측된 문장이 참조 문장과 얼마나 유사한지 측정한다[39]. BLEU는 예측된 문장이 참조 문장과 일치하는 정도를 n-그램 단위의 정확도(Precision) 기반으로 평가하며, 짧은 문장에 패널티(Brevity Penalty, BP)를 부여하여 길이에 따른 편향을 방지한다. BLEU 점수 산출 수식은 [수식 1]과 같다.

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c \leq r \end{cases}$$

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log p_n)$$

[수식 1] BLEU 점수 산출 수식

BLEU 점수는 0에서 1 사이의 값을 가진다. 점수가 높을수록 예측 문장이 참조 문장과 유사함을 의미하며, BLEU 점수 구간에 따라 예측 모델의 성능을 평가할 수 있다[표 14]. BLEU 점수가 0.1 미만인 경우, 예측된 문장의 의미가 거의 없음을 나타내고, 0.1대 점수는 문장의 핵심을 파악하기 어려운 정도로 예측했음을 나타낸다. 0.2대 점수는 모델이 예측한 문장의 핵심을 파악할 수 있으나 문법적인 오류가 존재함을 의미하며, 0.3대의 점수는 이해할 수 있는 문장을 예측함을 의미한다. 0.4대의 점수는 고품질의 문장을 예측함을, 0.5대의 점수는 매우 우수한 품질의 자연

스러운 문장을, 0.6 이상의 점수는 모델이 대체로 사람보다 우수한 품질의 문장을 예측했음을 의미한다. BLEU 점수를 계산하기 위하여 하나의 문장만을 활용하는 것이 아니라 다중 참조 문장을 활용할 수 있어 다양한 표현을 인정하는 유연성을 갖추고 있다. 그러나 BLEU는 n-그램 단위의 표면적 일치만을 중시하여 의미적 유사성을 반영하지 못하며 문장 구조나 의미가 다른 경우에도 점수가 높게 나올 수 있다는 한계가 있다. 따라서 BLEU 성능 지표와 다른 평가 지표를 함께 고려하여 모델의 성능을 평가하는 것이 적절하다.

[표 14] BLEU 점수 구간별 모델 성능의 해석[53]

BLEU 점수	해석
0.1 미만	거의 의미 없음
0.1 ~ 0.2	핵심을 파악하기 어려움
0.2 ~ 0.3	요점은 명확하지만 많은 문법적 오류가 있음
0.3 ~ 0.4	이해할 수 있는 양호한 번역
0.4 ~ 0.5	고품질 번역
0.5 ~ 0.6	매우 우수한 품질의 적절하고 유창한 번역
0.6 초과	대체적으로 사람보다 우수한 품질

2) ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)[27]는 기계 번역, 요약, 질문 응답 시스템 등 다양한 자연어 처리 작업에서 모델의 성능을 평가하는 데 사용되는 지표로, 주로 요약 모델의 성능을 평가하는 데 활용된다[4,39]. ROUGE-N은 ROUGE의 하위 성능 지표 중 하나로, n-그램 기반의 평가를 제공한다. ROUGE-N은 예측된 문장과 참조 문장 간의 n-그램 기반 일치도를 계산하여 유사성을 평가한다. ROUGE-N을 활용한 모델 성능 평가는 일반적으로 1-그램, 2-그램 단위로 수행하는 것으로 알려져 있다. ROUGE-N 점수의 산출 수식은 [수식 2]와 같다.

$$ROUGE-N = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count(gram_n)}$$

[수식 2] ROUGE-N 점수 산출 수식

ROUGE-N 성능 지표[44]는 모델의 예측 문장과 참조 문장 간의 n-그램 일치율을 기반으로 한 재현율과 정확도, 그리고 조화 평균인 F1 점수로 구성된다. ROUGE-N 성능 지표를 통하여 모델이 참조 문장의 정보를 어느 정도 예측 문장에 반영하였는지를 평가할 수 있다. 성능 지표가 제공하는 각 항목에 대해 정의는 다음과 같다.

- **재현율 (Recall)** : 참조 문장에 포함된 n-그램 중, 모델이 예측한 문장에 포함된 n-그램의 비율이다. 즉, 모델이 참조 문장에서 어느 정도의 n-그램을 예측 문장에 반영하였는지 평가한다.

- **정확도 (Precision)** : 예측한 문장의 n-그램 중, 참조 문장에 존재하는 n-그램의 비율이다. 이는 모델이 예측한 문장이 참조 문장과 얼마나 일치하는지를 평가한다.

- **F1 점수 (F1 score)** : 재현율과 정확도의 균형을 맞춘 평가 지표로, 두 값의 조화 평균을 사용하여 모델 성능을 종합적으로 평가한다.

