지시문 및 번역 데이터셋을 활용한 Llama2 Cross-lingual 한국어 확장

장규식°[1], 이현민[1], 나승훈[1], 임준호[2], 김태형[3], 류휘정[3], 장두성[3]
전북대학교[1], 한국전자통신연구원[2], KT[3]
{jks4880, leehm, nash}@jbnu.ac.kr[1]
joonho.lim@etri.re.kr[2]
{taehyeong_2019.kim, hwijung.ryu, dschang}@kt.com[3]

Llama2 Cross-lingual Korean with instruction and translation datasets

Gyu-sik Jang°[1],Seung-Hoon Na[1], Joon-Ho Lim[2], Tae-Hyeong Kim[3], Hwi-Jung Ryu[3], Du-Seong Chang[3]
Jeonbuk National University[1], ETRI[2], KT[3]

요 약

대규모 언어 모델은 높은 연산 능력과 방대한 양의 데이터를 기반으로 탁월한 성능을 보이며 자연어처리 분야의 주목을 받고있다. 이러한 모델들은 다양한 언어와 도메인의 텍스트를 처리하는 능력을 갖추게 되었지만, 전체 학습 데이터 중에서 한국어 데이터의 비중은 여전히 미미하다. 결과적으로 이는 대규모 언어 모델이 영어와 같은 주요 언어들에 비해 한국어에 대한 이해와 처리 능력이 상대적으로 부족함을 의미한다. 본 논문은 이러한 문제점을 중심으로, 대규모 언어 모델의 한국어 처리 능력을 향상시키는 방법을 제안한다. 특히, Cross-lingual transfer learning 기법을 활용하여 모델이 다양한 언어에 대한 지식을 한국어로 전이시켜 성능을 향상시키는 방안을 탐구하였다. 이를 통해 모델은 기존의 다양한 언어에 대한 손실을 최소화 하면서도 한국어에 대한 처리 능력을 상당히 향상시켰다. 실험 결과, 해당 기법을 적용한 모델은 기존 모델 대비 nsmc데이터에서 2배 이상의 성능 향상을 보이며, 특히 복잡한 한국어 구조와 문맥 이해에서 큰 발전을 보였다. 이러한 연구는 대규모 언어 모델을 활용한 한국어 적용 향상에 기여할 것으로 기대 된다.

주제어: LLaMA2, Cross-lingual transfer learning, 한국어 특화 모델

1. 서론

최근의 언어 모델 발전, 특히 대규모 언어 모델인 PaLM2[1], GPT4[2], LLaMA2[3]의 등의 발전과 함께 한국어 지원 모델의 수도 늘어나고 있다. 대규모 언어 모델들은 사람들의 감정을 분류하는 작업을 비롯해서 번역, 문서 작성, 인과 관계 유추 등의 문장에 대한 내용을 이해해야 할 수있는 분야까지 발전하고 있으며 높은 정확도를 기록하고 있다. 그럼에도 불구하고, 영어에서는 작업을 잘 수행하는 반면 한국어로 모델이 작업을 수행할때는 다양한 작업에서 한계를 보이는 문제가 있다. 이는 사전 학습된 모델들의 학습데이터가 영어에 편중되어 있으며, 한국어 학습량이 적기 때문에 발생하는 현상이다.

이러한 문제점을 해결하기 위해서 한국어 데이터셋을 기반으로 대규모 언어 모델을 개발하려는 시도가 있다. 하지만 이러한 방식은 데이터 구축 비용과 계산 비용이 많이 들며 연구에 한계를 직면하게 되며, 연구에 있어 하나의 제약이 된다. 이런 제약을 해결하기 위해 최근에는 기존의 언어 모델의 성능을 다른 언어로 전이하여 적은 데이터를 활용한 학습으로 높은 성능을 구축하는 방식의 연구가 진행되고 있다. 본 논문은 LLaMA2 모델을 이용한 Cross-lingual transfer learning 기법을 도입하여, 영어 중심의 사전 학습된 모델을 한국어 성능 향상을 위해 추가학습을 하는 방식을 제안한다.

