자연어 처리 태스크에 대한 기계와 인간의 성능 상관관계 연구

박서윤○, 김희재, 이성우, 강예지, 장연지, 김한샘‡
연세대학교 언어정보연구원[1], 언어정보학협동과정‡
{seoyoon.park, kimmejay, yjkang5009, yeonji3547, khss}@yonsei.ac.kr, sean05663@gmail.com

Exploring the Relationship Between Machine and Human Performance in Natural Language Processing Tasks

Seoyoon Park○, Heejae Kim, Seong-Woo Lee, Yejee Kang, Yeonji Jang, Hansaem Kim‡
Yonsei University, Institute of Language and Informatics
Yonsei University, Interdisciplinary Graduate Program of Linguistics and Informatics‡

요 약

언어 모델 발전에 따라 사람과 유사하게 글을 생성하고 태스크를 수행하는 LLM들이 등장하고 있다. 하지만 아직까지도 기계와 사람의 수행 과정에 초점을 맞추어 차이점을 드러내는 연구는 활성화되지 않았다. 본 연구는 자연어 이해 및 생성 태스크 수행 시 기계와 인간의 수행 과정 차이를 밝히고자 하였다. 이에 이해 태스크로는 문법성 판단, 생성 태스크로는 요약 태스크를 대상 태스크로 선정하였고, 기존 주류 사전 학습 모델이었던 transformer 계열 모델과 LLM인 ChatGPT 3.5를 사용하여 실험을 진행하였다. 실험 결과 문법성 판단 시 기계들이 인간의 언어적 직관을 반영하지 못하는 양상을 발견하였고, 요약 태스크에서는 인간과 기계의 성능 판단 기준이 다름을 확인하였다.

주제어: 문법성 판단, 요약, ChatGPT, transformers, 성능 상관관계

1. 서론

초거대 데이터로 사전 학습된 LLM의 등장에 따라 기존 주류 모델 학습 방법이었던 파인튜닝(fine-tuning) 과는 다른 방식의 프롬프트 러닝(prompt learning), 인간 피드백 기반 강화학습(reinforcement learning with human feedback) 혹은 인스트럭션 튜닝(instruction tuning)[1]들이 LLM 학습을 위해 사용되고 있다. 이러한 방법론들은 인간의 선호도를 LLM의 수행에 반영하는 한편, 언어 모델들이 사람과 거의 유사하게 글을 생성하고 태스크를 수행하는 것을 가능하게 하였다.

이에 따라 초거대 언어 모델들이 언어를 어떻게 '습득' 하는지에 대한 궁금증부터 사람의 언어 사용 양상과 어떤 점이 유사하고 어떤 점에서 차이가 드러나는지에 대한 관심이 증대되고 있다. 본 연구에서는 이러한 연구 동기를 토대로 자연어 처리 태스크 수행 과정 관찰을 통해 사람과 기계의 수행 방식에 대해 살펴보았다. 이를 위한 태스크로는 문법성 판단 태스크와 요약 태스크를 선정하였으며, 사람의 수행 방식 혹은 결과와 기계의 방식, 결과 간 상관관계를 비교함으로써 이러한 차이점에 대해 관찰하고자 한다. 실험 시에는 기존 주류 모델인 transformer 기반 언어 모델, 그리고 현재 가장 널리 쓰이고 있는 LLM인 ChatGPT 3.5를 대상으로 실험을 진행하였다.

