Super In-Context Learning을 활용한 생성 방법론

홍성태°†, 이승준†,김경민*§, 임희석*†§
고려대학교 컴퓨터학과†, Human-inspired AI 연구소§
{ghdchlwls123,dzzy6505,totoro4007,limhseok}@korea.ac.kr

Generation Methodology Using Super In-Context Learning

Seongtae Hong°†, Seungjun Lee†,Gyeongmin Kim*§, Heuiseok Lim*†§
Department of Computer Science and Engineering, Korea University†, Human-inspired AI Research§

요 약

현재 GPT-4와 같은 거대한 언어 모델이 기계 번역, 요약 및 대화와 같은 다양한 작업에서 압도적인 성능을 보이고 있다. 그러나 이러한 거대 언어 모델은 학습 및 적용에 상당한 계산 리소스와 도메인 특화 미세 조정이 어려운 등 몇 가지 문제를 가지고 있다. In-Context learning은 데이터셋에서 추출한 컨택스트의 정보만으로 효과적으로 작동할 수 있는 효율성을 제공하여 앞선 문제를 일부 해결했지만, 컨텍스트의 샷 개수와 순서에 민감한 문제가 존재한다. 이러한 도전 과제를 해결하기 위해, 우리는 Super In-Context Learning (SuperICL)을 활용한 새로운 방법론을 제안한다. 기존의 SuperICL은 적용한 플러그인 모델의 출력 정보를 이용하여 문맥을 새로 구성하고 이를 활용하여 거대 언어 모델이 더욱 잘 분류할 수 있도록 한다. Super In-Context Learning for Generation은 다양한 자연어 생성 작업에 효과적으로 최적화하는 방법을 제공한다. 실험을 통해 플러그인 모델을 교체하여 다양한 작업에 적응하는 가능성을 확인하고, 자연어 생성 작업에서 우수한 성능을 보여준다. BLEU 및 ROUGE 메트릭을 포함한 평가 결과에서도 성능 향상을 보여주며, 선호도 평가를 통해 모델의 효과성을 확인했다.

주제어: In-Contect learning, SuperICL, PLUG-IN

1. 서론

거대 언어 모델 (Large Language Models)이 등장하면서 기계 번역, 요약 및 대화 등의 다양한 자연어처리 태스크에서 최고 성능을 달성했다. OpenAI의 GPT-4[1]는 자연어 이해 및 생성 작업에서 인간 수준의 능력을 보여준다. 이러한 뛰어난 성과들로 인해 거대 언어 모델은 현대 인공지능 연구의 중요한 주제 중 하나로 각광받고 있다. 그러나 거대 언어 모델은 몇가지 중요한 문제와 한계를 가지고 있다. 먼저, 대규모 학습 데이터와 막대한 수의 모델 파라미터를 필요로한다. 이는 학습과 배포 과정에서 상당한 컴퓨팅 리소스와 비용을 요구하기 때문에 모델을 사용하는데 제약사항이 많이 존재한다. 또한, 거대 언어 모델은 다양한 도메인에 알맞게 미세 조정하기 어렵다는 문제도 있다. 특히 새로운 도메인의 전문지식을 필요로하거나 새로운 작업에 적응하는 것은 매우 어려운 과제이다.