LLaMA2[3]는 Meta AI에서 공개한 오픈소스 모델이다. 해당 모델은 2조개의 토큰을 사용하여 학습하였지만, 한국어 데이터셋은 전체 언어에 단 0.06% 밖에 되지 않는다. 이는 한국어에 대한 성능이 다른언어에 비해 상대적으로 낮을 수 밖에 없다.

위에 언급한 단점을 보완하기 위해 Cross-ligual Transfer Learning을 적용하였다. AI-hub에서 제공하는 한국어-영어 병렬 말뭉치 데이터를 사용하여 영어 기반의 Knowledge를 한국어로 전이하여 한국어에 대한 언어모델의 성능을 증가시키고자 하였다. 또한, 본 연구에서는 지시 데이터도 학습하는데 사용하였다. 이는 학습하지 않은 지시에 대한 언어모델의 Zero-Shot 성능 증가를 불러올 수 있다.[4] 본 연구에서는 지시 데이터로 Alpaca 데이터셋[5]과 한국어로 번역한 데이터셋을 사용한다. 이 과정은 모델에 여러가지 이점을 제공한다. 첫째로, 지시 데이터는 모델에 특정 작업에 대한 지시를 명확히 제공한다. 둘째로, 다양한 문맥의 지시를 포함하여 모델의 반응성을 향상시킨다. 마지막으로, Cross-lingual Instruction Tuning을 통해 모델이 지시를 이해하는 능력을 향상 시키며, 적은 양의 한국어 데이터만을 활용하여 높은 성능 향상을 달성할 수 있다.

본 논문의 기여는 다음과 같다.

- Cross-lingual transfer learning 기법이 영어와 유사성이 적은 한국어에 대해서도 적용을 확인하였으며, 그에 대한 성능을 확인 했다.
- 사전학습된 언어 모델을 상대적으로 적은 자원을 사용하여, 모델의 성능을 향상 시킬 수 있음을 보였다.
- 모델의 한국어 이해능력의 향상을 보이는 방법을 제시하였다.

