https://doi.org/10.15749/jts.2023.24.4.005

번역학연구
24권 4호
(2023.12.31.)

AI Hub의 학습용 말뭉치 데이터의 활용가능성 모색: ChatGPT의 번역 품질평가를 중심으로

곽 은 주 · 노 재 훈 · 박 미 진 · 전 현 주
(세종대 · 와이즈에스티글로벌 · 국제통역번역협회 · 신한대)

1. 서론

1.1 연구의 배경

최근 몇 년간 인공지능(AI) 기술의 급속한 발전은 다양한 분야에 혁신적인 변화를 가져왔다. 특히, 언어 처리와 번역 분야에서 AI의 역할은 더욱 주목받고 있다. 2022년 11월, OpenAI의 ChatGPT 3.5 버전 공개를 계기로 전 세계는 순식간에 ChatGPT의 열풍에 휩싸이기 시작했다[1]. 프롬프트에서 쌍방향의 대화를 통해 원하는 정보를 실시간으로 제공되는 서비스 체험은 AI에 대한 막연했던 기대감을 일순간 확신으로 변화시켰다. 21세기의 인류는 드디어 AI 세상이 도래하였음을 직감하는 특별한 경험을 한 것이다. MS의 공동설립자 빌 게이츠(Bill Gates) 역시 이러한 경험을 통해 “ChatGPT는 인류의 삶에 혁신적인 변화

1) “최신 데이터에 따르면 현재 ChatGPT의 사용자 수는 1억 8천만 명 이상이다. ChatGPT가 출시된 후 단 5일 만에 사용자 100만 명을 달성하였다. 현재 ChatGPT 웹사이트의 월 방문자 수는 16억 명에 이른다. 참고로 페이스북의 사용자 100만 명 확보에 10개월이, 넷플릭스는 약 3.5년이 소요되었다”(NamePepper 2023. 11. 5.).

를 가져올 것이며, 인터넷의 발명에 버금갈 만큼 획기적이라는 격찬을 아끼지 않았다[2].

ChatGPT는 복잡한 언어 모델을 기반으로 사용자의 프롬프트 요청에 대해 실시간으로 대화 형식의 응답을 제공한다. 이러한 기술은 AI가 인간의 언어를 이해하고 생성하는 방식을 근본적으로 변화시켰으며, 특히 번역의 영역에서 그 잠재력을 주목받고 있다. ChatGPT 사용 경험과 관련하여 20대부터 50대의 1,000명을 대상으로 한국언론진흥재단의 미디어연구센터에서 진행한 '챗GPT 이용 경험 및 인식 조사[3]'에 관한 설문조사(2023년 3월 29일~4월 2일) 결과가 상당히 흥미롭다. 무엇보다도 ChatGPT의 활용성 예측 관련 문항에서 번역, 녹취, 자료정리(88.1%), 글쓰기(84.5%), 그리고 어학공부(80.4%)를 위한 목적으로 사용한다는 응답률이 매우 높게 나타났다. 반면, ChatGPT로 인하여 대체 가능성이 가장 높은 직업군을 번역가와 통역가로 응답한 비율이 90%를 상회하였다(양정애 2023: 9-13).

ChatGPT를 사용하는 주요 목적에 번역이 포함된 사실과 동시에 통번역사의 직업 전망이 가장 불투명하다는 상반된 결과는 좀 더 진지하게 생각해볼 여지가 있다. 이는 번역이 인간의 고유한 영역이냐 아니냐의 문제가 아니라 실제로 ChatGPT의 번역 성능이 어느 정도이기에 이러한 결과를 도출하는가에 관한 문제이다. ChatGPT는 구글 번역(Google Translate)이나 파파고(Papago) 혹은 딥엘(DeepL)과 같은 번역에 특화된 범용의 엔진이 아니라 대량의 텍스트 데이터로 학습한 대규모의 언어 생성 모델(Large Language Model, LLM)이다. 그럼에도 언어 이해와 생성은 물론 번역 성능 또한 상당한 수준에 이르렀다는 함의는 번역학적인 관점에서 검증의 필요성을 강력하게 소환하기 때문이다.

1.2 연구의 목적 및 방법

ChatGPT[4]의 번역 품질 수준에 대한 신뢰할만한 타당성을 확보하려면 무엇

2) "Microsoft co-founder Bill Gates believes ChatGPT, a chatbot that gives strikingly human-like responses to user queries, is as significant as the invention of the internet ..."(REUTERS 2023. 2. 10).

3) 한국언론진흥재단 미디어연구센터 설문조사.