이러한 문제를 해결하기 위한 방안으로 In-Context Learning 방법론이 자연어 처리 분야에서 주목을 받고 있다. In-Context Learning은 주어진 데이터의 컨텍스트 정보를 활용하여 거대 언어 모델을 더 효과적으로 조정하고 최적화하는 방법론이다. 이 방법론은 실시간으로 새로운 정보나 패턴을 학습하는 방식이기 때문에 매우 유연하고, 적은 양의 데이터 셋에서도 효율적인 새로운 지식을 습득할 수 있어 효율성이 높다. 그러나 In-Context Learning의 몇 가지 한계점[2]이 존재한다. 첫째, 컨텍스트로 주어지는 샷의 개수와 순서에 매우 민감하며, 형식이나 패턴을 완벽하게 따르지 않을 수 있다는 점이다. 둘째, 언어 모델의 최대 입력 토큰 길이이다. 많은 컨텍스트 정보를 제공하기 위해서는 많은 토큰이 필요하며, 언어 모델의 최대 입력 토큰 길이로 인해 정보를 모두 담기에 한계가 있다. 이러한 제한을 극복하기 위해 본 연구에서는 SuperICL을 생성 태스크에 적용하므로써 Super In-Context Learning for Generation (SuperICL4Gen)를 제안한다. SuperICL4Gen은 SuperICL을 생성 태스크로 확장하여 성능을 개선한 방법론이다. 본 연구는 실험을 통해 플러그인 모델의 교체를 통해 다양한 작업에 적응할 수 있는 가능성을 확인하고, 자연어 생성 작업에서 우수한 성능을 평가했다. 실험 결과에서는 BLEU, ROUGE의 평가 지표를 사용하여 SuperICL4Gen 방법론을 사용한 모델을 기존 모델과 비교했다. 평가 지표를 통해 제안 모델이 기존 모델을 상당히 능가하는 결과를 얻었으며, 또한 선호도 평가에서도 모델의 효과성을 확인했다. 이로써 SuperICL4Gen 방법론은 자연어 생성 작업에서의 혁신적인 성과를 입증하고, 자연어 처리 분야에 새로운 가능성을 제시하는 연구임을 보여준다.

2. 관련 연구

2.1 In-Context Learning

In-Context Learning(ICL)은 처음으로 GPT-3[3] 논문에서 소개된 개념으로, 모델의 수정이나 업데이트 없이 거대 언어

*교신저자 (Corresponding author)