2. 관련 연구

2.1. 문법성 판단 태스크

'문법성 수용 가능성 판단 태스크(grammaticality judgement task)' 은 생성 언어학자들이 인간의 언어 능력과 문법 지식을 관찰하고 이해하기 위해 사용하는 주요 평가 척도이다[2]. 자연어 처리에서 문법성 판단 태스크는 신경망이 문법적 개념을 익혔는지를 판단하는 태스크로써 이때 문법적 개념은 인간의 언어적 능력 측면에서의 개념이다. [2]의 CoLA(Corpus of Language Acceptability) 데이터셋은 언어학 문헌에서 발췌한 영어 문장으로 구성되어 있으며, 해당 문장이 정문인지 비문인지 라벨링 되어 있어 문장에 대한 수용성을 판단하는 이진 분류 태스크를 수행할 수 있다. CoLA는 GLUE 벤치마크[3]에 포함되어 언어 모델의 문법적 수용성 판단 능력을 평가하는 기준 태스크로 활용되고 있다. 다만 [2]에 따르면 기계의 CoLA 성능은 인간의 성능과 비교했을 때 인간의 성능에 크게 미치지 못하며, 이에서 알 수 있듯이 문법성 판단은 까다로운 태스크이기도 하다. 이처럼 문법성 판단 태스크는 인간의 언어적 직관을 반영하고 있는 태스크이기 때문에 본 연구에서도 기계의 '자연어 이해' 측면을 측정하고자 문법성 판단 태스크를 사용하였다.

2.2. 요약 태스크

최근의 요약은 원문에서 내용을 그대로 옮겨오는 추출 요약(extractive summary)이 아닌 '추상 요약(abstractive summarization)' 으로 이루어진다 [4]. 그

동안의 추상 요약은 어려운 태스크로 받아들여졌으나 [5], BART[6] 등 자연어 생성에 특화된 모델들의 개발과 ChatGPT[1]와 같은 생성 AI의 등장으로 점차 수월해지고 있다. 요약을 평가하는 전통적인 지표로는 대표적으로 BLEU[7], ROUGE[8] 등을 들 수 있다. 특히 ROUGE는 recall에 초점을 맞춘 지표로, 분모로 참조 요약에서 가능한 모든 n-gram을 포함하고 있어 복수의 후보 요약문에 대해서도 측정을 가능하게 하였다. 그러나 ROUGE는 인간의 평가와 상관관계가 낮기도 하며[9], 원문에 등장하는 표현이 패러프레이징 되어 제시되는 추상 요약을 제대로 평가할 수 없다는 문제점이 있다. 이에 BERTSCore[10] 등 원문과 요약문의 의미적 유사성을 측정하는 시도들이 있었으며, 최근에는 [11] 등과 같이 요약문을 측정하는 인간 평가 지표를 고안하고 실현하는 연구들이 많이 진행되고 있다[12]. 본 연구에서는 기계의 요약 성능을 ROUGE, BERTScore로 측정하여 정량적인 성능을 살펴보는 한편, 인간 평가를 통해 생성된 요약이 인간의 언어적 직관에도 부합하는지를 살펴보았다.

2.3. 자동 평가 지표의 문제점

BLEU, ROUGE 등 자동 평가 지표들은 인간 평가와 상관 관계가 높지 않다[13]. [13]에 따르면 자동 평가 지표만 사용 시 자동 측정 항목에 대한 성능 지표는 양호하나, 생성된 텍스트 내 정보가 부족할 수 있는 것은 발견하지 못하는 문제점이 있다. 또한 자동 평가는 인간 평가와 낮은 상관 관계를 자주 보이며, 다양한 연구에 반복적으로 나타나기도 한다. 이처럼 자동 평가 지표들은 언어적 특성을 평가하는데는 적합하지 않으나[14], 텍스트 품질을 대략적으로 계산할 때 실용적이고 빠르게 반복해서 사용할 수 있다는 점에서 널리 쓰인다[15]. 다만 근본적으로 자연어 처리 결과 질을 향상하기 위해서는 인간 평가에 대한 연구가 꾸준히 필요하다. [16]은 계획적으로 잘 실행된 태스크에서 얻은 인간 평가는 학습 기반 측정 항목을 훈련할 수 있고 이러한 측정 항목의 품질을 향상시키는 데 도움이 될 수 있다고 언급하였다. 실제로 최근에는 인간 평가의 품질과 일관성에 대한 데이터가 기계학습 모델의 품질과 신뢰성을 제고하는 데에 적극적으로 활용되고 있는 추세이다.

