한국어 맞춤법 교정을 위한 초거대 언어 모델의 잠재적 능력 탐색

구선민[1,○], 박찬준[2], 박정배[3,*], 임희석[1,3,*]
고려대학교 컴퓨터학과[1], Upstage[2], Human-inspired AI 연구소[3]
{fhdahd, insmile, limhseok}@korea.ac.kr, chanjun.park@upstage.ai

Examining the Feasibility of Utilizing a Large Language Model for Korean Grammatical Error Correction

Seonmin Koo[1,○], Chanjun Park[2], JeongBae Park[3,*], Heuiseok Lim[1,3,*]
Department of Computer Science and Engineering, Korea University[1], Upstage[2], Human-inspired AI Research[3]

요 약

최근, 대부분의 태스크가 초거대 언어 모델로 통합되고 있을 정도로 많은 관심 및 연구되고 있다. 초거대 언어 모델을 효과적으로 활용하기 위해서는 모델의 능력에 대한 분석이 선행되어야 하나, 한국어에 대한 분석 및 탐색은 상대적으로 부족하다. 본 논문에서는 한국어 맞춤법 교정 태스크를 통해 초거대 언어 모델의 능력을 탐색한다. 맞춤법 교정 태스크는 문장의 구조 및 문법을 이해하는 능력이 필요하며, 사용자의 만족도에 영향을 미칠 수 있는 중요한 태스크이다. 우리는 맞춤법 세부 유형에 따른 ChatGPT의 제로샷 및 퓨샷성능을 평가하여 초거대 언어 모델의 성능 분석을 수행한다. 실험 결과 제로샷의 경우 문장부호 오류의 성능이 가장 우수했으며, 수사 오류의 성능이 가장 낮았다. 또한, 예제를 더 많이 제공할수록 전체적인 모델의 성능이 향상되었으나, 제로샷의 경우보다 오류 유형 간의 성능 차이가 커지는 것을 관찰할 수 있었다.

주제어: 초거대 언어 모델, 한국어 맞춤법 교정, 성능 분석

1. 서론

최근 자연어처리 분야에서 ChatGPT [1], FLAN [2], LLaMA [3] 등 초거대 언어 모델(Large Language Models) 의 등장으로 다양한 태스크에서 관심받고 있다 [4]. 초거대 언어 모델은 의미 정보를 이해하기 위한 언어 지식을 가지고 있다고 평가 받고 있다. 맞춤법 교정, 질의 응답, 텍스트 요약, 기계 번역, 논리 추론, 코드 디버깅 등 다양한 자연어 처리 작업에서 엄청난 성능을 보이고 있으며, 이로 인해 기존의 각기 다르게 연구되었던 태스크들이 초거대 언어 모델의 사용으로 통합되고 있다.

초거대 언어 모델을 효과적으로 활용하기 위해서는 모델의 능력에 대한 분석이 선행되어야 한다. 이를 위해, 기존 연구에서 다양한 태스크에서 초거대 언어 모델의 능력에 대한 탐색이 이루어져 왔다. 초거대 언어 모델의 전통적인 지식 기반 질의 응답 시스템 대체 가능성에 대하여 탐색한 연구에서 일부 데이터셋에 대해서는 SoTA(State-of-the-art)를 넘었지만, 최신 데이터셋에 대해서는 아직 기존의 시스템에 미치지 못한다는 결과를 얻었다 [5]. 텍스트 요약 분야에서는 초거대 언어 모델이 생성한 뉴스가 인간과 동등한 수준이라는 평가를 받았다 [6, 7]. 특히, 도메인을 세분화한 쿼리 기반 분석에서는 레딧 게시물, 뉴스 기사, 대화 도메인 전반에 걸쳐 ChatGPT가 생성한 요약이 정답보다 우수함을 보였다 [8]. 기계 번역 분야에서도 고자원 언어에 대해 우수한 성능을 보였다 [9].