ROUGE-N은 n-그램 기반의 표면적 일치도에 초점을 맞추므로, 의미적 유사성을 고려하지 못한다는 점에서 BLEU 성능 지표와 유사한 한계가 있다. 즉, 문장 구조가 달라지거나, 동의어가 사용되었을 때 낮은 점수를 산출할 수 있다. 예를 들어, '나는 비행기에 방금 탑승했다.'와 '나는 비행기에 지금 탔다.'라는 문장은 의미적으로 유사하나 n-그램의 차이로 인하여 ROUGE-N 성능 지표를 활용한 평가에서 낮은 점수를 받을 수 있다. 또한, n-그램의 수가 적을수록 일치율이 낮아지므로 문장의 길이에 따른 변동성을 고려하지 않는다면 정확한 평가가 어려울 수 있다는 단점이 있다.

3) 코사인 유사도

코사인 유사도(Cosine Similarity)[44]는 두 벡터 간의 유사도를 측정하는 방법으로, 텍스트 데이터나 고차원 데이터를 비교할 때 사용되는 성능 지표이다. 코사인 유사도는 두 벡터가 이루는 각도를 기준으로 유사도를 측정하며 주로 정보 검색, 추천 시스템, 문서 유사도 분석 등에서 두 객체 간의 유사도를 평가하기 위하여 사용된다. 두 벡터가 동일한 방향을 가질수록 유사도가 높고, 다른 방향일수록 유사도가 낮다고 판단된다. 코사인 유사도를 산출하는 수식은 [수식 3]과 같다.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

[수식 3] 코사인 유사도(Cosine Similarity) 산출 수식

코사인 유사도는 크기보다 방향에 중점을 둔 성능 지표이다. 유사도를 측정할 문장의 길이와 관계없이 문장의 핵심이나 주제의 유사도를 평가하는 데 유용하다. 또한, 코사인 유사도는 정규화된 값을 사용하므로 벡터의 길이나 크기에 또한 영향을 받지 않는다는 장점이 있다. 그러나 코사인 유사도는 단어 순서를 고려하지 않는다는 한계가 있다. 코사인 유사도는 벡터의 방향성만을 기준으로 유사도를 평가하므로 문장에서 단어들의 순서가 변경되더라도 높은 유사도를 부여할 수 있다. 예를 들어, '친구가 방금 학교에 도착했다.'와 '방금 학교에 도착한 친구'는 단어 순서에 차이가 있으나 의미상으로 유사하므로 코사인 유사도가 높게 평가될 수 있다.

4) 평가 결과

각 실험에 대한 BLEU, ROUGE-N, 코사인 유사도 평가 지표에 따른 결과는 다음과 같다.

- **대규모 언어 모델(LLM)을 활용한 AAC 상징 시퀀스의 한국어 문장 변환 실험**

대규모 언어 모델을 활용한 AAC 상징 시퀀스의 한국어 문장 변환 실험 결과, GPT-2와 T5 모델을 기반으로 한 AAC 상징 시퀀스의 한국어 문장 변환 모델이 모두 준수한 성능을 보였으며, 특히 GPT-2 모델이 BLEU 0.6519, ROUGE-1 0.7930, ROUGE-2 0.6976, 코사인 유사도 0.7403으로 모든 성능 지표에서 T5보다 우수한 성능을 보임을 알 수 있다[표 15].

[표 15] 대규모 언어 모델(LLM) 기반 AAC 상징 시퀀스의 한국어 문장 변환 실험 결과

성능 지표	GPT-2	T5
BLEU	**0.6519**	0.4657
ROUGE-1	**0.7930**	0.7502
ROUGE-2	**0.6976**	0.5826
코사인 유사도	**0.7403**	0.6082

- **추가 학습 데이터(TD2)에 대한 미세 조정 방식별 성능 및 연산 자원 사용량(GPU) 비교 실험**

추가 학습 데이터(TD2)에 대한 미세 조정 방식별 성능 및 연산 자원 사용량 비교 실험의 결과는 [표 16], [표 17]과 같다. GPT-2의 경우[표 16], 모든 모델이 모든 성능 지표에서 우수한 성능을 보였으며, 특히 초기 학습 데이터(TD1)와 추가 학습 데이터(TD2)를 합친 데이터셋으로 LLM의 사전 학습 모델을 전체 미세 조정한 모델(M1)이 BLEU 0.6519, ROUGE-1 0.7930, ROUGE-2 0.6976, 코사인 유사도 0.7403으로 가장 높은 성능을 보였다. 그러나 M0 모델을 추가 학습 데이터(TD2)만으로 전체 미세 조정한 모델(M2)의 학습 소요 시간이 9분 이내로 80% 이상 적은 시간이 소요되었고, QLoRA 미세 조정을 한 모델(M3)의 학습 소용 시간 또한 16분대로 약 75% 적은 시간이 소요되었다.