3. 실험 데이터셋

3.1. 문법성 판단 데이터셋

문법성 판단 데이터셋으로는 국립국어원에서 2021년 공개한 '문법성 판단 말뭉치 v1.1[2]' 을 사용하였다. 문법성 판단 말뭉치는 문헌에서 발췌된 총 19,940개 문장에 대해 문헌 저자들이 판단한 문법성 라벨(비문:0, 정문:1)과, 20대부터 50대까지의 평가자들이 비문/정문을 7점 리커트 척도로 판단한 내용이 포함되어 있다. 기본적인 문법성 라벨 비율은 50:50으로 설정되었다. 리커트 척도는 전체 인원의 평균 리커트 척도 점수와 표준편차, 그리고 연령별 평균과 표준편차로 제시되었다.

본 연구에서는 19,940개 문장을 모두 사용하였으며 실생활에서의 언어 현상을 잘 반영할 수 있도록 원전의 문법성 라벨 대신 전체 인원의 리커트 척도 점수 평균을 사용하여 3.5점을 기준으로 정문(3.5~7점), 비문(1~3.5점)으로 이진 라벨링을 진행하였다. 정문은 1, 비문은 0으로 표시하였고 최종적으로 라벨 비율은 정문 2, 비문 1의 비율로 재편되었다. 실험 시 데이터 분할 비율은 train/dev/test 각각 8:1:1로 상정하여 진행하였다.

source	acceptability _label	source _annotation	sentence	mean (total)	sd (total)	mean (20s)	sd (20s)	mean (30s)	sd (30s)	mean (40s)	sd (40s)	mean (50s)	sd (50s)
T00001	1		높은 달이 떴다.	4.981	2.273	5.371	2.059	5.273	2.198	4.095	2.508	4.667	2.41
T00001	0	+	달이 몸이 높았다.	2.223	1.794	2.5	1.905	2.065	1.459	1.85	1.785	2.32	2.036
T00002	1		실업는 사람이 까불까불한다.	5.884	1.493	6.306	1.238	5.71	1.575	5.95	1.276	5.44	1.781
T00003	1		나는 철수에게 공을 던졌다.	6.784	0.836	6.971	0.169	6.923	0.27	6.55	1.395	6.353	1.367
T00004	1		내가 순이와 물에서 다툰다.	4.286	2.187	4.735	2.079	4.152	2.181	3.583	1.976	4.714	2.644
T00004	0	+	내가 순이와 무리가 다툰다.	1.621	1.112	1.595	1.257	1.688	1.148	1.476	0.68	1.769	1.235
T00005	1		나는 부지런히 뛰었다.	6.68	0.942	6.892	0.315	6.781	0.608	6.429	1.326	6.231	1.691
T00005	0	?	나는 부지런히 뛰어졌다.	2.351	1.795	2.775	2.118	2.314	1.694	1.833	1.543	2.048	1.359
T00006	1		사랑이 죄는 아니다.	6.605	1.086	6.85	0.427	6.2	1.712	6.556	0.922	6.857	0.359
T00006	0	+	죄는 사랑이 아니다.	4.339	2.433	4.576	2.513	4.706	2.316	4.393	2.283	3.059	2.536
T00007	1		철수는 영이를 가면적으로 미워하였다.	3.554	2.151	3.909	2.31	3.059	1.969	3.464	2.027	4	2.345
T00007	0	+	철수에게는 영이가 가면적으로 미웠다.	3.02	2.044	2.813	1.975	3	2.118	3.5	1.96	2.867	2.232
T00008	1		그가 그에게 감기를 옮겼다.	6.465	1.162	6.563	1.19	6.206	1.321	6.85	0.489	6.333	1.291
T00008	0	+	그녀가 그에 의해서 감기를 옮겨졌다.	2.382	1.752	2.472	1.89	2.394	1.657	2.3	1.78	2.231	1.739

그림 1 국립국어원 문법성 판단 말뭉치(2021)

3.2. 요약 데이터셋

요약 데이터셋으로는 AIHUB에서 2020년 공개한 '도서자료 요약[3]' 을 사용하였다. 해당 데이터셋은 20만개의 문단(paragraph)과 이에 대한 요약문으로 구성되어 있으며, 요약은 추출(extractive) 요약이 아닌 생성 요약으로 이루어졌다. 데이터셋 내 요약문 길이는 300자에서 1000자 사이이며, 본 연구에서는 LLM 학습을 고려하여 평균 요약문 길이가 300 이내인 데이터 약 2만 건을 샘플링하여 활용하였다. 샘플링 시 도서 주제 등의 데이터 특성을 고르게 배분하였고 문법성 판단과 마찬가지로 요약 데이터셋도 8:1:1 비율로 분할하여 각각 train/dev/test 데이터로 사용하였다.