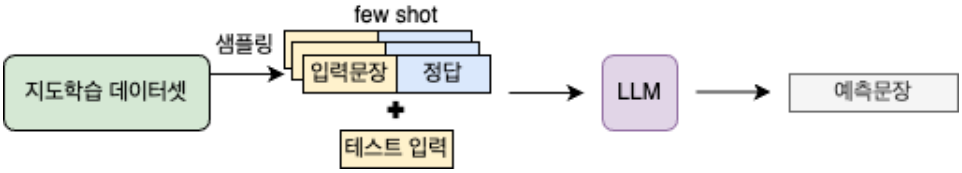

그림 1. ICL을 이용한 컨택스트 구성 및 학습 방법

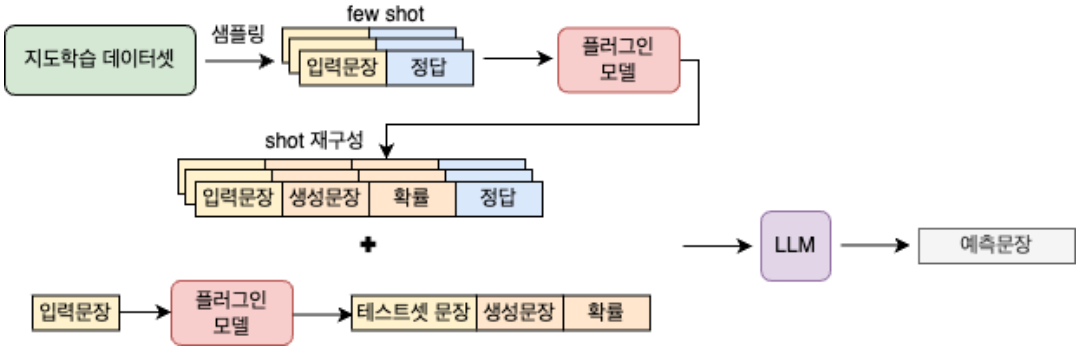

그림 2. SuperICL4Gen을 이용한 컨택스트 구성 및 학습 방법

모델을 새로운 태스크에 활용할 수 있게 도와주는 새로운 접근 방식이다. 일반적으로 특정 문맥에서 정보나 기술을 습득하는 학습 방법을 의미한다. In-Context Learning은 거대 언어 모델이 이 방법을 찾고 예측하는 학습을 가능하게 하기 위해 테스트 입력 앞에 몇 가지 학습 샘플을 컨텍스트로 추가하는 방법을 사용한다. 해당 방법론은 기계 번역 및 데이터 생성과 같은 태스크에서 널리 활용되고 있다. 이는 실시간으로 새로운 정보나 패턴을 학습하는 방식으로, 데이터 양이 적어도 효율적으로 새로운 주어진 컨택스트를 충분히 이해하고 따라하는 능력을 가질 수 있다. 그러나 In-Context Learning에는 몇 가지 제한사항이 있다. 첫째, 컨텍스트로 주어지는 샷의 개수와 순서에 민감하며, 완벽하게 따르지 않을 수 있다. 둘째, 언어 모델의 최대 입력 토큰 길이로 인해 많은 컨텍스트 정보를 담는 데 제한이 있다. 셋째, 거대 언어 모델이 충분한 컨텍스트 이해를 위해 필요한 샷의 품질이 매우 중요하며, 이를 수작업으로 샘플링하는 것은 번거로운 작업이다. 이러한 단점들을 극복하기 위한 여러 연구들[4]이 진행되고 있다.

2.2 Super In-Context Learning

Super In-Context Learning (SuperICL) [5]은 거대 언어 모델과 미세 조정된 모델을 결합하여 지도학습 작업의 성능을 향상시키는 방식을 제안하고 있다. 이 접근 방식은 두가지 단계로 수행되는데, 먼저 SuperICL의 첫 번째는 사전 학습 모델을 작업별 레이블이 지정된 데이터에서 미세 조정하는 것이다. 이 모델들은 플러그인 역할을 하며 분류 태스크에 대한 지식과 예측 결과를 제공한다. 그 다음 플러그인 모델이 제공한 지식을 활용하기 위해 거대 언어 모델에 대한 컨텍스트를 재구성한다. 재구성된 컨텍스트에는 예측된 레이블과 그에 따른 확률 값이 모두 포함된다. 추가적인 정보를 컨텍스트에 포함시켜 거대 언어 모델은 이를 고려하여 예측을 생성한다. SuperICL은 분류 태스크에서 미세 조정된 베이스라인 모델 및 In-Context Learning 모델에 비해 우수한 성능을 달성하고 불안정성 문제를 해결한다. 또한, 플러그인 모델이 정보를 흡수하도록 레이블을 예측하도록 하고 거대 언어 모델은 더 일반적인 언어 이해에 중점을 둘 수 있도록 하며 사전 학습 모델을 거대 언어 모델에 통합하여 지도 학습을 수행할 수 있는 방법을 보여주고 데이터 셋에서 적절한 샷을 선택해야하는 노고를 줄여준다. 이러한 결과와 함께 SuperICL은 거대 언어 모델 시대의 지도 학습에 대한 희망적인 새로운 패러다임을 제시한다. SuperICL4Gen과 SuperICL의 주요 차이점은 추가되는 정보에 각 분류 레이블 확률 대신 각 예측 문장의 토큰 생성 확률의 평균 값을 활용하여 기존의 확률값을 대체한다. 동일한 방법론을 생성 작업을 포함한 다양한 자연어 태스크에 보다 적합하게 활용될 수 있도록 변경하고 범위를 확장한 것이다.

3. 제안 모델

3.1 플러그인 모델 미세 조정

본 연구에서 제안 모델은 지도 학습 데이터셋을 활용하여 미세 조정된 생성 모델과 거대 언어 모델을 결합한다 . 이때, “플러그인 모델”이라는 용어는 다양한 자연어 생성 태스크에 대해 학습된 모델로서, 필요한 작업에 따라 다른 모델로 대체할 수 있는 기능을 갖춘 모델을 말하며 이러한 플러그인 모델은 SuperICL4Gen의 핵심 요소로 작용한다. 미세 조정된 모델은