4) 생성형 AI는 ChatGPT(OpenAI), LLaMa(Meta), BARD(Google), Bing(MS), Hyper

보다도 객관적인 번역품질평가(TQA, Translation Quality Assessment) 지표를 적용하여 항목별로 계량화된 결과치를 도출하는 등의 검증 절차가 필요하다. 번역 품질평가 수행 과정에서 우선적으로 고려해야 할 사항은 무엇보다도 ST와 TT를 기반으로 TT의 비교 대상 텍스트 쌍의 설정 범위와 방식에 관한 구체적인 개념을 적용해야 한다. 기본적으로 ST를 상수로 두고 범용의 기계번역을 적용한 MT(machine translation), 인간 번역사의 번역과 감수를 진행한 MTPE(machine translation post-editing), 그리고 ChatGPT의 번역결과 버전 등 최소한 세 쌍의 TT를 변수로 설정해야 한다. 비교 대상 텍스트를 선정할 때 인간 번역사가 수행한 MTPE를 번역 품질평가 기준으로 삼아야 하므로 객관적으로 신뢰할만한 테스트용 데이트 세트 확보 여부는 평가 결과의 신뢰성과 타당성을 담보하는 매우 중요한 관건이다. 특히 한국어와 영어의 언어 쌍을 기준으로 번역평가를 수행하기 위해서는 ChatGPT에 적용된 학습용 데이터의 불균형성을 염두에 두어야 한다. 다시 말해서 LLM을 기반으로 학습되는 생성형 AI의 학습용 데이터는 규모 측면에서 영어 텍스트가 절대적인 우위를 차지하며, 한국어는 상대적으로 그 격차가 매우 크다는 점이다. 따라서 한국어와 영어의 학습용 데이터의 불균형성을 최소화하기 위해서는 번역 품질 테스트용 데이트 세트의 선정 및 확보를 번역 품질평가 시 최우선적인 고려사항에 포함해야 한다.

이런 점에서 AI Hub(aihub.or.kr)에서 공개하는 '인공지능(AI) 학습용 말뭉치'를 테스트용 데이터로 활용하여 ChatGPT의 번역 품질을 평가하고 테스트용 데이터로서의 활용 가능성을 진단하고자 한다. 해당 말뭉치는 데이터의 수집, 가공, 처리 과정에서부터 신뢰성을 확보할 수 있는 체계적인 코퍼스 품질 관리 프로세스를 거쳤다. 그리고 번역을 수행하는 과정에서도 ST에 대한 기계번역을 거친 MT 데이터를 기반으로 인간 번역사의 MTPE 과정과 검수 및 번역 품질평가 과정을 거친 데이터이다. 이 과정으로 그친 것이 아니라 다시 공신력 있는 번역 품질평가 기관에서 시행하는 코퍼스 품질관리 가이드라인에 의거한 항목별 평가기준에 부합하여 일정 수준 이상의 검증 결과를 획득한 데이터에 해당한다. 이러한 객관성과 신뢰성을 기반으로 해당 말뭉치는 ChatGPT를 포함한

CLOVA X(Naver) 등 다양한 종류가 있지만 본고에서는 대표성을 지닌 ChatGPT를 중심으로 논의를 진행한다.

생성형 AI 플랫폼의 번역 성능 평가용 데이터로서 활용 가치가 매우 높을 것이라 상정하고 본격적인 논의를 전개하고자 한다.

본 연구의 목적은 AI Hub에서 공개하는 ‘AI 학습용 말뭉치’는 ChatGPT가 제공하는 번역 서비스의 품질평가용 테스팅 데이터로서 그 활용 가능성과 유용성을 모색하는 것이다. 이를 통해 AI 번역 기술의 현재 수준과 잠재적인 한계를 탐구한다. 그리고 본 연구는 다음과 같은 핵심 논제에 대한 해법을 모색하고자 한다. “ChatGPT는 번역 분야에서 어느 정도의 품질과 정확성을 제공하는가?”, “BLEU 평가지표를 적용한 번역 품질평가 결과는 어떻게 해석해야 하는가?”, “번역 품질평가용 테스팅 데이터는 어떠한 기준으로 선정되어야 하는가?”, 그리고 “번역 품질평가용 테스팅 데이터의 품질 향상 방안과 보다 적극적인 활용방법은 무엇인가?”

2. ChatGPT 번역 품질평가

2.1 선행연구

현재까지 진행된 ChatGPT의 번역 품질평가 관련 선행연구 사례는 매우 제한적이다. 국내 연구로는 우선 문학 장르와 사용자의 인식 및 번역 가능성에 대한 연구 사례가 있다. 전자와 관련하여 ChatGPT 3.5를 활용한 김소월 시의 번역결과물을 중심으로 AI 번역의 오류 양상을 분석한 연구(이유정 2023)와 ChatGPT가 문학 장르의 아이러니 번역의 활용 가능성(박수정과 최은실 2023)을 고찰한 연구가 있다. 그리고 후자와 관련해서는 학부 번역 전공자의 ChatGPT 관련 인식과 ChatGPT 번역 및 포스트에디팅에 관한 실험 연구(지윤주, 이상빈, 이선우 2023) 사례가 있다.