4. 실험 및 결과

4.1. 문법성 판단 태스크 수행 결과

앞서 언급한 바와 같이 문법성 판단 태스크는 주어진 문장에 대해 비문인지 정문인지를 각각 0과 1로 라벨링하는 이진 분류 태스크로 자연어 이해에 속한다. 본 연구에서는 human, transformer 그리고 LLM을 대상으로 실험을 진행하였고, 모델은 transformer 기반인 Electra와 ChatGPT 3.5를 사용하였다. 각 모델별 분류 성능은 f1 score로 측정하였다.

표 2 문법성 판단 정량 평가 지표(f1-score)

	human baseline	transformer	chatGPT 3.5
f1	**93.5**	62.3	59.1

실험에 앞서 test 데이터에 대해 human baseline을 구하였다. 기존 test 데이터 라벨과 비교한 결과 f1 score는 93.5로 측정되었다. 다만 향후 기계 성능 측정 시에는 정답에 대한 정확도를 높이고자 원래의 test 데이터를 사용하지 않고 human baseline 수행 결과에 대해 데이터 구축 및 검수 경험이 있는 3인이 논의하여 재라벨

1) https://chat.openai.com/

2) https://corpus.korean.go.kr/request/reausetMain.do

3) https://aihub.or.kr/

링한 데이터를 정답으로 활용하였다.

transformer 실험은 koelectra[4]를 사용하였고 해당 모델에 문법성 판단 데이터를 파인튜닝하여 이진 분류를 수행하였다. 하이퍼파라미터는 학습률 2e-5, 총 epoch은 5이다. 수행 결과 f1 score는 62.3으로 나타나 human baseline을 크게 밑도는 결과를 보였다. ChatGPT 실험의 경우 ChatGPT 3.5 turbo(0613)을 사용하였는데, 사전 실험에서 제로샷(zero-shot) 수행 시 특정 라벨로 편향되는 결과를 보여 최종적으로는 30개 예시에 대한 퓨샷(few-shot) 러닝 후 이진 분류를 진행하였다. 수행 결과 f1 score는 59.1로 나타나 transformer 모델보다 낮은 성능을 나타냈다.

4.1.1. 문법성 판단 경향성 분석

태스크 수행 과정에 대한 실험 간 상관관계를 분석하고자 카이제곱 검정(chi-square test)를 수행하였다. 카이 제곱 검정의 경우 빈도분석을 기반으로 두 집단이 동질한지 아닌지를 검정하는 방법으로[17], 본 연구에서는 문법성 판단 라벨 0, 1에 대한 빈도를 기반으로 검정을 진행하였다. 검정 시 원가설은 '두 실험 과정에는 차이가 없다.'로 설정하였고, transformer-human, ChatGPT-human,그리고 transformer-ChatGPT 실험을 비교하였다. 카이 제곱 검정 결과는 아래 표와 같다.

표 3 실험 과정별 카이 제곱 검정

실험	χ^2 결과	p value
transformer-human	$\chi^2(2, N=1051)=9.64$	$p<.001$***
ChatGPT-human	$\chi^2(2, N=1051)=5.49$	$p<.05$**
transformer-ChatGPT	$\chi^2(2, N=1051)=0.58$	$p=.446$

카이 제곱 검정 결과 transformer 및 chatGPT의 실험 결과는 영가설을 기각하여 human baseline 실험 과정과 차이가 있었음을 알 수 있었다. 반면에 transformer와 ChatGPT의 실험 과정은 영가설을 기각하지 못해 두 실험 과정에 차이가 없었다고 이야기할 수 있으며, 검정을 통해 인간의 태스크 수행 과정과 기계의 수행 과정에는 차이가 있음을 추론할 수 있다.