표 1. SuperICL4Gen 방법론을 이용한 입력 및 결과 예시

Context	Input: 사우디아라비아의 원유생산시설이 지난 14일(현지시간) 예멘 반군의 드론에 피격돼 하루 570만배럴의 생산 차질을 빚게 됐다는 소식에 주식 시장에서 관련 업종... T5-Base Prediction:사우디의 원유생산시설이 예멘 반군의 드론에 피격돼 하루 570만배럴의 생산 차질을 빚게 됐다는 소식에 주식 시장에서 관련 업종의 희비가 엇갈렸다. Confidence: 0.8450 Gold: 사우디아라비아의 원유시설이 지난 14일 드론에 피격되면서 생산 차질로 인해 관련 업종인 정유·조선 업종은 오르고, 화학 업종의 경우 불안한 흐름을 보이고 있으며, 이로 인해 사우디는 16일까지 피격으로 인해 줄어든 원유 생상량의 30%를 복구할 계획이라고 밝혔다. ...
Test	Input:건축·부동산 전문가, 시민단체 등과 주택정책 토론회 개최 청주시는 29일 오후 4시 상당도서관 다목적실에서 ... T5-Base Prediction:청주시는 29일 오후 4시 상당도서관 다목적실에서 '주택정책 토론회'를 개최하여 청주시 주택시장의 문제점을 공론화하고 합리적인 주택정책 방향을 모색하기 위해 추진되었으며, 이근복 시 공동주택과장은 토론회를 통해 시민을 이해하고 시민과 함께하는 청주시가 되도록 노력하겠다고 말했다. Confidence: 0.8146
Prediction	Ground Truth: 청주시는 29일 '주택정책 토론회'를 개최하여 청주시 주택시장의 문제점을 공론화하고 합리적인 주택정책 방향을 모색하기 위해 노력할 것이라고 밝혔다.

언어 생성 작업에 필요한 특정 지식과 능력을 제공하며, 성능을 향상시키기 위해 함께 작동한다. 필요 시 다른 태스크를 수행하는 모델로 대체하여 다양한 자연어 생성 작업에 활용할 수 있다. 예를 들어, 요약 작업에 특화된 플러그인 모델을 사용하거나, 번역 작업을 위해 또 다른 플러그인 모델을 적용할 수 있다. 이는 SuperICL4Gen이 각 작업에 맞게 특정 모델을 유연하게 선택하여 활용하는 다목적 모델 역할을 수행할 수 있음을 말한다.

3.2 거대 언어 모델을 위한 입력 재구성

그림 1과 같이 기존의 In-Context Learning은 특정 작업에 관련된 컨텍스트를 입력으로 사용하는데 주로 데이터셋에서 샘플링한 예제들로 구성된다. 이러한 예제들은 입력 텍스트와 해당 작업의 정답을 사용하여 해당 태스크에 대한 컨텍스트을 구성한다. 즉, In-Context Learning은 데이터셋에서 가져온 예제들을 활용하여 문맥을 형성하고 특정 작업에 활용하는 방식을 채택하고 있다. 반면, 그림 2과 같이 SuperICL에서는 플러그인 모델의 생성 결과를 다시 거대 언어 모델의 입력으로 추가적인 정보를 제공한다. 각 분류 레이블 확률 값이 컨텍스트의 정보로 추가되는 SuperICL과 달리 SuperICL4Gen은 플러그인 모델을 통해 입력에 대한 예측 문장을 생성하고, 이 생성된 예측 문장을 활용하여 각 토큰 생성 확률의 평균 값을 계산하고 이를 레이블 확률 값으로 대체한다. 이 과정에서 학습 데이터셋에서 무작위로 샘플링된 샷을 사용하며, 이로써 다양한 문맥과 예제를 포함하므로 모델이 여러 상황에서 작동하고 적응할 수 있도록 도와준다. 각각의 샷은 다음과 같은 정보를 포함한다.

- 입력 문장: 특정 작업에 대한 입력 문장
- 생성 문장: 플러그인 모델에 의해 생성된 결과 문장
- 문장 생성 확률: 플러그인 모델이 예측한 결과에 대한 생성 확률
- 정답: 해당 샷에 대한 실제 정답 문장

테스트 입력 문장 또한 마찬가지로, 플러그인 모델을 사용하여 입력문에 대한 출력과 확률을 고려하여 입력을 재구성한다. 이러한 방식을 통해 SuperICL4Gen에서 거대 언어 모델은 플러그인 모델의 예측 결과와 생성 확률을 고려하여 최종 예측을 생성한다. 불필요한 정보가 경우 주어진 예측 문장을 참고하지 않고 새로운 생성 결과를 재생성함으로써, 정확하고 신뢰할 수 있는 예측 달성이 가능하다.