그리고 해외 연구로는 BLEU 및 CHRF와 같은 자동 평가지표를 사용한 ChatGPT의 번역 품질평가 및 ChatGPT 번역이 외국어 교육 부문에서 유발하는 기회와 과제(Geng and Jian 2023)에 관한 논의가 있다. 또한 루돌프 외(Rudolph et al. 2023) 및 방 외(Bang et al. 2023)는 중국어, 영어, 독일어, 루마니아어와 관련된 번역 작업에서 ChatGPT의 번역 품질이 Google Translate, DeepL 및 기

타 범용의 번역 엔진과 비교했을 때 서로 비슷하다는 결과를 제시한 바 있다. 그리고 ChatGPT 기반의 중국어와 영어 쌍의 번역 품질을 한자어 단어를 사례로 비교 분석한 연구(Wu 2023) 논문이 있다. 이와 같이 지금까지 언급한 선행 연구는 문학번역, 외국어 교육, 인공지능 기계번역과의 다국어 번역 비교, 그리고 한자어 단어 사례 중심의 번역 품질 비교 등으로 ChatGPT의 번역 품질에 관한 평가를 시도한 점에서 의미가 있다. 하지만 번역 품질평가에 적용하는 테스팅 데이터 선정과 그 기준에 관한 객관적인 논의는 거의 찾아볼 수 없다. 때문에 해당 선행 연구사례는 ChatGPT의 번역 품질에 관한 논의를 미시적인 범주에만 국한하는 한계가 있다.

한편 ChatGPT의 번역 품질을 LLM의 규모 측면에서 접근한 관심을 끄는 연구가 있다. 앞서 언급한 바와 같이 ChatGPT와 같은 LLM(대형 언어 모델)은 대규모의 학습용 데이터 리소스를 적용한 고자원(high-resource) 언어의 경우에는 번역 또한 탁월한 성능을 보이는 모델이다(Jiao et al. 2023). 하지만 학습용 데이터 리소스가 제한적인 저자원(low-resource) 언어에 대한 적용성은 여전히 의문이며, 특히 저자원 언어로의 번역평가를 위해 연구용으로 구축한 M2M100 모델[5]의 번역 결과가 ChatGPT보다 양호하다(Stap and Ali 2023: 164)고 하였다. 이 연구에서 시사하는 바는 ChatGPT의 번역 성능은 앞서 언급한 바와 같이 ChatGPT에 적용된 학습용 데이터의 불균형성을 반드시 염두에 두어야 한다는 점이다. 따라서 한국어와 영어처럼 저자원과 고자원의 언어 쌍을 기준으로 번역평가를 수행하기 위해서는 두 언어 사이의 학습용 데이터의 불균형성을 최소화할 수 있는 번역 품질 테스트용 데이트 세트를 활용할 필요가 있다.

이런 점에서 본 연구는 빅테크 기업들이 개발하는 각종 생성형 AI 플랫폼과 번역 엔진의 번역 품질평가를 위해서는 평가의 주체가 한국어를 모국어로 사용하는 연구자나 기관이 수행하는 것은 물론이며 반드시 한국어와 영어 기반의 테스팅 데이터 세트를 적극 활용할 것을 제안한다.

5) 아델라니 외(Adelani et al. 2022)에서 적용한 스페인어에서 원주민 언어로의 번역을 위해 다국어 M2M100 모델(Fan et al. 2021)을 미세 조정하여 적용한 모델이다(Stap and Ali 2023: 164).

2.2 번역 품질평가 과정 설계

번역 품질평가는 다음과 같은 과정으로 이루어진다. 먼저 본 연구진이 참여한 (한영) ‘AI 학습용 말뭉치 구축사업’에서 수행한 최종결과물 데이터 중 무작위로 추출한 5개 분야의 상위 300개 세그먼트의 데이터를 ChatGPT 번역 품질 검증을 위한 테스트용 데이터 세트로 활용한다. 그리고 이 테스팅 데이터에 포함되어 있는 MT와 MTPE 외에도 비교 대상 데이터인 ChatGPT 번역 결과물을 추가하여 번역 품질평가 대상 데이터를 구축한다. 이에 대한 객관적인 지표를 측정하기 위하여 BLEU(Bilingual Evaluation Understudy) 스코어를 사용하여 계량적인 수치를 도출한다. 이를 기반으로 테스팅 데이터와 비교 대상 ChatGPT 번역과의 상관관계를 중심으로 통계치의 의미를 해석하고 각각의 번역 특성을 진단한다. 종합적으로 ChatGPT의 번역 품질 수준에 관하여 논의하며 생성형 AI를 포함한 번역 서비스를 제공하는 다양한 플랫폼의 번역 품질검증 및 향상을 위한 ‘AI 학습용 말뭉치’의 테스팅 데이터로서의 활용 가능성을 논의한다.

2.3 번역 품질평가용 테스팅 데이터 구축

ChatGPT의 번역 품질평가용 테스팅 데이터는 구축 목적과 과정, 그리고 품질평가 등의 측면에서 타당성과 신뢰성, 그리고 규모의 적절성을 포함한 객관적인 유효성을 확보해야 한다. 이에 관하여 다음과 같이 항목별로 보다 자세히 논의하기로 한다.