그러나 이러한 연구들은 대부분 영어와 같은 고자원 언어를 대상으로 분석이 이루어져 왔다. 언어마다 고유한 특성이 다르며, 일부 연구에서 초거대 언어 모델이 저자원 언어에서는 고자원 언어보다 능력을 발휘하지 못한다는 것이 밝혀졌다 [9]. 따라서 초거대 언어 모델을 한국어에 더욱 효과적으로 활용하기 위해서는 한국어를 대상으로 탐색 및 분석이 필요하다.

이를 위해 우리는 맞춤법 교정 태스크를 대상으로 한국어의 초거대 언어 모델의 능력을 탐색해보고자 한다. 맞춤법 교정은 주어진 문장에서 오류를 감지하고 교정하는 태스크이다. 문장의 의미를 명확히 전달하기 위해 중요하며, 문장 구조 및 문법 규칙을 이해하고 교정할 단어를 생성하는 능력이 필요하다. 실용적 관점에서는 후처리기에 활용되어 사용자의 만족도를 향상시키는 역할을 한다. 맞춤법 교정 결과는 다양한 다운스트림 태스크의 입력으로 사용된다 [10]. 양질의 입력을 줄 수록 모델 성능에 긍정적 영향을 미치며 이는 나아가 최종 사용자의 만족도에도 영향을 미치게 된다. 한국어 언어에 대해 초거대 언어 모델의 잠재적 활용가치를 파악하기 위해 맞춤법 교정 태스크의 성능을 탐색 및 분석해보고자 한다.

구체적으로, 초거대 언어 모델 중 ChatGPT를 대상으로 다양한 오류 유형에 대해 분석해보고자 한다. 한국어 맞춤법 교정 데이터셋에 대하여 총 4가지(띄어쓰기, 문장부호, 수사, 맞춤법 및 문법) 오류에 대하여 제로샷 및 퓨샷 성능을 분석한다. 퓨샷의 예제는 데이터셋에서 미리 샘플링하여 성능 측정의 대상이 되는 문장들과 구분한다. 실험 결과 제로샷의 경우 문장부호 오류의 성능이 가장 우수했으며, 수사 오류의 성능이 가장 낮

*교신저자 (Corresponding author)

았다. 또한, 예제를 많이 제공할 수록 전체적인 모델의 성능이 향상되었다. 그러나 제로샷의 경우보다 오류 유형 간의 성능 차이가 더 커지는 것을 관찰할 수 있었다. 이를 통해 한국어 맞춤법 교정 태스크에 대한 초거대 언어 모델의 능력 및 한계점에 대한 이해를 제공하며, 이는 연구자들에게 유용한 지침이 되기를 기대한다.

2. 관련 연구

2.1 맞춤법 교정

맞춤법 교정은 주어진 문장에서 오류를 감지하고 교정하는 태스크이다. 맞춤법 교정 모델은 주로 시퀀스-투-시퀀스 모델을 이용하여 수행되었다 [11, 12]. 초거대 언어 모델 등장 이전에 다양한 방법론을 이용하여 맞춤법 교정 모델의 성능 향상을 위한 연구들이 이루어져 왔다. 자기 지도 학습을 이용하여 학습 손실을 통해 데이터의 난이도를 측정 및 이용하여 성능을 향상시킨다 [13]. 또한 모델의 교정 성능 향상을 위해 디코딩 전략을 적용하여 교정 속도를 향상시키기도 하였다 [14].

그러나 모델 학습을 위해서는 대용량 병렬 말뭉치가 필요하며, 저자원 언어에서는 성능이 상대적으로 낮다. 따라서 이를 완화시키기 위해 저자원 언어에서의 성능을 향상시키기 위해 대조적 학습 방법을 적용하기도 하였다 [15]. 또한 병렬 말뭉치 생성을 위해 합성 데이터 생성 방법론을 적용하기도 하였다 [16]. 맞춤법 교정 태스크는 실용적 관점에서 교정 결과가 다양한 다운스트림 태스크의 입력으로 활용될 수 있으므로 모델 성능에 영향을 미칠 수 있으며, 교육적 관점에서도 학습자에게 오류 원인을 함께 제공해주어 학습을 돕는 역할로 활용될 수 있는 중요한 태스크이다 [10, 17].