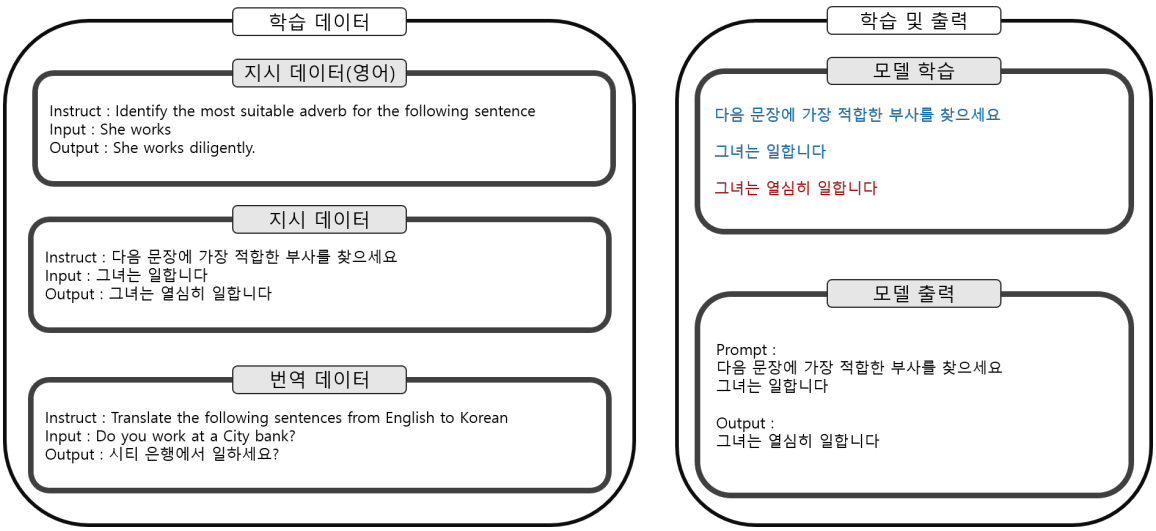

그림 1. 학습 예제

검정 : 데이터 포맷, 파랑 : 학습 하지 않는 데이터, 빨강 : 학습 하는 데이터

2. 관련 연구

2.1 한국어 특화 언어 모델

카카오 및 SKT와 같은 기업은 GPT를 기반으로 한국어 특화 모델인 KoGPT를 개발하여 진행해 왔으며, SKT의 KoGPT의 경우 다양한 공개 데이터셋을 사용한것 뿐만 아니라 이모티콘 등을 활용하여 인식 능력을 향상한 한국어 특화 모델을 공개했으며, 카카오의 경우 GPT3를 벤치마킹 하여, 200B의 한국어 데이터로 6B의 모델을 발표하였다. [6] 네이버는 HyperCLOVA X 라는 한국어 특화 모델을 개발하여, 공개하였고 ChatGPT의 한국어 양보다 6500배가 되는 한국어 데이터 셋을 사용하여 한국어 언어 모델에 대한 성능과 활용도 측면에서 다양성을 향상 시켰으며, 한국어 자연어 처리 분야에서 큰 기여를 하였다. [1] 대기업이 아닌 국내 대학으로는 고려대에서는 Polyglot-ko를 기반으로 한국어 지시 데이터셋을 이용해 LoRA 방식으로 추가학습한 한국어 특화 대규모 언어모델 구름을 공개하였다[7].

2.2 Cross-lingual transfer

다양한 언어 간의 사전 훈련 모델에서 특정 언어로의 전이 학습을 고려할 때 중요한 점은 원래 학습된 언어의 기능을 손상시키지 않아야 한다는 것이다.[8] 이 문제를 해결하기 위해 하나의 언어로만 학습된 모델에서 모든 트랜스포머 계층을 동결하고, 새로운 임베딩 메트릭스만을 동일한 마스크된 언어 모델링 목표를 사용하여 학습을 시키는 방법이 제안되었다. 이 접근법은 단일 언어 모델이 다양한 언어를 통하여 일반화 하는 점에 대한 연구를 하였다.[9] 하지만 이는, 모델의 사전 학습된 일반화 되는 추상화를 학습을 하는 것에 의존하지만, 이는 모델의 추상화 학습이 정확하지 이해하지 못한다는 단점을 지니고 있다. 현재 다국어에 대한 연구는 기존에 잘 학습된 언어에서 타겟 언어에 대한 성능 향상으로의 전이하기 위한 방법을 모색하고 있다. 특히, 양방향 언어 사전을 활용하여 소스 언어의 토큰 임베딩을 대상 언어로 매핑하는 방법[10]이나 다양한 언어의 의미적 이해를 통해 영어 외의 언어에서도 사전 학습된 모델의 성능을 극대화하는 방식이 주목받고 있다.[11] 본 연구는 이러한 최근의 연구 트렌드를 바탕으로 언어 간 의미적 이해를 깊게 파악하고, 한국어의 특성을 모델에 반영하여 사전 지식을 효과적으로 활용하는 방식으로 성능 향상을 추구하였다.