2.3.1 테스팅 데이터의 타당성

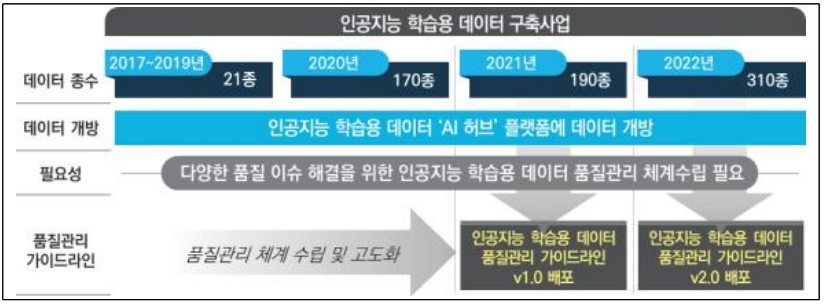
본 연구에서 적용하는 번역 품질 테스팅 데이터는 2017년부터 2023년 현재까지 과학기술정보통신부(이하 ‘과기부’)와 한국지능정보사회진흥원(National Information Society Agency, 이하 ‘NIA’)이 주관하는 한국형 디지털 뉴딜의 대표 사업인 데이터 댐 프로젝트[6]의 일환으로 수행된 AI 학습용 데이터 구축 사

6) 데이터 수집 가공 결합 거래 활용을 통해 데이터경제를 가속화하고 5G 전국망에 기반해 모든 산업으로 5G와 인공지능(AI) 융합서비스를 확산하려는 사업이다.

업[7]의 결과물과 관련된다.

AI 학습용 데이터[8] 구축사업은 양질의 인공지능 학습용 데이터를 대규모로 구축하여 중소기업 및 스타트업 등 민간의 인공지능 기술개발 촉진 및 관련 산업을 육성하고, 일자리 창출 등 민간 참여 기반의 인공지능·데이터 선순환 생태계를 조성하고자 하는 목적으로 진행되고 있다. 그동안의 추진경과를 살펴보면 2017년부터 2022년까지 총 691종의 AI 학습용 데이터의 구축이 완료되었으며, 품질검증, 개발자·전문가 등을 대상으로 사전공개를 통한 오류 및 유효성 검수 과정을 거친 데이터를 AI Hub에 개방하고 있다. AI 학습용 데이터 구축사업에 대한 전체적인 이해를 돕고자 연도별 데이터 구축 종수, 데이터 개방, 품질관리 체계 수립의 필요성, 그리고 품질관리 가이드라인 등을 포함한 품질관리체계 수립 추진배경에 관한 내용을 소개하면 다음의 <그림 1>과 같다(과기부 외 2023a: 1).

〈그림 1〉 AI 학습용 데이터 품질 관리체계 수립 추진 배경(과기부 외 2023a: 1)



위의 그림에서 보다시피 해당 사업은 양질의 데이터를 구축하기 위하여 「인

7) 데이터 댐의 가장 기초이자 핵심으로 AI 스피커, 자율주행차, 정밀의료 등 AI 서비스 개발에 필수적인 AI 학습용 데이터를 대규모로 구축 개방(aihub.or.kr)하는 사업이다.

8) 인공지능 학습용 데이터(AI Dataset)란 머신러닝, 딥러닝 등 인공지능 모델 학습을 위해 활용되는 데이터를 총칭한다. 특히 '인공지능 학습용 데이터 구축' 사업의 경우 비정형 또는 반정형 데이터를 수집하고 참값(GT, Ground Truth) 어노테이션을 통해 라벨링하여, 지도학습(Supervised Learning)에 쓰이는 데이터를 구축하는 데 초점을 둔다. 즉, AI Dataset = AI Training Data + Validation Data + Test Data이다(과기부 외 2023a: 5).

공지능 학습용 데이터 품질관리 가이드라인」에 제시된 요건에 부합하는 엄정한 품질관리 시스템을 가동하고 있다. 그리고 이러한 시스템을 통하여 다음과 같이 데이터 품질의 고도화를 지향하고 있다. 단계별 '품질관리 프레임워크'의 준비 계획, 구축, 운용활용의 3단계 프로세스를 기반으로 공통된 공정에 따라 체계적으로 각각의 산출물 및 품질관리 활동이 다음의 <그림 2>와 같이 이루어진다(과기부 외 2023a: 2).

〈그림 2〉 프로세스 기반 산출물 및 품질관리 활동(과기부 외 2023a:2)

이처럼 인공지능 학습용 데이터 품질관리 체계 수립 추진 배경과 프로세스 기반 산출물 및 품질관리 활동에 관한 내용을 통하여 본 연구에서 활용하고자 하는 테스팅 데이터는 인공지능 학습용으로 구축된 한영 병렬 말뭉치 데이터로서 사용목적에 부합하는 합목적성을 지니고 있음을 강조하고자 한다. 그리고 특히 코퍼스의 구성 측면에서도 한국어 ST, 영어 TT(MT를 바탕으로 수행된 인간번역사의 MTPE 버전)를 포함하고 있는 병렬 말뭉치는 다른 범용의 MT는 물론 생성형 AI의 번역 품질 테스트용으로 활용할 수 있는 효용가치가 매우 크기 때문이다. 왜냐하면 인간 번역사의 MTPE 버전을 기준으로 테스팅 목적에 따라 상호 비교 검증할 수 있는 대상 데이터를 가감하여 탄력적으로 활용할 수 있도록 기본 세팅이 되어 있기 때문이다. 이런 점에서 해당 말뭉치 데이터는 번역 품질 테스팅 데이터로서 활용할 수 있는 충분한 목적성과 타당성을 지닌다고 할 수 있다.