QLoRA 미세 조정을 수행한 모델이 전체 미세 조정을 수행한 모델에 비하여 긴 시간이 소요된 데에는 QLoRA 미세 조정 과정 중 양자화 과정을 거치면서 보다 많은 시간이 소요된 것으로 분석하였다. 또한, QLoRA 미세 조정한 모델의 경우, 준수한 성능을 보임과 동시에 GPU 사용량이 3.71GB로 다른 모델에 비하여 현저히 적은 양의 GPU 자원을 학습에 활용하였음을 알 수 있다.

[표 16] 미세 조정 방식별 GPT-2 기반 AAC 상징 시퀀스의 한국어 문장 변환 실험 결과

모델	M0	M1	M2	M3
데이터셋	초기 훈련 데이터셋 (TD1)	전체 훈련 데이터셋 (TD1 + TD2)	추가 훈련 데이터셋 (TD2)	
미세 조정	전체 미세 조정			QLoRA
BLEU	0.5512	**0.6519**	0.6351	0.5690
ROUGE-1	0.7293	**0.7930**	0.7902	0.7503
ROUGE-2	0.6005	**0.6976**	0.6832	0.6259
코사인 유사도	0.6673	**0.7403**	0.7367	0.6870
GPU 사용량	6.87GB			**3.71GB**
소요 시간	52분 11초	1시간 1분 40초	**8분 58초**	16분 27초

[표 17]은 T5 모델 기반의 미세 조정 방식별 성능 및 연산 자원 사용량을 실험한 결과이다. T5 모델 기반의 모든 미세 조정 모델이 네 가지 성능 지표에서 준수한 성능을 보였으며, 추가 데이터(TD2)만으로 전체 미세 조정한 모델(M2)이 BLEU 0.4951, ROUGE-1 0.7642, ROUGE-2 0.6123, 코사인 유사도 0.6395로 가장 좋은 성능을 보였다. 그러나 추가 데이터(TD2)만을 활용하여 QLoRA 미세 조정한 모델(M3)이 학습에 2.54GB만을 사용하며 70% 적은 양의 GPU 자원을 사용하면서도 준수한 성능을 보였다는 점에서 QLoRA 미세 조정을 수행한 모델이 제한된 연산 자원 환경에서도 충분히 학습할 수 있도록 함을 알 수 있다.

[표 17] 미세 조정 방식별 T5 기반 AAC 상징 시퀀스의 한국어 문장 변환 실험 결과

모델	M0	M1	M2	M3
데이터셋	초기 훈련 데이터셋 (TD1)	전체 훈련 데이터셋 (TD1 + TD2)	추가 훈련 데이터셋 (TD2)	
미세 조정	전체 미세 조정			QLoRA
BLEU	0.4925	0.4657	**0.4951**	0.4521
ROUGE-1	0.7484	0.7502	**0.7642**	0.7427
ROUGE-2	0.5926	0.5826	**0.6123**	0.5820
코사인 유사도	0.6324	0.6082	**0.6395**	0.6155
GPU 사용량	8.43GB			**2.54GB**
소요 시간	36분 35초	57분 46초	**8분 49초**	24분 18초

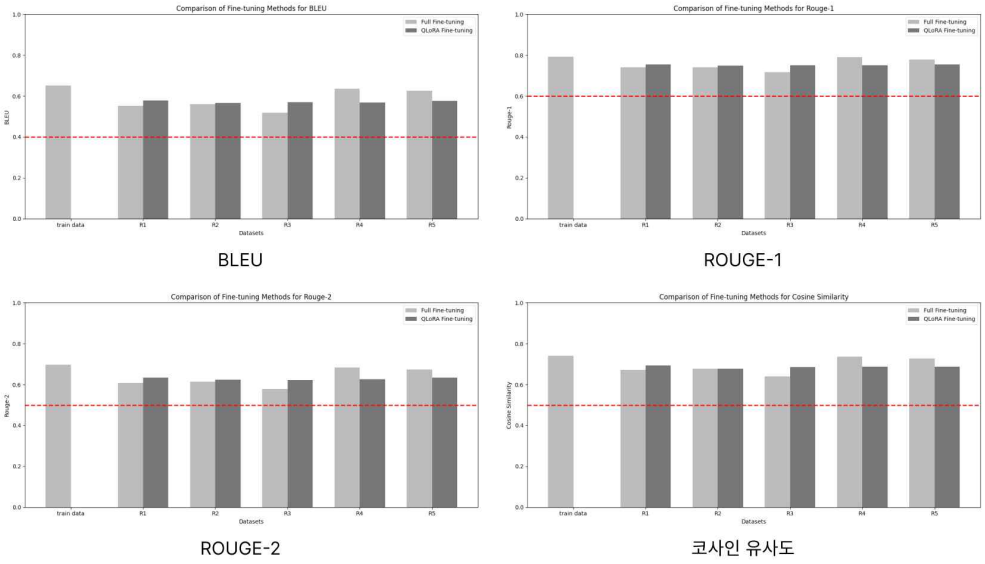
- **기존 학습 데이터(TD1)와 추가 학습 데이터(TD2)의 비율별 미세 조정 성능 비교 실험**