4.1.2. 문장 난이도에 따른 수행 분석

통계적 검정 외에도 실험 결과 간 차이를 분석하고자 문장 난이도별 수행 결과를 분석하였다. 난이도별로 분석을 수행하는 이유는 태스크 수행 시 기계가 인간의 언어적 직관 및 수행 경향성을 따르는지 확인할 수 있기 때문이다. 난이도는 앞서 human baseline을 수행한 사람들이 문제 해결 시 느낀 난이도를 '상', '중', '하'로 주석하였다. 분석은 a) 모든 실험 정답, b) human 정답-기계(transformer & LLM) 오답, c) transformer만 오답 세 가지 경우의 수에 대해 진행하였다. 기본적으로 데이터셋의 난이도별 사례 및 비율은 상 87개 (8.28%), 중 174개(15.56%), 하 790개(75.2%)로 난이도 '하'의 비중이 높다.

분석 결과 모두 정답을 맞춘 a)의 경우 난이도별 비율이 데이터 전체의 비율과 거의 비슷하게 나타나는 것을 확인할 수 있다. 그러나 사람만이 정답을 맞춘 b)의 경우 난이도 '상, 중'의 비율은 늘어난 반면 '하'의 비율은 줄었다. 이는 곧 기계가 사람에 비해 난이도가 높은 문제는 잘 해결하지 못하는 경향이 있다고도 말할 수 있으며, 기계의 태스크 수행 경향이 사람의 언어적 직관과는 다르다는 것을 추론해볼 수 있다. c), d)의 경우 transformer와 chatGPT의 정량적 성능 지표와도 관련이 있으며, 두 모델의 편향성을 드러내는 결과임을 확인하였다.

표 4 난이도별 문법성 판단 태스크 수행 결과 분석

실험	정답: 0	정답: 1	난이도별 사례(비율)
a) 전체 정답	244	199	상: 39(8.8%) 중: 77(17.38%) 하: 327(73.81%)
b) human만 정답	98	70	**상: 22 (13.1%)** **중: 35 (20.8%)** 하: 111 (66.1%)
c) transformer 만 오답	44	**142**	상: 7 (3.76%) 중: 27 (14.52%) 하: 152 (81.72%)
d) ChatGPT만 오답	**198**	55	상: 19 (7.48%) 중: 35 (13.78%) 하: 200 (78.74%)

먼저 d)의 경우 c)에 비해 많은 오답 수를 보이는데 이는 ChatGPT의 정량적 지표가 transformers보다 떨어지는 것과 관련된다. 또한 상대적으로 ChatGPT는 transformers보다 난이도가 높은 문제를 잘 해결하지 못하며, 이에 대한 원인으로는 한국어 지식 부족, 혹은 불충분한 퓨샷 학습 등을 생각해볼 수 있다. 이외에도 transformer는 오답 생성 시 0을, chatGPT는 1을 리턴하는 편향성을 드러내었다.

이처럼 문법성 판단이라는 자연어 이해 태스크 수행 시 사람과 기계의 태스크 수행 과정은 차이가 있으며, 기계의 경우 인간의 언어적 직관에 비추어 봤을 때 어려운 문제들을 잘 수행하지 못하는 것을 확인할 수 있었다. 또한 같은 기계라도 transformer와 chatGPT의 수행에는 편향성이나 성능 차이 등의 차이점이 존재하였다.