2.3.2 테스팅 데이터의 신뢰성

본 연구에서 활용할 테스팅 데이터는 구체적으로 2022년 인공지능 학습용 데이터 구축 지원 사업의 ‘AI 허브 데이터 활용을 위한 기계 번역앱 구축과 번역기 평가’ 부문 중 본 연구팀이 참여한 ‘한영 신규 말뭉치 데이터 구축 프로젝트’(이하 ‘한영 말뭉치 구축 프로젝트’)의 결과물을 대상으로 한다. 해당 말뭉치는 명칭에서 제시하는 바와 같이 한국형 디지털 뉴딜의 대표 사업인 데이터 댐 프로젝트[9]의 일환으로 AI 학습용으로 구축된 명시적인 목적성을 지닌 데이터이다.

그리고 무엇보다도 해당 데이터는 이미 ST와 기계번역 기반의 TT, 즉 MT를 기반으로 인간 번역사가 수행한 MTPE 결과물의 번역 품질에 관하여 공식적인 인증평가를 받았다는 사실이 매우 중요하다. 다시 말해서, 한국정보통신기술협회(TTA) 산하 AI융합시험연구소가 시행한 ‘인공지능 학습용 데이터 품질검증’ 기준을 통과한 데이터로서 객관적인 신뢰성을 확보하고 있다는 점이다. 검증항목은 통계적인 다양성(한영 데이터의 어절 수, 길이, 카테고리, NER 테깅 분포), 요건의 다양성(도메인 분포, 서브 도메인 분포), 구문의 정확성, 의미의 정확성, 그리고 유효성 등 5가지로 구성되어 있다. ‘한-영 번역 정확도 검사’에서 1,071,228개(33-1. 번역 말뭉치 데이터 고도화 866,415개 문장, 33-3. 신규 말뭉치 데이터 204,813개 문장)의 데이터 중 1,400개의 데이터를 샘플링하여 검증하였으며, 6가지의 검증항목 모두 기준 스코어 이상을 충족하여 품질검증 인증을 받았다. 품질검증 항목별 구체적인 내용 및 기준은 다음의 <그림 3>과 같다.

9) 데이터 수집 가공 결합 거래 활용을 통해 데이터경제를 가속화하고 5G 전국망을 기반으로 모든 산업분야로 5G와 인공지능(AI) 융합서비스를 확산하려는 사업이다.

〈그림 3〉 인공지능 학습용 데이터 품질검증 항목 및 주요 내용(과기부 외 2023a: 67)

● 품질검증항목

번호	품질지표	주요 내용	기준
1	다양성(통계)	• 관심객체, 카테고리, 수집환경 등 인공지능 학습용 데이터의 주요특성을 통계적 방법으로 확인	충분성, 균등성, 편향성 여부 확인
2	다양성(요건)	• 사업수행기관의 구축목표 대비 구축결과물 비율(수량)이 충족하는지 통계적 방법으로 확인	구축목표 대비 구축결과물 비율(수량) 기준 설정
3	구문 정확성	• 라벨 데이터 포맷과 값이 정확하게 입력되어 있고 필수항목 누락 여부 검사	정확도 99.5% 이상 권고
4	의미 정확성	• 어노테이션 값이 의미적으로 정답(GT)인지 확인하는 항목으로 실제적인 정답 비율 확인	정확도 95% 이상 권고
5	유효성	• 구축한 데이터세트를 잘 알려진 인공지능 학습 모델로 훈련시킨 후 목표로 했던 수준의 성능 달성이 가능한지 확인	인공지능 학습 모델 별 적정 기준 설정

뿐만 아니라 2023년 최근에 AI허브 데이터를 활용한 연구에서 외부의 혹은 자체 개발한 다른 종류의 테스팅 세트를 적용하여 말뭉치 데이터의 품질에 대한 검증이 이루어진 바 있다(박찬준, 임희석 2020; 고원희, 최진혁, 최규동 2023). 두 연구 모두 검증 대상 말뭉치는 다르지만 공통적으로 AI Hub에 공개된 병렬 말뭉치 데이터에 대한 품질 재검증 절차를 통하여 해당 데이터의 품질에 대한 보다 객관적인 신뢰성을 확보했다는 사실이 중요하다. 물론 본 연구에서 활용하는 말뭉치 데이터를 검증 대상 데이터로 삼은 것은 아니지만 'AI 학습용 말뭉치' 데이터의 품질은 물론 품질 검증 시스템에 대한 공신력도 재확인하였다고 할 수 있다. 이런 점에서 본 연구에서 'AI 학습용 말뭉치'를 테스팅 데이터로 활용할 수 있는 근거 및 타당성 확보는 물론 신뢰성 측면에서도 보다 객관적인 검증을 받았다고 할 수 있다.