GPT-2 모델을 활용한 학습 데이터셋의 비율별 미세 조정 방식에 따른 성능 비교 실험의 결과는 [표 18]과 [그림 13]과 같다. [그림 13]의 각 그래프에 표시된 붉은 점선은 모델이 준수하다고 평가되는 성능 지표별 일반적인 기준이다. 모든 데이터셋에 대한 학습 모델이 각 성능 지표에서 매우 준수한 성능을 보였으며 특히 추가 데이터셋의 비율이 30%(R3), 40%(R2), 50%(R1)일 때 QLoRA 미세 조정한 모델이 전체(Full) 미세 조정한 모델보다 높은 성능을 보였다. 이를 통하여 추가 데이터셋의 비율이 높아질수록 QLoRA 미세 조정한 모델의 성능이 전체 미세 조정한 모델

의 성능보다 높을 수 있음을 예측할 수 있다. 따라서 추가 학습 데이터셋을 모델에 반영하기 위한 학습을 수행하는 과정에서 추가 데이터의 비율이 높을 때, QLoRA 미세 조정을 적용한 모델이 적은 양의 연산 자원을 활용하면서도 보다 우수한 성능을 보임을 알 수 있다.

[표 18] 학습 데이터 비율별 GPT-2 기반 미세 조정 성능 비교

데이터셋	미세 조정	BLEU	ROUGE-1	ROUGE-2	코사인 유사도
R1	Full	0.5529	0.7403	0.6083	0.6722
	QLoRA	**0.5774**	**0.7554**	**0.6330**	**0.6932**
R2	Full	0.5596	0.7401	0.6135	**0.6783**
	QLoRA	**0.5670**	**0.7483**	**0.6238**	0.6778
R3	Full	0.5186	0.7172	0.5786	0.6408
	QLoRA	**0.5702**	**0.7513**	**0.6224**	**0.6856**
R4	Full	**0.6351**	**0.7902**	**0.6832**	**0.7367**
	QLoRA	0.5690	0.7503	0.6259	0.6870
R5	Full	**0.6252**	**0.7794**	**0.6743**	**0.7272**
	QLoRA	0.5756	0.7550	0.6330	0.6870



BLEU ROUGE-1

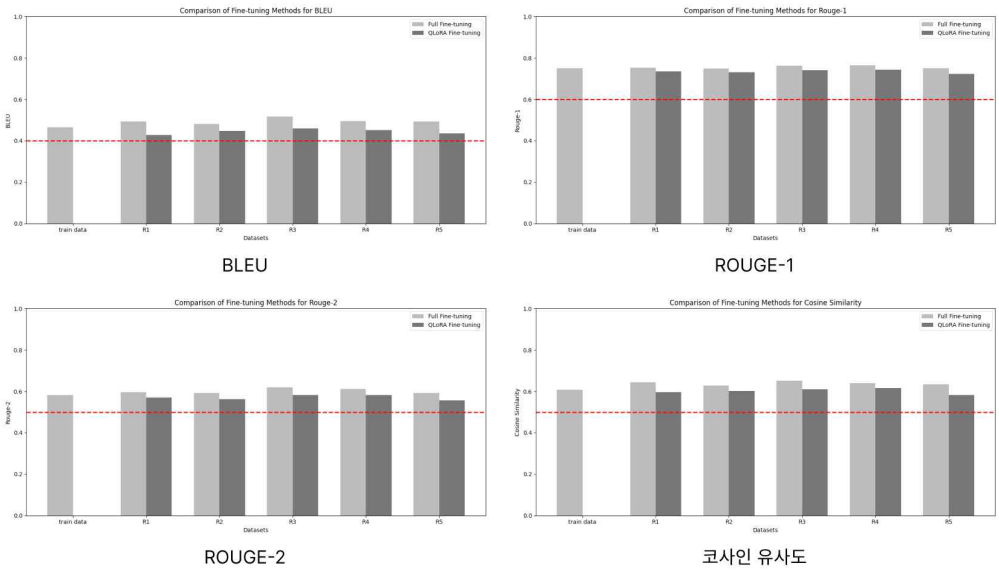
ROUGE-2 코사인 유사도

[그림 13] 학습 데이터 비율별 GPT-2 기반 미세 조정 성능 그래프

T5 모델을 활용한 실험의 결과는 [표 19]와 [그림 14]와 같다. [그림 14]에서 붉은 점선은 성능 지표별 준수하다고 평가되는 일반적인 기준이며, QLoRA 미세 조정 모델이 전체(Full) 미세 조정한 모델의 성능에 미치지 못했으나, 모든 데이터셋 비율에서 준수한 성능을 갖추었다는 것을 알 수 있다. [표 18]과 [표 19]의 학습 데이터 비율별 성능 지표 결과를 비교하였을 때, T5 모델을 기반으로 한 미세 조정 모델의 성능이 GPT-2 기반 모델에 비하여 낮은 성능을 보였다. 이러한 T5 모델에 대하여 성능을 높일 수 있는 하이퍼파라미터 실험이 필요하다.