2.2 초거대 언어 모델

최근 ChatGPT [1], FLAN [2], LLaMA [3] 등과 같은 초거대 언어 모델이 자연어 처리 분야에서 상당한 관심을 받고 있다. 이러한 모델은 대량의 텍스트 데이터를 기반으로 학습되었으며 맞춤법 교정, 질의 응답, 분류, 기계 번역을 포함한 다양한 자연어처리 태스크에서 높은 성능을 달성하였다.

초거대 언어 모델의 맞춤법 교정 성능에 대한 잠재력을 파악하기 위하여 분석이 수행되었다. 상용화 시스템과의 성능을 비교하여 초거대 언어 모델의 높은 성능을 보여주거나, 기존의 다양한 맞춤법 교정 모델들을 비교 대상으로 분석을 수행하기도 하였다 [18, 19]. 이를 통해 맞춤법 교정에 대한 초거대 언어 모델의 잠재력이 강조되었다.

그러나 기존 연구들은 대부분 고자원 언어를 대상으로 초거대 언어 모델의 잠재력을 조사하였으며, 또한 오류 유형을 세분화여 분석하지 않았다. 통합적인 성능만으로는 초거대 언어 모델의 잠재력을 파악하기 어려우며 언어적 특성이 다르기 때문에 한국어를 대상으로한 잠재적 능력 탐색이 필요하다. 이를 위해 우리는 한국어를 대상으로 오류 유형을 세분화하여 모델의 세부적인 성능을 분석한다.

Do grammatical error correction on all the following sentences I type in the conversation. Always answer in Korean. —— Referring to the example, do grammatical error correction that fit the given sentences. An example of doing grammatical error correction is as follows: {{examples}}

표 1. 한국어 맞춤법 교정 프롬프트. 답변은 한국어로 생성해야 하며, 예제가 주어지는 경우 분석하는 오류 유형과 관련된 예제가 주어진다.

3. 제안하는 방법

맞춤법 교정 태스크는 문장 구조 및 문법을 이해하고 적절한 교정 단어를 생성하는 능력이 요구된다. 맞춤법 교정 결과는 다양한 다운스트림 태스크의 입력으로 활용될 수 있으므로 모델의 성능 및 사용자 만족도에 영향을 미칠 수 있다. 본 논문에서는 이러한 중요성을 기반으로 ChatGPT를 대상으로 초거대 언어 모델의 한국어 맞춤법 교정 태스크의 성능을 검증하고자 한다.

검증의 대상은 문장에서 나타나는 맞춤법 오류를 4가지(띄어쓰기, 문장부호, 수사, 맞춤법 및 문법) 유형으로 세분화하여 검증한다. 띄어쓰기 오류의 경우 문장 내에 띄어쓰기 규칙에 위배되는 경우가 포함된 경우이다. 문장부호 오류의 경우 문장 사이에 문장 부호가 부착되지 않거나, 잘못된 위치에 부착하는 경우이다. 수사 오류의 경우 기수를 서수로 잘못 변환한 경우 등 수사 부분에서 오류가 나타난 경우이다. 이는 문장 내에 숫자가 포함되어 있을 경우에만 발생 가능하다. 맞춤법 및 문법 오류의 경우 문장 내의 일부 어절이 삭제, 추가, 교체, 분리 등의 맞춤법 및 문법 오류가 발생한 경우이다.

또한 예제를 함께 주었을 때 성능에 미치는 영향을 분석하기 위하여 데이터셋에서 오류 유형 별로 퓨샷 예제를 랜덤 샘플링 및 분할하여 검증 문장들과 중복되지 않도록 모델의 프롬프트에 포함시킨다. 한국어 언어의 맞춤법 교정 성능을 분석을 위해 OpenAI에서 제공하는 API를 사용한다. 태스크 수행을 위해 사용한 프롬프트는 표 1와 같다. 이후 연구자들이 고자원 언어와 한국어 성능을 비교하기 용이하도록 고자원 언어에서 ChatGPT의 맞춤법 교정 성능을 분석할 때 사용한 프롬프트를 기반으로 구성하였다 [18]. 기존의 프롬프트에 ChatGPT가 한국어로 답변을 생성 및 퓨샷 성능 측정을 위한 예제 제공 프롬프트를 추가하였다.