2.3 Instruction tuning

Instruction tuning에 관한 연구는 사람의 피드백을 통한 강화 학습을 활용하여 사람들이 선호하는 모델을 개발하는 방식을 통하여 주목을 받기 시작하였다[12]. 이 접근법은 기존 모델보다 높은 사람들의 선호도와 개선된 일반화 성능을 보였지만 추가적인 데이터셋 구축에 사람의 개입이 필요하므로 비용이 크게 소모되는 문제가 있었다. 이러한 단점을 해결하기 위해 이미 뛰어난 성능을 가진 기존 모델을 활용하여 스스로 지시 데이터를 생성하고 학습하는 방법이 제안되었다[13]. 이 방법은 사람의 개입을 최소화하고자 하였으며, 그로 인하여 모델을 만드는데 있어 비용의 부담을 줄일 수 있었다. 이러한 방식들의 지시 데이터를 활용한 미세조정은 더욱 확장되어 다양한 작업과 명령 데이터셋 구축에 대한 연구로도 이어 졌으며, 더 많은 지시데이터를 활용한 모델의 미세조정을 통해 다양한 작업에 대한 명령 변경에 대한 모델의 민감도를 줄이는 효과가 있었다[14]. 이런 연구를 바탕으로 하여 Open AI의 api인 text-

[1] https://channeltech.naver.com/contentDetail/18

davinci-003를 활용하여 지시 데이터를 생성하는 방법으로 데이터를 증강하여 LLaMA에 instruction tuning을 한 Alpaca 모델이 공개가 되었다. 본 논문은 이러한 기존 연구들을 기반으로 alpaca 모델의 지시 데이터를 활용하여 미세조정 방법에 대한 연구를 수행하였다.

3. 방법

3.1 Cross-lingual transfer learning

대규모 언어 모델을 활용하여 영어 지식을 한국어로 전환하는 데 번역 데이터셋을 활용했다. 이를 통해 모델은 한국어에 대한 언어 이해를 바탕으로, 번역 지시 데이터를 추가 학습하여 지시사항에 더욱 정밀하게 반응하게 된다. 하지만 모델이 가진 사전 지식을 활용하여 모델을 성능을 증가 하기 위해서는 타겟 언어로만의 학습이 아닌 기존에 학습된 언어에 대한 데이터로도 학습을 진행해 주어야 한다.[15]

따라서 번역데이터들 뿐만 아닌, 지시 데이터를 활용하여 모델이 가진 사전지식의 망각을 최소화 하고 사전 지식을 효과적으로 활용하게 되어 전체적인 학습 효율성의 향상을 이루게 된다.

전처리 단계에서는 “Translate the following sentences from English to Korean”라는 지시문을 사용해 번역 데이터셋을 처리하며 학습했다. 이는 본래의 번역 데이터셋의 경우 지시에 해당하는 프롬프트가 존재하지 않아 추가한 문장이다. 해당 지시문을 통해 모델은 기존에 가진 지식을 활용하여 영어 문장을 이해하고, 한국어 문장을 학습함으로써 한국어 생성 능력을 향상 시킬 수 있다.

본 논문에서는 LLaMA2-7B full fine-tuning을 진행하였다. 학습 데이터에 한국어만 사용하는 것이 아닌 영어 지시 데이터도 학습함으로써 full fine-tuning시 발생할 수 있는 성능 trade-off를 최소화 하고자 하였다. 또한, 번역 및 지시 데이터를 통해 Cross-lingual transfer learning의 효과를 가져올 수 있음을 보이고자 하였다.

3.2 데이터 전처리

지시 데이터는 지시(instruction), 입력(input), 그리고 출력(output)의 세 부분으로 이루어져 있다. 지시, 입력, 출력 각 부분 사이에 개행 문자를 삽입함으로써 각각의 요소들을 구분할 수 있도록 하였다. 또한, 학습시에 지시, 입력을 제외한 출력 부분만 Cross-entropy Loss를 측정해 학습을 진행하였다. 이를 통해, 언어모델은 사람의 지시에 대한 출력만 학습함으로써 더 좋은 응답을 할 수 있게된다.

4. 실험 및 평가 데이터

4.1 학습 데이터

학습 데이터는 표 1와 같다.

지시 데이터는 Alpaca와 KoAlpaca 데이터셋을 사용하였다.[2] [3] 본 연구에서는 전체 데이터셋을 사용하여 한국어에 대한 성능을 향상시키는 것뿐만 아니라, 영어 데이터의 성능을 유지하면서 한국어도 잘 수행하게끔 전이를 시키는 것을 목적으로 하여 전체 데이터셋을 사용하는 것보다 적은 데이터셋을 사용하여 Cross-lingual transfer learning의 성능을 보고자 한다.