[표 19] 학습 데이터 비율별 T5 기반 미세 조정 성능 비교

데이터셋	미세 조정	BLEU	ROUGE-1	ROUGE-2	코사인 유사도
R1	**Full**	**0.4940**	**0.7536**	**0.5957**	**0.6442**
	QLoRA	0.4285	0.7345	0.5705	0.5962
R2	**Full**	**0.4812**	**0.7497**	**0.5914**	**0.6283**
	QLoRA	0.4472	0.7318	0.5623	0.6027
R3	**Full**	**0.5162**	**0.7622**	**0.6205**	**0.6510**
	QLoRA	0.4585	0.7407	0.5832	0.6100
R4	**Full**	**0.4951**	**0.7642**	**0.6123**	**0.6395**
	QLoRA	0.4521	0.7427	0.5820	0.6155
R5	**Full**	**0.4931**	**0.7502**	**0.5930**	**0.6333**
	QLoRA	0.4351	0.7238	0.5567	0.5829



[그림 14] 학습 데이터 비율별 T5 기반 미세 조정 성능 그래프

VI. 결론 및 향후 연구

보완대체의사소통(Augmentative and Alternative Communication, AAC)은 언어 표현 및 이해에 어려움을 겪는 사람들을 위해 다양한 도구와 방식을 통해 의사소통을 지원한다. AAC 도구들은 AAC 사용자가 일상에서 자율적이고 적극적인 사회적 상호작용과 의사소통을 수행하도록 돕는다. 특히, AAC 그림 상징은 직관적인 그림 이미지와 텍스트로 구성되어 있어 AAC 사용자가 이를 통해 다양한 표현을 할 수 있다. 그러나 비장애인은 AAC 상징에 대한 이해도가 낮기 때문에 AAC 사용자가 일상에서 AAC 상징을 활용하여 소통하는 것은 제한적이다. 이로 인하여 AAC 상징 시퀀스를 자연스러운 한국어 문장으로 변환하는 기능의 필요성이 높아지고 있다. 또한, AAC 사용자가 모델을 활용하면서 축적되는 데이터가 사용자의 요구와 실시간 피드백을 담고 있기 때문에 이를 모델에 주기적으로 반영하는 것은 사용자 경험 개선에 매우 중요한 역할을 한다는 점에서 정기적으로 모델을 업데이트하는 것은 필수적이다.

본 연구는 제한된 연산 자원 환경에서 대규모 언어 모델(Large Language Model, LLM)을 활용하여 AAC 상징 시퀀스를 자연스러운 한국어 문장으로 변환하고자 하였으며, AAC 사용자가 모델을 활용하면서 축적되는 추가 데이터를 주기적으로 모델에 반영하기 위한 미세 조정을 수행하고자 하였다. 따라서 본 연구에서 수행한 실험 내용은 다음과 같으며, 모든 실험에서 성능 비교를 위하여 n-그램 기반 성능 평가(BLEU[30], ROUGE-N[27])와 의미적 유사성 기반 성능 평가(코사인 유사도[44])를 함께 수행하였다.

첫째, 대규모 언어 모델인 GPT-2[32]와 T5[33] 모델의 사전 학습 모델을 활용하여 AAC 상징 시퀀스를 한국어 문장으로 변환하는 모델을 설계하고 성능을 평가하였다. 두 모델 모두 준수한 성능을 보였으며, GPT-2 기반 모델이 T5 기반 모델에 비해 모든 성능 지표에서 높은 성능을 보였다.

둘째, 추가 학습 데이터셋에 대한 미세 조정 방식별 성능과 연산 자원 사용량을 비교하였다. GPT-2와 T5 모델에 대하여 1) LLM의 사전 학습 모델을 기존 학습 데이터와 추가 학습 데이터를 합친 데이터셋으로 전체 미세 조정한 모델, 2) 초기 모델을 추가학습 데이터셋만으로 전체 미세 조정한 모델, 3) 초기 모델을 추가학습 데이터셋만으로 QLoRA 미세 조정[29]한 모델로 분류하여 실험하였다. GPT-2 기반 모델은 기존 학습 데이터셋과 추가 학습 데이터셋을 합친 데이터셋으로 전체 미세 조정한 모델의 성능이 가장 높았으며, QLoRA 미세 조정을 수행한 모델은 50% 적은 GPU 메모리를 사용하면서도 준수한 수준의 성능을 보였다. T5 기반 모델은 초기 모델을 추가 학습 데이터셋만으로 전체 미세 조정한 모델의 성능이 가장 우수하였으며, QLoRA 미세 조정을 수행한 모델은 30% 미만의 GPU 메모리를 사용하면서도 준수한 성능을 보였다.