2.3.3 테스트용 데이터 규모의 적절성

2023년 10월 31일 현재 본 연구팀이 참여한 '기계번역 평가 프로젝트'[10]를 포함한 '2022년 인공지능 학습용 데이터(319종)'가 AI Hub에 추가로 개방되었

10) AI Hub > 데이터 찾기> 분야별 보기 > 한국어 > AI 허브 데이터 활용을 위한 기계 번역앱 구축과 번역기.

다.[11] 이는 데이터의 적극적인 활용성 확장을 위한 검증과정을 모두 통과한 데이터라는 의미를 함유한다. 본 연구에서 적용할 한영 말뭉치 구축 프로젝트의 데이터 개요는 다음과 같다.

〈표 1〉 AI 허브 데이터 활용을 위한 기계번역 앱 구축과 번역기 평가 데이터 개요[12]

항목별	지표	규모
구축량	문장 수 또는 단어	AIHUB TM 구축 870,022문장
		용어 사전 10,000건
		NER 데이터 10,000건
		신규 말뭉치 데이터[13] (한-영, 한-일, 한-중) 635,110문장
		번역기 평가 데이터 600,000문장
		유사 문장 데이터 729,910문장
		MTPE 시험 데이터 101,673문장
주제분포	비율	15개 세부 분야별 데이터 분포 확인
문장길이 분포	수량 (어절 수)	평균 15어절 (최소 2어절~최대 30어절)

위의 표에 제시한 63만여 규모의 **'신규 말뭉치 데이터'** 중에서 본 연구팀은 '한-영 말뭉치 데이터' 구축 분야에서 10만 세그먼트 규모의 '한영 병렬 코퍼스의 MTPE 및 검수 과정'을 수행하였다. 그중에서 본 연구를 위하여 무작위로 선택한 농학, 수학, 의약학, 사회과학, 그리고 환경을 포함한 5개 분야에서 각각 상위 300개의 세그먼트를 수합한 1,500개의 세그먼트로 번역 품질 평가용 데이터를 구축한다. 해당 말뭉치 데이터는 한국어 ST, 영어 MT, 그리고 MTPE를 포함하여 3개 열(column)의 레이블링으로 구성되어 있다[14]. 이를 테스팅 데이

11) AI Hub > 공지사항 > 2022년 인공지능 학습용 데이터 정식 개방 안내; 본 연구에서 적용한 데이터 부문은 2023년 현재 부분 개방 중이다. 따라서 TTA '인공지능 학습용 데이터 품질검증' 평가 시 제출한 최종 결과물을 활용하였음을 밝혀둔다.

12) AI Hub > 데이터 찾기> 분야별 보기 > 한국어 > AI 허브 데이터 활용을 위한 기계 번역앱 구축과 번역기 > 데이터 통계.

13) 본 연구팀의 프로젝트 참여 부문: 한-영(재료, 정보/통신, 생명과학, 농림수산 식품, 보건의료, 수학, 물리학, 화학, 사회 등 10개 영역) 10만 세그먼트에 해당한다.

14) AI Hub에서 공개하는 AI 학습용 말뭉치 데이터의 최종결과물에는 명시되지 않지만 본 연구에서 범용의 MT와 ChatGPT를 비교하는 TQA를 수행하기 위하여 MTPE 작업 시 사전 참고용으로 구축한 중간결과물인 MT_Basic을 포함하였음을 밝혀둔다.

터로 활용하여 ChatGPT의 번역 품질을 평가하기 위하여 ChatGPT 레이블링을 하나 더 추가하여 번역 결과물[15]을 포함한 4개의 행으로 칼럼을 구성한다. 즉, 각 세그먼트는 ST와 TT(MT_Basic, MTPE, 그리고 ChatGPT)로 구성하였으며, 각각의 번역 결과물 간의 직접적인 비교 및 대조가 가능하다. 그리고 보다 정확한 번역 품질평가를 가능하게 하는 고품질의 병렬코퍼스는 번역평가 테스팅 데이터로서 원천 텍스트 수집부터 MTPE 최종 결과물 구축에 이르는 전체 과정에서 「인공지능 학습용 데이터 품질관리 가이드라인」의 준수 및 인증을 포함하고 있으므로 신뢰성 측면에서 더욱 중요성을 지닌다. 참고로 구축된 테스트용 데이터는 다음과 같이 구성되어 있다.