	번역데이터	지시문 데이터	
데이터 타입	영한 데이터	영어	한국어
개수	200,302	52,002	49,620

표 1. 학습 데이터 개수

4.2 모델 학습

본 연구에서는 Meta의 LLaMA2-7B 모델을 full fine-tuning 하여 진행을 하였으며 모델의 성능을 향상[16]시키기 위해 LLaMA2 tokenizer에 사전을 더 확장하여 실험을 진행한 beomi님의 Ko LLaMA2 모델과 같은 tokenizer[17]를 사용하여 실험을 진행하였다.

실험 환경은 Nvidia RTX A6000 * 8를 사용하여 진행하였다. 이 외의 하이퍼 파라미터는 표2와 같다.

하이퍼 파라미터	값
Epoch	3
Learning Rate	2e-5
Total Batch size	128
AdamW beta	0.9, 0.95
Cosine Warmup Ratio	0.05

표 2. 모델 파라미터 개수

4.3 비교 모델

본 연구에서는 original LLaMA2-7b[4], 고려대 지시 데이터로 학습한 KoLLaMA2-7b[5], LLaMA2(beomi)[17]를 베이스 라인으로 설정하고, 이들을 학습된 모델과 비교하여 성능 평가를 수행하였다. KoLLaMA2-7b 모델은 153,000개의 지시 데이터를 사용하여 LoRA 방식으로 미세 조정이 이루어진 모델이며,

[2]https://github.com/Beomi/KoAlpaca
[3]https://github.com/tatsu-lab/stanford_alpaca
[4]https://huggingface.co/meta-llama/Llama-2-7b-hf
[5]https://huggingface.co/psymon/KoLlama2-7b

LLaMA2(beomi) 모델은 토크나이저 확장 및 다양한 한국어 데이터를 활용하여 추가 학습이 진행된 한국어 LLaMA2 모델이다.

본 연구에서의 학습 방식이 기존에 추가학습을 진행하여 개발된 모델들과 비교하여 어떠한 점에서 향상이 되었는지 비교되는지 성능을 평가 하였다.

4.4 상식 추론 평가

HellaSwag 데이터셋은 영문으로 구성되어 있으며, 이야기나 지시문 집합의 끝맺음 문장을 예제에서 선택해야하는 데이터셋이다. 해당 데이터셋을 사용하여 언어모델들을 평가하여 한국어를 학습함으로써 일어나는 catastrophic forgetting 정도를 측정 및 비교하고자 한다.

4.5 대안 선택 평가

COPA 데이터셋의 경우 500개의 질문의 테스트 데이터를 사용한 영어 데이터셋이며, 하나의 전제와 두개의 대안으로 구성되어 더 타당한 대안을 고르는 작업이다. 해당 성능 평가를 통하여 영어에 대한 인과 추론을 평가하고자 한다.

4.6 한국어 감성 분류 평가

NSMC 데이터셋은 네이버에서 제작한 영화 리뷰 분류용 데이터셋으로, 리뷰 내용이 긍정적이면 1, 부정적이면 0으로 라벨링 되어 있다. 본 연구에서는 모델이 한국어를 얼마나 잘 이해하는지 평가하기 위해 해당 데이터셋에서 500개의 샘플을 활용하여 평가를 수행하였다.

4.7 번역 성능 평가

AI-hub 일상생활 번역 병렬 말뭉치 데이터 중 1000개의 sample만을 뽑아내어 번역 성능을 테스트 하는데 사용하였다. 본 실험에서는 영어에서 한글로 번역하는 성능과 한글을 영어로 번역하는 성능에 대해 실험해 보았다. 해당 실험에서는 영어에서 한국어로의 번역과 한국어에서 영어로의 성능을 모두 평가한다.