셋째, 기존 학습 데이터셋과 추가 학습 데이터셋의 비율별 미세 조정 성능을 비교 실험하였다. BLEU, ROUGE-N, 코사인 유사도 성능 지표를 바탕으로 한 실험 결과, 모든 모델, 모든 데이터셋 비율에 대한 미세 조정 모델이 준수한 성능을 보였다. 특히, GPT-2 기반 모델을 활용한 실험에서 추가 데이터셋의 비율이 전체 데이터셋의 30%, 40%, 50%일 때, 전체 미세 조정한 모델보다 좋은 성능을 보임을 확인하였다.

세 가지 실험을 통하여 제한된 물리적 연산 자원 환경에서도 대규모 언어 모델을 활용하여 AAC 상징 시퀀스를 자연스러운 한국어 문장으로 변환하였다. 이는 시퀀스-투-시퀀스 모델을 활용하여 AAC 상징 시퀀스를 한국어 문장으로 변환한 선행 연구[2]를 확장하였고, [2] 연구에서 제안한 모델 중 가장 높은 성능을 보인 BERT[22] 모델의 임베딩 레이어를 적용한 시퀀스-투-시퀀스 모델의 BLEU[30,53] 점수 0.5826과 비교하였을 때, 대규모 언어 모델을 활용하여 AAC 상징 시퀀스를 한국어 문장으로 변환하는 모델의 BLEU 점수를 0.6519로 높이며 보다 좋은 성능의 모델을 제안했다는 점에서 의의가 있다. 또한, 효율적인 미세 조정 방식인 QLoRA 미세 조정 방식을 활용하여 기존 데이터셋뿐만 아니라 새롭게 축적되는 데이터셋을 모델에 반영하였다는 점에서 사용자로부터 데이터를 수집하고 이를 지속적으로 반영하는 과정에서의 효율적인 학습과 준수한 성능을 확인하였다. 성능 지표 결과에 있어 주로 LLM의 사전 학습 모델을 전체 데이터셋을 처음부터 다시 학습하는 미세 조정이 가장 높은 성능을 보였지만, 제한된 연산 자원 환경에서도 대규모 언어 모델을 활용하여 미세 조정을 수행했다는 점에서 추후 적은 연산 자원을 사용하면서도 더욱 큰 규모의 파라미터 수를 가지는 모델을 활용하여 더욱 우수한 성능을 낼 수 있을 것으로 기대한다.

추후 연구로서 보완대체의사소통 상징 시퀀스의 문장 변환 모델의 성능을 향상하기 위한 하이퍼파라미터 실험이 필요하다. 모델의 정확도를 더욱 극대화하기 위하여 새로운 대규모 언어 모델을 활용하거나 프롬프트 엔지니어링(Prompt)을 수행하는 실험과 앙상블(Ansemble) 기법을 적용한 실험 또한 추후 연구의 방향으로 제시한다. 매우 방대한 파라미터 수를 가지는 대규모 언어 모델을 활용하고, 이를 재훈련하지 않고도 도메

인에 적합한 모델로 활용할 수 있도록 프롬프트 엔지니어링을 활용하거나, 여러 모델을 결합하여 예측을 서로 보완할 수 있는 방법을 모색할 필요가 있으며 이를 통하여 더욱 정확하고 유기적인 예측 결과를 얻을 수 있을 것으로 기대한다.

참 고 문 헌

[1] 김수미. “AAC를 활용한 함께 책 읽기 중재가 복합의사소통장애 학 생의 의미 관계 표현과 어휘다양도 변화에 미치는 효과,” 국내석사 학위논문 창원대학교 대학원, 2019.

[2] 안서영. (2023). 딥러닝 기반 보완대체의사소통 상징의 다의성을 반영한 상징 시퀀스의 한국어 문장 변환 [석사학위논문, 성신여자대학교]. https://www-riss-kr.libproxy.sungshin.ac.kr/link?id=T16832829

[3] 천춘경. “보완.대체 의사소통 (AAC) 체계 활용을 위한 지역사회 중심의 기초어휘 및 문장조사,” 국내석사학위논문 단국대학교 대학원, 2000.

[4] 김영석, 이병훈, 강민지, & 한성원. (2022). 시퀀스-투-시퀀스를 활용한 데이터 기반 문장 생성 모델 비교. *Journal of the Korean Institute of Industrial Engineers-Vol*, 48(5), 464-476.

[5] 박은혜, & 김정연. (2006). 손짓기호체계와 그림의사소통판을 이용한 의사소통 중재가 중도뇌성마비학생의 의사소통 능력에 미치는 효과. 지체. 중복. *건강장애연구*, 47, 265-289.