4. 실험 환경

SuperICL4Gen 방법론을 이용하여 요약문 생성 태스크에 대한 성능을 평가하기 위해 실험을 수행한다. 본 실험에서 플러그인 모델로 KE-T5 Base를 기반으로 사용한다. KE-T5 Base[6]은 [7]모델을 한국어와 영어 코퍼스를 이용하여 사전학습한 모델이다. 미세조정을 위해 사용된 데이터셋은 다양한 주제에 대

한 요약문 데이터를 포함하고있는 AIHub 문서요약 텍스트 데이터셋을 사용하였으며 학습데이터는 23,505개, 검증 데이터 6,715개를 사용한다. 제안 모델의 평가를 위해 테스트 데이터 100개를 사용하였으며 랜덤으로 추출된 3개의 샷을 컨텍스트로 제공하며 각 모델마다 주어진 샷은 동일하다.

4.1 학습 설정

플러그인 모델인 KE-T5-Base[6]는 멀티헤드,레이어 수는 각 12개이고 768차원의 벡터를 가지며 사전 단어의 크기는 64,128개이다. 실험을 위한 모델의 하이퍼 파라미터는 배치 사이즈 128, 학습률 1e-4, 입력 최대 길이 256, 출력 최대 길이 64, 조기 학습 종료 조건 5로 설정한다. 거대 언어 모델로 사용한 모델은 GPT-3.5-TURBO[8]이다. GPT-3[3]를 개선하여 자연어를 이해하고 생성할 수 있으며 최대 토큰 4,097개를 입력으로 사용 가능한 모델이다.

4.2 평가 방법

양적 평가 방법 본 연구에서는 생성된 자연어 요약문의 성능을 평가하기 위해 BLEU[9]와 ROUGE[10] 두 가지 주요 평가 메트릭을 사용한다. BLEU와 ROUGE는 자연어 처리 연구에서 널리 인정받는 평가 지표이다. BLEU는 기계 번역 및 요약 생성과 같은 작업에서 널리 사용되는 평가 메트릭 중 하나이다. BLEU는 생성된 텍스트와 참조 텍스트 간의 단어 및 구문 일치를 측정하여 평가한다. ROUGE는 주로 요약 생성 및 문서 비교 작업에서 사용되는 평가 방법으로, 생성된 요약문과 참조 요약문 간의 일치 정도를 측정한다. 이 두 평가 메트릭을 사용함으로써, 제안 모델의 자연어 요약문 생성 성능을 객관적으로 평가하고 비교할 수 있다.

선호도 평가 방법 본 연구에서 생성된 두 문장의 선호도 평가를 위해 GPT-3.5-Turbo 모델을 활용하며 temperature를 0으로 설정하여 정확한 선호도 평가를 수행한다. 최근 거대 언어 모델을 이용한 평가는 사람의 주석없이도 높은 상관관계를 보인다. 거대 언어 모델은 대량의 다양한 데이터에서 학습되는데 이는 모델이 다양한 언어와 주제를 이해하고 자연스럽게 생성할 수 있도록 도와준다. 학습 데이터의 다양성은 모델이 선호도 평가를 다양한 상황에서 수행할 수 있는 기반을 제공한다. 또, 거대 언어 모델은 동일한 입력에 대해 일관된 예측을 수행하며 주관적인 편견을 가지지 않기 때문에 모델이 사용자의 주관을 반영하지 않고 객관적인 기준에 따라 선호도를 판단한다는 점을 의미하므로 객관성 있는 선호도 평가 결과를 보장한다. 이러한 두 가지 측면은 언어 모델을 선호도 평가에 사용함으로써 더 신뢰할 만한 결과를 얻을 수 있다. 지시어로 사용한 문장은 다음과 같다.