5. 결과

5.1 상식 추론 평가

데이터 세트 HellaSwag를 사용하여 평가한 모델의 성능은 표 3에서 확인할 수 있다. 이 데이터 세트는 영어로 구성되어 있으며, 실험은 제로샷 방식으로 실험을 진행하였다. 그 결과 성능저하가 0.01 발생 된 것을 볼 수 있다. 그럼에도 불구하고 기존의 다른 모델들과 비교하였을때 한국어 데이터셋만을 활용하여 학습을 진행하는 것이 아닌 영어 데이터셋도 같이 학습을 하는 Cross-lingual transfer learning 방식이 기존의 다른 모델들의 한국어 데이터셋 만을 이용한 학습 방식과 비교하여 망각 현상 개선에 영향을 주었음을 시사한다.

5.2 대안 선택 평가

표3의 Copa dataset의 평가는 제로샷으로 진행하였으며, 기존 모델 대비 0.02 감소한 것을 확인 할 수 있다. 이는 망각 현상이 발생하였음을 보여주었으며, 한국어 학습으로 인하여 영어에 대한 성능저하가 일어났음을 시사한다.

5.3 감성 분류 평가

본 연구에서는 NSMC 테스트 데이터셋 중 500개의 샘플을 활용하여 모델들의 성능을 제로샷 방식으로 평가하였다. 평가는 다양한 프롬프트를 사용하여 수행을 진행하였으나, 본 표의 지시 프롬프트로는 "### Instruction:\n다음 문장은 긍정일까요 부정일까요?\n' + sent + '\n### Response:"를 활용하였다. 본 실험에서의 평가 기준은 모델이 단순히 '긍정' 혹은 '부정'이라는 단어만을 생성하는 것이 아닌, 더욱 확장된 문장 형태로 답변을 제시할 수 있는 경우를 고려하였다. 이는 모델이 생성한 텍스트에서 '긍정' 혹은 '부정'이라는 토큰의 존재 유무에 따라 각 리뷰를 긍정 혹은 부정으로 분류하였다. 모델이 정답 단어를 포함하지 않는 답변을 생성한다면 오답으로 간주가 되었기에 이진 분류임에도 불구하고 점수가 매우 낮은 경우가 발생하였다. 본 실험을 통해 깊이 파악하고자 한 주요 측면은 각 모델들이 한글 지시어에 대한 언어 이해 능력과 입력된 평가 문장에 대한 이해도였다. 연구를 통해 학습한 모델은 주어진 프롬프트를 높은 수준으로 이해하고 뛰어난 정답률을 보였으며, 한편으로는 다른 모델들이 지시어를 제대로 이해하지 못하고 부정확한 답변을 제시하는 경향을 강하게 보여주었다. 이러한 실험 과정을 통해 모델들이 한국어 지시어를 얼마나 정확히 이해하는지, 그리고 주어진 리뷰 문장에 대하여 어느 정도의 인식 능력을 가지고 있는지를 평가할 수 있었다.

데이터셋 언어	영어		한국어
데이터셋	HellaSwag	COPA	nsmc
LLaMA2	**0.57**	0.86	0.27
LLaMA2(Ko)	0.56	**0.88**	0.28
LLaMA2(beomi)	0.49	0.84	0.13
LLaMA2(Ours)	0.56	0.84	**0.54**

표 3. 각 데이터셋 별 결과 값

5.4 번역 성능 평가

표4번을 보면 비교 모델들과 비교 하였을때 번역 성능이 더 좋게 나온 것을 알 수 있다.