4.2. 요약 태스크 수행 결과

요약 태스크도 문법성 판단 태스크와 마찬가지로 사람과 transformer, 그리고 ChatGPT의 수행 결과를 비교하였다. 다만 human baseline을 상정했던 문법성 판단 태스크와 달리 요약에서는 사람의 태스크 수행이 아닌 결과물에 대한 human evaluation을 사용하여 사람과 기계의 결과를 비교하였다. 정량적 지표로는 ROUGE와 BERTScore를 사용하였는데, 이는 단어 시퀀스(sequence)에 대한 n-gram을 바탕으로 하는 ROUGE만으로는 생성 요약을 정확히 측정할 수 없기 때문이다. 즉, 원문에 등장하지 않는 표현으로 요약문이 작성될 가능성이 있으므로 단어 중복 정도와 요약문 정답-생성 요약문 간 의미적 유사성을 추가로 고려하여 정량 평가를 진행하였다.

transformer 모델로는 kobart모델[5]을 사용하였다. 학

4) monologg/koelectra-base-v3-discriminator 사용

습은 batch 16, 학습률을 2e-5로 설정하여 총 10 epoch을 수행하였다. 수행 결과 ROUGE와 BERTScore 성능은 아래 표와 같다.

표 5 transformer를 사용한 요약 태스크
정량 평가 지표(ROUGE, BERTScore)

bart	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
성능	0.176	0.06	0.168	0.758

생성 요약을 진행하였기에 ROUGE는 매우 낮은 성능을 보였다. 그러나 BERTScore의 경우 ROUGE보다 높은 성능을 보여, 요약 원문의 의미와 유사한 요약문이 생성되었음을 확인할 수 있었다. 점수 차이가 많이 나기 때문에 피어슨, 스피어만 상관 계수로 두 점수 간의 관계를 살펴본 결과, 피어슨 상관 계수는 .747***, 스피어만 상관계수는 .71***로 두 지표 간 비교적 강한 양의 상관관계가 존재했다. 즉, 두 지표는 같이 증가하거나 감소하는 경향을 보인다.

ChatGPT 3.5를 사용한 요약의 경우 transformer로 수행한 요약 중 ROUGE-1과 BERTScore에 대해 각각 점수 상위 20위, 하위 20위에 대해 수행하였다. 수행 시 프롬프트 엔지니어링을 사용하여 요약문을 생성하였고, 길이는 연구 대상인 300자 이내의 생성 요약문과 동일하도록 300자 이내로 제한을 두었다. 해당 요약문들에 대해 ROUGE와 BERTScore로 성능을 평가한 결과, 아래와 같은 결과를 보였다.

표 6 ChatGPT를 사용한 요약 태스크
정량 평가 지표(ROUGE, BERTScore)

gpt	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
성능	0.171	0.05	0.165	0.752

결론적으로 transformer와 ChatGPT 3.5를 사용하여 생성한 요약문에 대한 정량적인 평가 지표는 두 경우 모두 비슷한 수치를 보였다. 그러나 사람이 transformer와 ChatGPT의 결과물을 육안으로 검수한 결과, 두 모델의 결과물 사이에는 질적인 차이가 존재했다. 이는 곧 ROUGE와 BERTScore만으로는 요약문의 질을 온전히 평가할 수 없는 것을 의미한다.

4.2.1. 요약 결과 인간평가

요약문에 대한 기계와 인간의 차이를 분석하고자, 인간이 평가한 점수와 정량적인 지표 간의 상관관계를 구하였다. 인간 평가는 [11]에서 제안한 4개 지표 중 fluency를 제외한 coherence, consistency, relevance를 중심으로 7점 리커트 척도를 산정하였다.

- Coherence(일관성): 요약문 구조, 조직이 얼마나 잘 이루어졌는지, 그리고 글 전반적으로 주제에 대한 일관적인 정보를 전달하고 있는지 판단
- Consistency(일치성): 요약문이 원문의 사실과 일치하는가? 요약문은 원문에 등장하는 사실들만 다루어야 하며, 허구가 있어서는 안됨.
- Relevance(관련성): 정보가 중복되거나 과도하게 포함되어 있는가? 즉, 원문에서 중요한 정보만을 뽑아 요약문을 구성하였는지 판단

평가자는 데이터 구축 경험자 2인이며, 점수를 평균내어 최종 점수로 산정하였다. 리커트 점수와 정량적 지표 간의 상관관계는 스피어만(Spearman) 상관관계로 구하였다. transformer 생성 요약문과 ChatGPT 생성 요약문에 대한 인간 평가는 아래 표와 같았다.