1, 2번 중 어느 문장이 요약문 입력에 대해 잘 요약했는지 번호를 알려줘 (단, 비슷한 경우 tie)

1번 문장: {1번 문장 입력}
2번 문장: {2번 문장 입력}

결과:

5. 실험 결과

5.1 양적 실험 결과

표 2. 정량평가 결과 및 성능 비교

model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
KE-T5-Base	14.15	0.5845	0.4061	0.4502
GPT-3.5 ICL	14.45	0.6069	**0.4470**	0.4475
SuperICL4Gen	**15.88**	**0.6185**	0.4346	**0.4559**

표 3는 SuperICL4Gen 방법론을 사용한 모델의 요약문 생성 성능을 보여주는 표이고, 표 1는 모델의 결과를 보여주는 예시이다. 요약문 생성 테스크에서 제안 모델인 SuperICL4Gen은 베이스라인 모델보다 BLEU, ROUGE 스코어가 베이스라인 모델에 비해 각각 12.23%, 5.82%, 7.02% 및 1.27% 더 높은 성능을 보인다. 일반적인 In-Context Learning을 사용한 GPT-3.5 ICL 모델과의 비교에서 SuperICL4Gen을 사용한 모델이 GPT-3.5 ICL 모델에 비해 ROUGE-2에서 낮은 결과를 보이지만 BLEU 점수에서 9.9% 더 높은 점수를 기록하고, ROUGE-1, ROUGE-L에서 각각 1.91%, 1.88% 더 높은 결과를 보인다.

5.2 선호도 평가

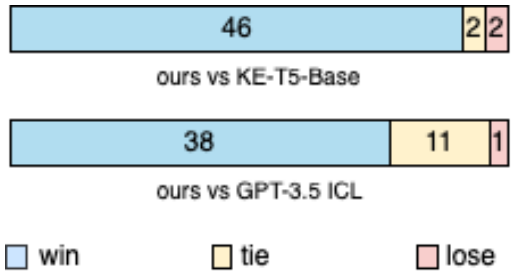

그림 3. SuperICL4Gen 선호도 평가 결과

선호도 평가를 위해 무작위로 선택된 테스트 데이터셋 50개에 대해 평가를 수행한다. 선호도 평가는 GPT-3.5-Turbo 모델을 사용하여 진행되었으며, 결과는 그림 3에서 확인할 수 있다. 베이스라인 모델인 KE-T5-Base 모델과 비교에서 제안 모델은 압도적인 성능 차이를 보인다. 이 비교에서 SuperICL4Gen은 46번의 승리를 기록하였으며, 무승부 2건을 포함한 48건의 긍정적인 결과를 확인할 수 있었다. 다음으로, GPT-3.5 ICL

과의 비교에서도 제안 모델은 높은 성능을 보인다. 11건의 패배가 존재하였으나 38번의 선호도를 차지하는 결과를 확인할 수 있다.

5.3 샷의 개수에 따른 성능 분석

In-Context Learning은 샷의 개수에 따라 성능이 큰 폭으로 차이 나고 거대 언어 모델이 패턴을 잘 이해할 수 있게 하려면, 샷 선택 시 높은 퀄리티의 컨텍스트를 수동으로 선택해야하는 단점이 있다. SuperICL4Gen에서는 적은 샷의 개수에서도 강건하며 컨텍스트에 추가로 제공되는 정보만으로도 거대 언어 모델이 효과적으로 잘 활용하는 능력을 보여준다.

표 3. 샷의 개수에 따른 성능 비교

model	n-shot	ROUGE-1	ROUGE-2	ROUGE-L
GPT-3.5 ICL	1-shot	0.5332	0.4185	0.3984
	3-shot	0.6069	0.4470	0.4475
SuperICL4Gen	1-shot	0.5920	0.4224	0.4534
	3-shot	0.6185	0.4346	0.4559

표 3는 샷의 개수에 따른 모델 간의 성능 비교이다. 3-shot에서 1-shot으로 샷의 개수를 감소시킬 때 ROUGE-1, ROUGE-2 및 ROUGE-L 점수는 각각 12.14%, 6.38%, 10.97%의 큰 폭의 성능 감소를 보인다. GPT-3.5 ICL은 주어지는 샷의 개수에 따른 큰 폭의 성능 차이를 확인할 수 있다. 그러나 SuperICL4Gen에서는 3-shot에서 1-shot으로 샷의 개수를 감소하더라도 일반적인 In-Context Learning에 비해 적은 폭의 차이를 나타내며 강건한 성능을 유지하는 것을 확인할 수 있다. 이는 컨텍스트 내 추가적인 정보만으로도 성능이 향상될 수 있으며, 샷의 개수에 변화에 민감하지 않음을 나타낸다.