	한영 번역	영한 번역
LLaMA2	0.032	0.002
LLaMA2(ko)	0.052	0.005
LLaMA2(beomi)	0.012	0.004
LLaMA2(Ours)	**0.057**	**0.111**
	한영 번역(한글)	영한 번역(한글)
LLaMA2	0.033	0.002
LLaMA2(ko)	0.001	**0.029**
LLaMA2(beomi)	-	-
LLaMA2(Ours)	**0.055**	0.014

표 4. 번역 성능 평가

영한 번역의 경우 매우 큰 성능 향상을 보여주었으며, 원래의 LLaMA2 모델과 비교하여 약 55배의 성능 향상을 보여주었다. 또한 다른 한국어로 학습된 모델들하고 비교하여도 큰 성능향상을 보여주어, 한국어 번역 성능이 매우 크게 향상됨을 보여주었다. 또한 다른 모델들의 경우 번역 작업을 수행한뒤 같은 문장을 반복해서 만드는 문제점을 보이거나 번역된 문장을 생성을 못하기도 하는 등의 문제를 보였다. 연구에서 만든 모델의 경우 번역 작업을 수행한 뒤 같은 문장을 반복하는 문제가 발생하지 않았으며 번역된 텍스트만을 출력으로 내보내는 등의 더 나은 모습을 보였다. 한영 번역의 경우에도 더 나은 성능을 보였으며 마찬가지로 다른 모델들의 경우 출력하지 못하는 등의 문제가 발생하였으나, 본 연구의 모델의 경우 번역된 문장만을 출력하는 모습을 보이며 안정적인 모습을 보여주었다. 지시 문장을 한글로 했을때도 번역 성능이 향상됨을 알 수 있었으며, 기존의 모델들 보다도 전반적인 텍스트에 대한 이해도가 올랐음을 알 수 있다. 번역 작업에 대한 prompt를 한글로 해서 줄 경우에도 연구에서의 모델은 한글 지시를 어느정도 이해하고 번역된 문장을 생성해 주었다. 그럼에도 불구하고 모델은 입력값을 그대로 출력하는 모습을 보이는 등의 취약한 모습을 보여주었다. 반면 한글로 영어 번역을 진행할 경우 모델은 잘 이해하지 못하는 모습을 보여주며, 성능 향상이 비교적 낮은 것을 알 수 있다.

6. 한계

본 연구에서의 한계는 번역 데이터셋으로 학습을 진행하였음에도 BLEU 점수가 높게 나오지 못하였으며, 이는 연구 과정에서 BLEU 점수를 측정할때 입력 값과 비교하여 다양한 번역 결과물이 나올 수 있지만 번역 실험에서는 하나의 문장만을 생성하게 하는 제약을 두고 실험을 진행하였기 때문에, 기존의 다른 연구들에 비해 점수 자체가 낮게 나왔다. 하지만 본 연구를 통해서 성능 향상이 있었다는 것을 증명한다.

번역 성능의 경우 한글로 지시하여도 영어로의 번역은 잘 수행 능력이 향상 되었음을 보였으며, 한글로 영어 지시를 입력할 경우 모델들은 영어 입력 값을 그대로 출력하거나, 부분적인 번역만 수행하는 등의 매우 취약한 모습을 보여주었으며, 영어로 지시했을때 나타났던 문제들이 한글로 지시할 경우 반복하는 등의 문제가 나타남을 확인할 수 있었다. 이는 모델이 한국어 지시를 모델에게 줄때 취약함을 볼 수 있는 부분이다.

7. 결론

본 연구는 대규모 언어 모델에서 Cross-lingual transfer learning의 중요성과 효과를 살펴보는 것에 집중하였다. 특히 한국어와 같이 영어와 구조적 유사성이 상대적으로 적은 언어에 대한 전이 학습의 진행이 가능한지에 대해 연구를 진행하였다. 또한 연구를 통하여 한국어 데이터셋만을 이용하여 학습을 진행하는 방법 보다, 한국어와 영어를 동시에 활용한 데이터셋을 활용하여 전이학습을 진행하였다. 이러한 접근 방식은 모델이 본래 가지고 있던 사전 지식을 망각을 최소화 하도록 하면서, 특히 한국어와 같이 영어와의 언어적 유사성이 차이 나는 언어에 대해 모델의 언어 이해 성능을 향상시킬 수 있는 방법을 제시하였다. 마지막으로 연구에서는 데이터셋의 일부만을 활용하였음에도 불구하고 상당한 성능 향상을 이룰 수 있음을 보였다.