ChatGPT의 한국어 맞춤법 교정 세부 성능 비교를 위해 오류 유형 별로 성능을 탐색한다. 또한 예제를 함께 주었을 때의 성능을 분석하기 위하여 4가지 샷(1, 4, 8, 16)으로 구성하여 검증 대상의 유형과 관련된 예제를 함께 주어 성능을 조사한다.

4. 실험

4.1 데이터셋

한국어 맞춤법 교정 데이터셋을 검증에 활용한다 [20]. 한국어 맞춤법 교정 검증용 데이터셋으로 다양한 오류 유형을 담은 3,000개의 문장으로 구성되어 있다. 각 유형 별로 퓨샷의 예제로 사용할 문장들과 실제 검증 대상의 문장들을 구분하여 사용한다.

4.2 실험 설계

우리는 OpenAI에서 제공하는 ChatGPT api를 사용하며, GPT-3.5를 대상으로 실험을 수행한다. temperature는 0.5로 세팅하였으며, 제로샷 및 퓨샷 세팅으로 구분하여 실험한다. 퓨샷 세팅에 대해서는 예제 별 성능을 탐색하기 위하여 4가지(1, 4, 8, 16)에 대해 성능을 탐색한다. 성능 측정은 다양한 맞춤법 교정 연구의 평가 지표로 사용되는 BLEU [21] 및 GLEU [22] 점수를 활용한다.

4.3 실험 결과

전체 실험 결과는 표 2과 같다. 각 오류 유형의 예제를 주지 않은 제로샷의 경우 평균 성능은 BLEU 46.23점, GLEU 45.84점이다. 오류 유형별 성능은 띄어쓰기 오류의 경우 BLUE 및 GLEU가 각각 45.93점, 45.84점, 문장부호 오류의 경우 48.97점, 49.14점이다. 수사 오류의 경우는 각각 BLEU 42.92점 및 GLEU 38.98점을 보이며, 맞춤법 및 문법 오류는 BLEU 47.08점 및 GLEU 45.95점이다. 동일한 조건에서도 오류 유형에 따라 성능 차이가 나는 것을 관찰할 수 있다. 특히, 가장 높은 성능을 보이는 문장부호 오류와 가장 낮은 성능을 보이는 수사 오류의 성능 차이는 BLEU 기준 6.05점으로 무시할 만한 차이가 아니다. 이를 통해 초거대 언어 모델을 효과적으로 활용하기 위한 목적으로 잠재적 능력을 탐색할 때 유형을 세분화하여 분석할 필요성을 보여준다.

그림 1은 각 오류 유형별 제로샷 및 퓨샷 BLEU 성능을 그래프로 나타낸 것이다. 초거대 언어 모델이 맞춤법 교정을 수행하도록 할 때 각 오류 유형과 관련된 예제를 포함하여 프롬프트를 구성한 퓨샷의 경우 모든 경우에서 제로샷보다 성능이 향상되는 것을 관찰할 수 있다. 함께 주어지는 예제의 수가 많아질수록 성능이 지속적으로 향상되나, 일정 수준의 예제 수를 넘어가면 성능 향상의 폭이 줄어든다. 평균 점수를 기준으로 1-샷과 4-샷의 차이는 BLEU 점수 기준 58.50점 및 70.97점으로 12.47점의 향상을 보였다. 그러나, 4-샷과 8-샷 간의 차이는 4.03점으로 성능 향상의 폭이 1/3 수준으로 줄어든 것을 관찰할 수 있다. 8-샷 및 16-샷 간의 차이는 2.17점으로 더 감소한다. 이를 통해 오류 유형과 관련된 예제가 성능 향상에 도움이 되나, 일정 성능 이상부터는 제공하는 예제의 수와 비례해서 성능이 향상되는 것은 아님을 보인다.