[6] 박은혜. 보완/대체의사소통체계를 위한 기초어휘조사: 뇌성마비 초등 저학년 학생을 중심으로, *특수교육논총*, vol. 13, no. 1, pp.91-115, 1996.

[7] 박종영. (2018). RNN 을 활용한 도시철도 역사 부하 패턴 추정. *전기학회논문지*, 67(11), 1536-1541.

[8] 배기민, 이학진, 김세옥, & 이장형. (2024). 온 디바이스 국방 AI 를 위한 PEFT 효용성 연구. *한국컴퓨터정보학회 학술발표논문집*, 32(1), 51-54.

[9] 서지우, 유세희, 홍기형, 연석정, 채수정, & 이희연. (2024). 위치 상황 기반 보완대체의사소통의 자원 계층 구조와 상황 전환 사용자 인터페이스 개발. *보완대체의사소통연구*, 12(1), 1-22.

[10] 이정은, 박은혜. “보완· 대체의사소통체계 적용을 위한 상황 중심 핵심어휘 개발 연구,” *재활복지*, vol. 4, pp. 96- 122, 200

[11] 이재헌, 황환이, 정태순, 김덕용, 안정빈, 이규찬, & 이승환. (2024). 시계열 데이터를 이용한 딥러닝 기반 용접 공정 모니터링 리뷰. *대한용접접합학회지*, 42(4), 333-344.

[12] 조희, & 홍기형. (2020). GeoAAC, 위치기반 보완대체의사소통 모바일 앱. *보완대체의사소통연구*, 8(1), 87-117.

[13] 최진, & 김종한. (2024). 시계열 데이터 예측을 위한 트랜스포머 경량화 설계. *한국항공우주학회 학술발표회 초록집*, 520-521.

[14] 허은정. (2010). 그림교환 의사소통 체계 (PECS) 프로그램이 자폐아동의 요구하기, 자발적 발화 및 눈맞춤에 미치는 효과. *정서· 행동장애연구*, 26(3), 179-208.

[15] Ali, A. H., Yaseen, M. G., Aljanabi, M., & Abed, S. A. (2023). Transfer learning: A new promising techniques. *Mesopotamian Journal of Big Data*, 2023, 29-30.

[16] Athukoralage, D., & Atapattu, T. Multi-Stage QLoRA with Augmented Structured Dialogue Corpora: Efficient and Improved Conversational Healthcare AI. In *NeurIPS 2024 Third Table Representation Learning Workshop*.

[17] Bahdanau, D. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[18] 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., & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

[19] Cho, K. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[20] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

[21] De Marco, P., Ricciardi, V., Montesano, M., Cassano, E., & Origgi, D. (2024). Transfer learning classification of suspicious lesions on breast ultrasound: is there room to avoid biopsies of benign lesions?. *European Radiology Experimental*, 8(1), 121.

[22] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[23] Graves, A. Long Short-term Memory. *Supervised Sequence Labelling with Recurrent Neural Networks*, 2012, 37-45.

[24] He, K., Girshick, R., & Dollár, P. (2019). Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4918-4927).

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

[26] Lewis, M. (2019). Bart: Denoising sequence-to-sequence pre-trainin

g for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

[27] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).

[28] McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109-165). Academic Press.

[29] Ni, H., Meng, S., Chen, X., Zhao, Z., Chen, A., Li, P., Zhang, S., Yin, Q., Wang, Y., & Chan, Y. (2024, August). Harnessing earnings reports for stock predictions: A qlora-enhanced llm approach. In *2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS)* (pp. 909-915). IEEE.

[30] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

[31] Radford, A. (2018). Improving language understanding by generative pre-training.

[32] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.

[33] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of mach*

ine learning research, 21(140), 1-67.sss

[34] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088), 533-536.

[35] Sanh, V. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[36] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404, 132306.

[37] Sutskever, I. (2014). Sequence to Sequence Learning with Neural Networks. arXiv preprint arXiv:1409.3215.

[38] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

[39] Wang, Y., Jiang, J., Zhang, M., Li, C., Liang, Y., Mei, Q., & Bendersky, M. (2023). Automated evaluation of personalized text generation using large language models. arXiv preprint arXiv:2310.11593.

[40] Xu, L., Xie, H., Qin, S. Z. J., Tao, X., & Wang, F. L. (2023). Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. arXiv preprint arXiv:2312.12148.

[41] Yeon, S. J., Kim, Y. T., & Park, E. H. (2016). Transparency and name agreement of Korean Ewha-AAC symbols: nouns, verbs, and adjectives. *AAC Research & Practice*, 4(1), 45-63.

[42] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, Hui., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43-76.

[43] 그림교환의사소통 체계(PECS). (n.d.). *Pyramid Educational Consultants*. https://pecs-korea.com/. [Accessed: 24-Dec-2024].

[44] 딥 러닝을 이용한 자연어 처리 입문. (2024, December 11). *Wiki Docs*. https://wikidocs.net/book/2155. [Accessed: 24-Dec-2024].