〈그림 4〉 'AI 학습용 말뭉치'를 활용한 ChatGPT 번역 품질평가용 테스팅 데이터 예시

AI 학습용 말뭉치 데이터[16]			생성형 AI 데이터
ST	TT		
	MT_Basic	MTPE	ChatGPT
참가투별과 나방등의 조사 주이는 1년차는 하리구입 66주, 2년차는 54주씩 조사하였다.	The number of weeks of investigation for powdery mildew and moths was 66 weeks per treatment group in the first year and 54 weeks in the second year.	The incidence of powdery mildew and moths was investigated at 66 weeks in the first year of treatment and 54 weeks in the second year of treatment.	The frequency of investigations for powdery mildew and moth species was 66 weeks per treatment group in the first year and 54 weeks in the second year.
참가투별 여명률은 1년차 여명주율, 2년차 여명과율을 조사하였으며 나방류는 피해주율을 조사하였다.	The powdery mildew morbidity rate was investigated in the first year and the second year morbidity rate, and for moths, the damage rate was investigated.	The morbidity rate of powdery mildew was investigated as the morbidity rate in the 1st and 2nd years, and the damage morbidity was investigated in moths.	The incidence rate of powdery mildew was investigated in the first and second year, and for moths, the rate of damage was investigated.
굽 강도의 경우 40 MPa 시험체에서 이러한 경향이 나타나지 않고 페로니켈 슬래그 혼입률의 증가할수록 강도가 저하되는 경향을 보였다.	In the case of bending strength, this trend did not appear in the 40 MPa test specimen, and the strength tended to decrease as the ferronickel slag mixing ratio increased.	In the case of flexural strength, this tendency did not appear in the 40 MPa specimen, and the strength decreased as the ferronickel slag mixing ratio increased.	the 40 MPa test specimen, and the strength tended to decrease as the ferronickel slag incorporation ratio increased.
휨 강도 시험은 페로니켈 슬래그 잔골재 치환량에 따라서 제작하였고, 시험체는 1,00 × 1,00 × 400 mm 사각기둥으로 제작하여 KS F 2408의 규정에 따라서 실시하였다.	The flexural strength test was produced according to the amount of ferronickel slag fine aggregate added, and the test specimen was manufactured as a 1,00 × 1,00 × 400 mm square column and conducted in accordance with the regulations of KS F 2408.	The flexural strength test was made according to the amount of fine aggregate added in ferronickel slag, and the test piece was manufactured in a square column of 1,00× 1,00×400 mm and carried out according to KS F 2408.	The flexural strength test was conducted based on the amount of ferronickel slag fine aggregate added, and the test specimen was created as a 1,00 × 1,00 × 400 mm square column, in accordance with the regulations of KS F 2408.
휨 강도 시험 결과 페로니켈 슬래그 잔골재를 혼입한 24 MPa, 30 MPa에 대해서 25%를 혼입한 시험체의 경우 압축 강도 시험과 같이 상대적으로 높은 강도를 나타내었다.	As a result of the bending strength test, the test specimen containing 24 MPa and 30 MPa of ferronickel slag fine aggregate mixed with 25% showed relatively high strength, as shown in the compressive strength test.	As a result of the flexural strength test, the specimen containing 25% of 24 MPa and 30 MPa of fine ferronickel slag showed relatively high strength as in the compressive strength test.	The bending strength test results showed that the test specimen, which comprised 25% ferronickel slag fine aggregate mixed with 24 MPa and 30 MPa, exhibited a relatively high strength, akin to the results of the compressive strength test.
회수율 시험은 두 가지 수준으로 농도를 선정하여 모처리 각 나물 시료 10 g에 각각 1 μg/mL의 표준 용액을 주입하여 최종 농도 0.1 및 0.5 mg/kg이 되도록 처리하여 3번씩 실험을 수행하였다.	For the recovery rate test, two levels of concentration were selected, and 1 μg/mL standard solution was injected into 10 g of untreated blood sprout sample, respectively, to achieve final concentrations of 0.1 and 0.5 mg/kg, and 3 repeated experiments were performed.	In the recovery rate test, the concentration is selected at two levels, and the standard solution of 1 μg/mL is injected into 10 g of the untreated ginseng sample. And the final concentration was 0.1 and 0.5 mg/kg, and three repeated experiments were performed.	For the recovery rate test, two levels of concentration were chosen, and a 1 μg/mL standard solution was injected into each 10 g sample of untreated blood sprout, aiming for final concentrations of 0.1 and 0.5 mg/kg, and the experiment was carried out three times.
회분식 반응기는 하나의 반응기에서 지속적인 흐름없이 배출없이 일정 시간 처리 후 배출시키는 시스템으로 하나의 반응기에서 여러 공정을 진행시킬 수 있는 장점이 있어 소규모 처리에 적합하다.	A batch reactor is a system in which treatment is performed for a certain period of time without continuous flow or discharge from a single reactor and then discharged. It has the advantage of being able to proceed with multiple processes in a single reactor, making it suitable for small-	Batch reactor is a system that discharges after treatment for a certain period of time without continuous flow or discharge from one reactor. It has the advantage of being able to proceed with several processes in one reactor, so it is suitable for small-scale treatment.	A batch reactor is a system that allows treatment for a specific period without continuous flow or discharge from a single reactor, before expelling the treated material. It has the advantage of facilitating multiple processes within one reactor, making it ideal for small-scale processing.