참고문헌

[1] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J. H. Clark, L. E. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, K. Robinson, S. Ruder, Y. Tay, K. Xiao, Y. Xu, Y. Zhang, G. H. Abrego, J. Ahn, J. Austin, P. Barham, J. Botha, J. Bradbury, S. Brahma, K. Brooks, M. Catasta, Y. Cheng, C. Cherry, C. A. Choquette-Choo, A. Chowdhery, C. Crepy, S. Dave, M. Dehghani, S. Dev, J. Devlin, M. Díaz, N. Du, E. Dyer, V. Feinberg, F. Feng, V. Fienber, M. Freitag, X. Garcia, S. Gehrmann, L. Gonzalez, G. Gur-Ari, S. Hand, H. Hashemi, L. Hou, J. Howland, A. Hu, J. Hui, J. Hurwitz, M. Isard, A. Ittycheriah, M. Jagielski, W. Jia, K. Kenealy, M. Krikun, S. Kudugunta, C. Lan, K. Lee, B. Lee, E. Li, M. Li, W. Li, Y. Li, J. Li, H. Lim, H. Lin, Z. Liu, F. Liu, M. Maggioni, A. Mahendru, J. Maynez, V. Misra, M. Moussalem, Z. Nado, J. Nham, E. Ni, A. Nystrom, A. Parrish,

M. Pellat, M. Polacek, A. Polozov, R. Pope, S. Qiao, E. Reif, B. Richter, P. Riley, A. C. Ros, A. Roy, B. Saeta, R. Samuel, R. Shelby, A. Slone, D. Smilkov, D. R. So, D. Sohn, S. Tokumine, D. Valter, V. Vasudevan, K. Vodrahalli, X. Wang, P. Wang, Z. Wang, T. Wang, J. Wieting, Y. Wu, K. Xu, Y. Xu, L. Xue, P. Yin, J. Yu, Q. Zhang, S. Zheng, C. Zheng, W. Zhou, D. Zhou, S. Petrov, and Y. Wu, "Palm 2 technical report," 2023.

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

[3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[4] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," 2022.

[5] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," https://github.com/tatsu-lab/stanford_alpaca, 2023.

[6] I. Kim, G. Han, J. Ham, and W. Baek, "Kogpt: Kakaobrain korean(hangul) generative pre-trained transformer," https://github.com/kakaobrain/kogpt, 2021.

[7] N. . A. Lab and H.-I. A. research, "Kullm: Korea university large language model project," https://github.com/nlpai-lab/kullm, 2023.

[8] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1568–1575, Nov. 2016. [Online]. Available: https://aclanthology.org/D16-1163

[9] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.421

[10] B. Minixhofer, F. Paischer, and N. Rekabsaz, "WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models," *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022. [Online]. Available: https://doi.org/10.18653%2Fv1%2F2022.naacl-main.293

[11] W. Zhu, Y. Lv, Q. Dong, F. Yuan, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li, "Extrapolating large language models to non-english by aligning languages," 2023.

[12] 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.

[13] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions," 2023.

[14] Z. Xu, Y. Shen, and L. Huang, "Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning," 2023.

[15] S. Schuster, S. Gupta, R. Shah, and M. Lewis, "Cross-lingual transfer learning for multilingual task oriented dialog," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3795–3805, Jun. 2019. [Online]. Available: https://aclanthology.org/N19-1380

[16] Q. Zeng, L. Garay, P. Zhou, D. Chong, Y. Hua, J. Wu, Y. Pan, H. Zhou, R. Voigt, and J. Yang, "Greenplm: Cross-lingual transfer of monolingual pre-trained language models at almost no cost," 2023.

[17] L. Junbum, "llama-2-ko-7b (revision 4a9993e)," 2023. [Online]. Available: https://huggingface.co/beomi/llama-2-ko-7b