5.4 실험 결과 및 분석

양적 실험 결과에서 제안 모델이 문장 수준의 일치도 및 요약 능력에서 강점을 보여주고 있음을 확인할 수 있다. 전반적으로 SuperICL4Gen 방법론을 적용한 모델은 KE-T5-Base와 GPT-3.5 ICL과의 비교 지표에서 일관된 우수한 성능을 보인다. 이는 컨텍스트에 추가적으로 제공하는 정보가 효과적 활용되어 요약하는데 도움이 되는 것을 알 수 있다. 선호도 평가 실험에서 단순히 In-Context learning 방법론을 사용하여 입력과 정답으로 구성된 컨텍스트를 제공하는 것보다 더 많은 정보(확률 및 예측 문장)를 제공하여 생성 태스크를 수행하는 방법론이 더 효과적임을 보여준다.

6. 결론

본 연구는 분류에서만 활용되던 SuperICL을 자연어 생성 분야에 적용했다. 제안한 SuperICL4Gen 방법론을 이용하여 플러그인 모델의 미세 조정과 거대 언어 모델의 입력 재구성을 통해 자연어 요약문 생성 작업에서 우수한 성능을 보였다. 실험 결과에서 확인된 정량적인 평가 지표에서 기존의 베이스라인 및 In-Context Learning 방법론을 채택한 모델을 상당히 능가하는 결과를 얻었으며, 선호도 평가에서도 우수한 성능을 확인했다. 대화 시스템, 문서 요약, 기계 번역 등의 생성 태스크에도 활용 가능성을 시사한다. 향후 연구에서는 거대 언어 모델이 플러그인 모델의 예측 결과와 확률 값을 더 잘 해석하여 생성하는 연구가 필요하다. 이를 통해 모델의 더 나은 설명력을 확보하고, 예측 결과를 신뢰할 수 있는 방향으로 이끌 수 있을 것이다. 결론적으로, 사전 학습 모델을 거대 언어 모델에 통합하여 지도 학습을 수행할 수 있는 방법론을 제시하고 미세 조정된 모델의 활용성을 보여준다. 이는 자연어 생성 작업에 새로운 가능성을 열어주며 자연어 처리의 다양한 응용 분야에서 활용할 수 있을 것이다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기술기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2022-2018-0-01405)

이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2021R1A6A1A03045425)

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기술기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-00368, 뉴럴-심볼릭(neural-symbolic) 모델의 지식 학습 및 추론 기술 개발)

참고문헌

[1] OpenAI, “Gpt-4 technical report,” 2023.

[2] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, “What makes good in-context examples for gpt-3?” *arXiv preprint arXiv:2101.06804*, 2021.

[3] 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.

[4] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, “A survey for in-context learning,” *arXiv preprint arXiv:2301.00234*, 2022.

[5] C. Xu, Y. Xu, S. Wang, Y. Liu, C. Zhu, and J. McAuley, “Small models are valuable plug-ins for large language models,” *arXiv preprint arXiv:2305.08848*, 2023.

[6] K. AIRC, “Ke-t5: Korean english t5,” Mar. 2021. [Online]. Available: https://github.com/AIRC-KETI/ke-t5

[7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485–5551, 2020.

[8] OpenAI-Blog, “Chatgpt: Optimizing language models for dialogue,” 2022. [Online]. Available: https://openai.com/blog/chatgpt/

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Jul. 2002. [Online]. Available: https://aclanthology.org/P02-1040

[10] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” *Text Summarization Branches Out*, pp. 74–81, Jul. 2004. [Online]. Available: https://aclanthology.org/W04-1013