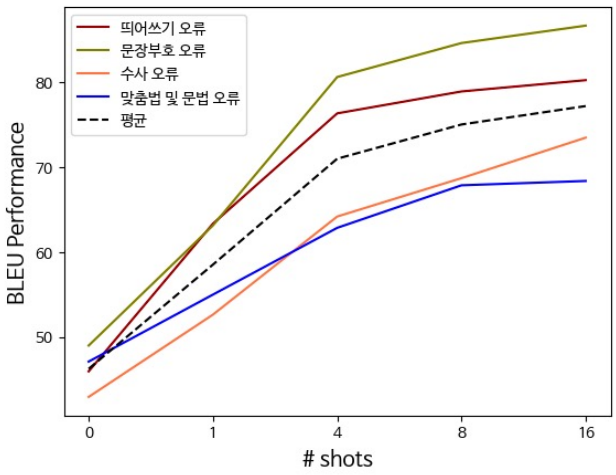

그림 1. 한국어 맞춤법 교정 태스크의 제로샷 및 퓨샷 성능

제로샷과 비교하였을 때 오류 유형 간 성능 차이가 커지는 것을 관찰할 수 있다. 제로샷의 경우 성능이 가장 우수한 유형과 저조한 유형 간의 차이가 BLEU 기준 6.05점이었으나, 16-샷의 경우 18.29점으로 유형 간 성능 차이 폭이 커지는 것을 관찰할 수 있다. 또한 동일한 수의 예제를 프롬프트에 포함시켰더라도 오류 유형에 따라 효과성의 차이가 나는 것을 보여준다. 1-샷의 경우 수사 오류의 성능이 가장 낮았으나, 16-샷의 경우 맞춤법 및 문법 오류의 성능이 가장 낮은 것을 확인할 수 있었다. 특히, 이 둘 간의 차이는 1-샷의 경우 BLEU 2.36점의 차이를 보였으나, 16샷의 경우 BLEU 5.1점으로 2배 이상의 차이를 보인다. 이를 통해 퓨샷 세팅에서 오류 유형 간 효과성의 차이가 있음을 보여준다. 즉, 초거대 언어 모델이 관련된 예제과 함께 한국어 맞춤법 교정을 수행할 때 모델의 능력을 더 잘 이끌어내는 오류 유형과 비교적 덜 이끌어내는 유형이 있음을 보여준다.

따라서 맞춤법 교정 태스크를 포함하여 초거대 언어 모델의 잠재적 능력을 탐색하기 위해서는 통합된 성능이 아닌 분류를 세분화하여 개별 성능을 분석하여 모델의 장점 및 장점을 분석하는 것이 중요하다. 동일한 세팅이이어도 오류 유형에 따른 모델의 강건성의 차이를 보이기 때문에 초거대 언어 모델을 효과적으로 활용하기 위해서는 세분화된 분석을 통해 모델의 능력 및 한계점을 이해해야 한다.

Type	# shots									
	0		1		4		8		16	
	BLEU	GLEU	BLEU	GLEU	BLEU	GLEU	BLEU	GLEU	BLEU	GLEU
temperature=0.5										
띄어쓰기 오류	45.93	45.84	63.32	61.40	76.31	74.07	78.89	76.18	80.22	77.71
문장부호 오류	48.97	49.14	63.11	60.69	80.59	78.75	84.59	82.71	86.64	85.14
수사 오류	42.92	38.98	52.61	46.95	64.16	58.34	68.68	63.85	73.45	67.63
맞춤법 및 문법 오류	47.08	45.95	54.97	53.34	62.83	54.85	67.84	65.92	68.35	66.46
Avg.	46.23	44.98	58.50	55.59	70.97	68.02	75.00	72.17	77.17	74.24

표 2. 한국어 맞춤법 교정 태스크의 오류 유형 별 ChatGPT-3.5 성능 분석 결과. Avg. 는 평균 성능을 나타낸다.