표 7 상, 하위 20위에 대한 인간 평가

	transformers	ChatGPT
상위 20위	8.98(SD=2.72)	15.63(SD=1.73)
하위 20위	4.3(SD=1.78)	12.85(SD=3.38)

평가 결과 전반적으로 상위 20위의 결과에 대한 인간 평가가 하위 20위의 결과에 대한 평가보다 더 좋게 나타났다. 또한 transformer가 생성한 요약문은 상, 하위에 대한 인간 평가 점수차가 2배 넘게 차이 나는 반면 ChatGPT 3.5가 생성한 요약문의 경우 상, 하위 20위에 대한 점수 간극이 좁혀졌다. 또한 ChatGPT의 인간 평가 점수보다 transformer에 대한 평가보다 높게 나타났는데 이러한 사실들은 transformer 기반 모델들이 아직도 인간의 언어적 직관에 부합하는 요약문을 생성하는 것에 미흡하며, 반면에 ChatGPT는 사용자가 만족할 만한 요약문을 생성하고 있는 것을 의미한다. 이는 ChatGPT가 글을 '생성' 하는 목적을 가진 생성AI의 한 종류이기 때문인 것으로 추측된다.

4.2.2. 요약 태스크 경향성 분석

육안 검수 시 요약문의 정량적 지표가 좋다고 하여 좋은 요약이 아니었던 것에 착안해 엄밀한 분석을 진행하고자 전체 항목에 대한 상관관계 대신 상위 20위, 하위 20위 항목 각각에 대한 상관관계를 구하였다.

transformer 생성 요약문의 경우 특히 정량적 성능과 육안 검수 간의 간극이 컸었다. 아래 표는 transformers 생성 요약문 정량 평가 지표 상, 하위 20위와 인간 평가 점수 간의 스피어만 상관관계이다.

표 8 transformer 정량 평가 지표와
인간 평가 간 스피어만 상관관계

transformer	ROUGE-1 vs human	BERTScore vs human
상위 20위	-0.101(p=.673)	0.349(p=.131)
하위 20위	-0.09(p=.72)	0.08(p=.731)

상관관계 분석 결과, p value가 모두 유의 수준을 초과하여 결론적으로는 정량적 지표와 인간 평가 간의 상관관계가 관찰되지 않았다. 이는 transformer가 생성한 요약문의 정량적 평가 지표가 높다 하더라도 해당 요약문이 인간 평가 시 반드시 좋은 점수를 얻는 것을 의미

5) https://huggingface.co/ainize/kobart-news

하지 않으며, 4.2.1.과 마찬가지로 transformer 생성 요약문이 인간의 직관에 부합하지 않게 생성된다는 것을 의미한다.

ChatGPT에 대한 인간 평가와 정량적 평가 지표 간의 상관관계 역시 유의미한 관계를 나타내지 않았다.

표 9 ChatGPT정량 평가 지표와
인간 평가 간 스피어만 상관관계

chatGPT	ROUGE-1 vs human	BERTScore vs human
상위 20위	-0.185(p=.43)	0.54(p=.014)
하위 20위	0.30(p=.2)	0.38(p=.093)

이는 ChatGPT도 transformer 결과와 마찬가지로 높은 정량적 점수가 반드시 높은 인간 평가로 이어지지 않는다는 것을 의미하며, 인간의 평가는 ROUGE, BERTScore 등의 지표에서 측정하는 글자 순서, 의미적 유사성과는 다른 양상으로 진행됨을 알 수 있다.