[45] “마이토키,” Mytalkie.co.kr. [Online]. Available: http://www.mytalkie.co.kr/. [Accessed: 10-Dec-2024].

[46] 몸짓상징-손담. (n.d.). *장애자녀 부모 지원 종합시스템*. https://www.nise.go.kr/onmam/front/index.do. [Accessed: 10-Dec-2024].

[47] "인식기술-언어지능," Aihub.or.kr. [Online]. Available: https://aihub.or.kr/keti_data_board/language_intelligence. [Accessed: 10-Dec-2024].

[48] 한국정보화진흥원, Aihub.or.kr [Online]. Available https://aihub.or.kr. [Accessed: 10-Dec-2024].

[49] AdamW. (2023). *PyTorch*. https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html. [Accessed: 24-Dec-2024].

[50] c4. (2022, December 6). *Tensorflow*. https://www.tensorflow.org/datasets/catalog/c4. [Accessed: 24-Dec-2024].

[51] Christopher Olah. (2015). Understanding LSTM Networks. *colah's blog*. https://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed: 24-Dec-2024].

[52] CrossEntropyLoss. (2023). *PyTorch*. https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html. [Accessed: 3-Dec-2024].

[53] Evaluating Models. (2024, August 13). *Google Cloud*. https://cloud.google.com/translate/automl/docs/evaluate. [Accessed: 3-Dec-2024].

[54] OpenAI GPT2. (n.d.). *Hugging Face*. https://huggingface.co/docs/transformers/model_doc/gpt2. [Accessed: 24-Dec-2024].

[55] paust/pko-t5-base. (n.d.). *Hugging Face*. https://huggingface.co/paust/pko-t5-base. [Accessed: 24-Dec-2024].

[56] SKT-AI/KoGPT2. (n.d.). *GitHub*. https://github.com/SKT-AI/KoGPT2. [Accessed: 24-Dec-2024].

ABSTRACT

Transformation of AAC Symbol Sequences into Korean Sentences Based on Large Language Model and QLoRA Fine-Tuning

Seo Jiwoo
Department of Future Convergence Technology Engineering
Graduate School of Sungshin University

In modern society, communication is becoming increasingly important, however, communication for individuals with speech disabilities remains limited. Augmentative and Alternative Communication(AAC) is a supplementary tool or method that supports their communication more effectively. AAC users use AAC symbols to represent their intended messages. AAC symbols consist of intuitive symbolic images and their textual expressions. AAC users can arrange these AAC symbols to convey their intentions or emotions. However, since non-disabled people have a low understanding of AAC symbols, translation models from the AAC symbol sequences into natural Korean sentences are essential for smooth communication. For the continuous improvement of service quality and user experience of the translation models, periodic model updates based on data accumulated from AAC users using the models are required.

This study aims to transform AAC symbol sequences into Korean sentences using large language models(LLMs). The study also aims to use QLoRA fine-tuning to efficiently update the model periodically, even in environments with limited computational resources.

We used pre-trained models of transformer-based LLMs with high performance in the field of natural language processing, to ensure competitive performance with limited data while handling new data flexibly. As LLMs grow rapidly, they require lots of computational resources like GPUs for training. To overcome this limitation, this study uses QLoRA fine-tuning, a Parameter Efficient Fine-Tuning(PEFT) method, to reduce the usage of resources while maintaining competitive performance.

We conducted three experiments to evaluate model performance. First, we evaluated models converting AAC symbol sequences into Korean sentences using GPT-2 and T5 Korean models, and compared their performance. Both models showed competitive performance, GPT-2 shows BLEU 0.6519, ROUGE-1 0.7930, ROUGE-2 0.6976, and cosine similarity 0.7403 and is better than T5. Second, we performed the additional training using full fine-tuning and QLoRA fine-tuning with the additional training data, and compared the performances of the two fine-tuning methods. For GPT-2, the full fine-tuned model using both existing and additional training data achieved the highest performance, while the QLoRA fine-tuned model demonstrated similar performance with 50% less GPU memory usage. For T5, the full fine-tuned model achieved the best performance on additional training data, and the QLoRA fine-tuned model demonstrated competitive performance while utilizing less than 30% of GPU memory. Third, we compared the performance of fine-tuning based on the ratio of additional training data. Results showed that

both fully fine-tuned and QLoRA fine-tuned models using GPT-2 and T5 achieved strong performance across all ratios. In experiments using the GPT-2 model, the QLoRA fine-tuned model performed better than the full fine-tuned model when the additional dataset was 30%, 40%, or 50% of the total training dataset.

In conclusion, this study successfully converted AAC symbol sequences into natural Korean sentences using LLMs, achieving better performance compared to previous studies. Furthermore, we also suggested a way to keep adding new data to the model, even in environments with limited resources.