5. 결론

본 연구에서는 맞춤법 교정 태스크를 대상으로 한국어의 초거대 언어 모델의 능력을 오류 유형 4가지(띄어쓰기, 문장부호, 수사, 맞춤법 및 문법)에 대해 탐색하였다. 프롬프트를 구성할 때 오류 유형과 관련된 예제의 효과성을 알아보기 위해 제로샷 및 퓨샷으로 구성하여 분석을 수행하였다. 실험 결과 전체적으로 제로샷보다 퓨샷의 성능이 높았으나 세부 유형 간 예제를 포함시켰을 때 효과성의 차이를 보였다. 이는 초거대 언어 모델을 활용할 때 오류 유형에 따라 세분화된 접근이 필요함을 보여준다. 즉, 맞춤법 교정 태스크를 포함한 다양한 자연어 처리 태스크에서 초거대 언어 모델의 능력을 효과적으로 활용하려면 세분화된 분석을 통해 잠재적 능력을 탐색해야 함을 시사한다. 추후 연구에서는 더 세분화된 오류 유형 및 효과적인 프롬프팅 방법 및 예제의 변형을 통해 초거대 언어 모델의 잠재적 능력을 이끌어내 성능을 높이고자 한다.

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). 본 연구는 2022년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임 [S3310507].

참고문헌

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

[2] 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," *arXiv preprint arXiv:2109.01652*, 2021.

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

[4] L. Zhang, M. Wang, L. Chen, and W. Zhang, "Probing gpt-3's linguistic knowledge on semantic tasks," *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 297–304, 2022.

[5] Y. Tan, D. Min, Y. Li, W. Li, N. Hu, Y. Chen, and G. Qi, "Evaluation of chatgpt as a question answering system for answering complex questions," *arXiv preprint arXiv:2303.07992*, 2023.

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

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

[8] X. Yang, Y. Li, X. Zhang, H. Chen, and W. Cheng, "Exploring the limits of chatgpt for query or aspect-based text summarization," *arXiv preprint arXiv:2302.08081*, 2023.

[9] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla, "How good are gpt models at machine translation? a comprehensive evaluation," *arXiv preprint arXiv:2302.09210*, 2023.

[10] L. Feng, J. Yu, D. Cai, S. Liu, H.-T. Zheng, and

Y. Wang, "Asr-robust spoken language understanding on asr-glue dataset," 2022.

[11] A. Solyman, Z. Wang, and Q. Tao, "Proposed model for arabic grammar error correction based on convolutional neural network," *2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, pp. 1–6, 2019.

[12] A. Kuznetsov and H. Urdiales, "Spelling correction with denoising transformer," *arXiv preprint arXiv:2105.05977*, 2021.

[13] Z. Gan, H. Xu, and H. Zan, "Self-supervised curriculum learning for spelling error correction," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3487–3494, 2021.

[14] X. Sun, T. Ge, F. Wei, and H. Wang, "Instantaneous grammatical error correction with shallow aggressive decoding," *arXiv preprint arXiv:2106.04970*, 2021.

[15] H. Cao, W. Yang, and H. T. Ng, "Grammatical error correction with contrastive learning in low error density domains," *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4867–4874, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.419

[16] C. Park, J. Seo, S. Lee, C. Lee, H. Moon, S. Eo, and H.-S. Lim, "Bts: Back transcription for speech-to-text post-processor using text-to-speech-to-text," *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pp. 106–116, 2021.

[17] M. Kaneko, S. Takase, A. Niwa, and N. Okazaki, "Interpretability for language learners using example-based grammatical error correction," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7176–7187, May 2022. [Online]. Available: https://aclanthology.org/2022.acl-long.496

[18] H. Wu, W. Wang, Y. Wan, W. Jiao, and M. Lyu, "Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark," *arXiv preprint arXiv:2303.13648*, 2023.

[19] T. Fang, S. Yang, K. Lan, D. F. Wong, J. Hu, L. S. Chao, and Y. Zhang, "Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation," *arXiv preprint arXiv:2304.01746*, 2023.

[20] S. Koo, C. Park, J. Seo, S. Lee, H. Moon, J. Lee, and H. Lim, "K-nct: Korean neural grammatical error correction gold-standard test set using novel error type classification criteria," *IEEE Access*, Vol. 10, pp. 118 167–118 175, 2022.

[21] 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, 2002.

[22] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, "Ground truth for grammatical error correction metrics," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 588–593, 2015.