5. 결론 및 향후 연구

본 연구에서는 사람과 기계의 자연어 이해, 생성 태스크 수행 과정을 비교하고자 문법성 판단 태스크와 요약 태스크의 각 주체 간 수행 결과를 분석하였다. 문법성 태스크 수행 결과, 사람과 기계 간 태스크 수행 과정에는 차이가 있음을 확인할 수 있었다. 또한 문법성 판단은 사람의 언어적 직관이 가장 잘 반영된 태스크이기에 사람은 높은 난이도의 문장도 잘 판별하는 반면, 기계의 경우 높은 난이도의 문장을 잘 분류하지 못하였다. 아울러 같은 기계라 하더라도 ChatGPT보다는 파인튜닝이 이루어진 transformer의 성능이 근소하게 높았으며 오답 생성 시 transformer와 ChatGPT 간 양상이 다르다는 것을 라벨 편향을 통해 알 수 있었다.

요약 태스크의 경우 transformer와 ChatGPT 간의 정량적 지표 간의 차이가 매우 적게 나타났다. 생성 요약 특성상 ROUGE보다 BERTScore가 높아 기계 생성 요약문이 원문의 의미는 비교적 잘 보존하고 있음을 알 수 있었으나, 정량적 지표가 높다고 하여 인간의 정성적 평가 결과와 반드시 유의미한 관계가 있지는 않았다. 즉, 기계 입장에서의 높은 요약 성능과 사용자 입장에서의 높은 요약 성능 간에는 차이가 있으며, 기존 의미적 유사성 통념에도 불구하고 인간의 평가는 의미적 유사성을 크게 고려하지 않는다.

이처럼 본 연구에서는 사람과 기계의 이해, 생성 태스크 수행 양상을 문법성 판단 태스크와 요약 태스크에 비추어 각각 살펴보았다. 연구 결과 사람과 기계의 수행에는 차이가 있는 것을 확인할 수 있었다. 그러나 파인튜닝 여부, 생성 AI 특성 등 차이에 대한 원인에 대한 추측만 가능하였을 뿐 구체적으로 어떤 요인으로 인해 사람과 기계 간의 수행이 차이가 나는지는 정확히 밝힐 수 없었다. 향후 연구에서는 사람과 기계의 수행을 다르게 하는 요인을 밝히는 한편 실험 과정 속에서 이를 관찰할 수 있는 방법론을 모색할 예정이다.

참고문헌

[1] OUYANG, Long, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[2] WARSTADT, Alex; SINGH, Amanpreet; BOWMAN, Samuel R. Neural network acceptability judgments. Transactions of the Association for Computational Linguistics, 2019, 7: 625-641.

[3] WANG, Alex, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461, 2018.

[4] Som Gupta, S. K Gupta, Abstractive summarization: An overview of the state of the art, Expert Systems with Applications, Volume 121, Pages 49-65, ISSN 0957-4174, 2019.

[5] Gehrmann, S., Deng, Y., & Rush, A. M. Bottom-up abstractive summarization. arXiv preprint arXiv:1808.10792., 2018.

[6] LEWIS, Mike, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019.

[7] PAPINENI, Kishore, et al. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002. p. 311-318

[8] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74-81, Barcelona, Spain. Association for Computational Linguistics.

[9] F. Liu and Y. Liu, "Exploring Correlation Between ROUGE and Human Evaluation on Meeting Summaries," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 1, pp. 187-196, Jan. 2010, doi: 10.1109/TASL.2009.2025096.

[10] ZHANG, Tianyi, et al. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675, 2019.

[11] FABBRI, Alexander R., et al. Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 2021, 9: 391-409.

[12] BOMMASANI, Rishi; CARDIE, Claire. Intrinsic evaluation of summarization datasets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. p. 8075-8096.

[13] VAN DER LEE, Chris, et al. Human evaluation of automatically generated text: Current trends and best practice guidelines. Computer Speech & Language, 2021, 67: 101151.
[14] SCOTT, Donia; MOORE, Johanna. An NLG evaluation competition? eight reasons to be cautious. In: Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation. 2007. p. 22-23.
[15] REITER, Ehud; BELZ, Anja. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Computational Linguistics, 2009, 35.4: 529-558.
[16] WELTY, Chris; PARITOSH, Praveen; AROYO, Lora. Metrology for AI: From benchmarks to instruments. arXiv preprint arXiv:1911.01875, 2019.
[17] 이용훈, R을 활용한 코퍼스언어학과 통계학, 한국문화사, 2016