15) ChatGPT 번역 결과물 추출 일자는 2023년 11월 13일이다.

16) 혼선을 방지하기 위하여 "MT_Basic은 말뭉치 구축과정에서 MTPE를 수행하기 위하여 적용된 기초 자료이며, 최종결과물에는 명시되지 않았음"을 다시 한 번 밝혀둔다. 그리고 범용의 MT와 ChatGPT와 비교하는 TQA를 수행하기 위하여 MTPE를 기준으로 좌우로 MT_Basic과 ChatGPT를 병렬 배치하였으며, 이는 작업 수행의 순서와도 관련이 있다.

2.3 번역 품질평가 툴

AI Hub의 'AI 학습용 말뭉치'를 테스팅 데이터로 활용하여 ChatGPT의 번역 품질을 평가할 때 BLEU 스코어를 적용하였다. BLEU 스코어는 인간번역과 기계번역의 유사성을 살펴볼 수 있는 대표적인 지표 중 하나이다. BLEU 스코어는 주로 N-gram 일치를 기반으로 기계 번역의 결과물이 사람이 수행한 번역과 유사한 정도를 측정한다. 대규모의 데이터에 대한 빠른 평가가 가능하며, 연구 및 산업 분야에서 널리 사용되는 툴인 반면에 문맥적인 뉘앙스나 유창성을 평가하는 데는 한계가 있다. 그리고 때로는 오역을 잘못된 정답으로 간주하기도 한다. 그럼에도 본 연구에서 BLEU 스코어를 적용하는 이유는 「AI 학습용 말뭉치 데이터 품질 관리 가이드라인」에도 제시된 평가 툴(과기부 외 2023a: 118-120)이며, 번역학계 전반에서 가장 널리 사용되는 툴인 점을 감안하였기 때문이다. 그리고 이와 같이 유사한 평가 툴을 사용하면 관련 논의를 할 때 비교평가 군으로 활용할 수 있다는 판단도 작용하였기 때문이다.

3. 번역 품질평가 결과

3.1 BLEU 스코어를 활용한 분석

본 연구에서 활용할 번역 품질평가용 테스팅 데이터는 한영 병렬 코퍼스로 두 개의 언어 자료가 포함되어 있다. <그림 4>에서 제시한 바와 같이 가로축[17]은 한국어 원문(ST), TT인 기계번역(MT_Basic), 기계번역 포스트 에디팅(MTPE), 그리고 ChatGPT를 포함하여 4개의 열로, 그리고 세로축은 세그먼트 행으로 구성되어 있다. 가로 축을 다시 요약하면 ST 1열, TT는 3열(MT_Basic vs MTPE vs ChatGPT)로 구성되어 있다. 여기서 언급해야 할 중요한 사항은 MT는 코퍼스 구축 초기 단계에서 범용의 기계번역 엔진을 활용하여 ST를 기반으로 추출한 TT이며, ST와 TT의 미스매칭, 누락, 생략, 첨가, 구두점 등 가시

17) 가로축에 분야(field) 및 세그먼트 번호(Num.) 열이 포함되어 있어나 본 연구에서는 논의의 범주에 포함하지 않는다.

적으로 드러나는 번역 오류에 대한 소극적인 포스트에디팅 과정이 추가되었다는 점이다. 다시 말해서 번역 품질평가용 테스팅 데이터에 적용한 MT의 원시 추출물에 사전 공정 과정이 추가된 결과물임을 미리 밝혀둔다. 따라서 평가 시 혼선을 피하기 위하여 해당 MT는 MT_Basic으로 레이블링 하였다[18]. 다만 이후 논의 과정에서 표현하는 MT_Basic을 연구자들이 또 다른 종류의 번역 엔진 혹은 플랫폼으로 오인 혹은 혼동하지 않도록 각별한 유의가 필요하다는 점을 미리 밝혀둔다.

3.1.1 평균 및 누적 N-gram BLEU 스코어[19]

이제 본격적인 분석을 시작해 보기로 한다. BLEU 툴을 적용할 때 N-gram 및 누적(cumulative) N-gram 스코어 추출 방식은 효과적인 번역 품질평가 방법이다. BLEU 스코어는 기본적으로 비교 대상 번역(MT_Basic과 ChatGPT)과 기준이 되는 참조 번역(MTPE) 간의 N-gram 일치도를 기준으로 평가한다. 평균(average) N-gram BLEU 스코어는 MTPE와 비교 대상인 MT_Basic과 ChatGPT의 N-gram(N 단어 시퀀스)의 정밀도를 측정한다. 예를 들어 1-gram BLEU 스코어는 각각의 단어를 확인하는 반면, 4-gram BLEU 스코어는 4개 단어 시퀀스의 일치 여부를 평가한다. 그리고 누적 N-gram은 1-gram, 2-gram 등을 개별적으로 보는 대신 1-gram, 2-gram, 3-gram, 4-gram 일치 항목의 조합을 동시에 고려하여 누적 BLEU 스코어를 계산한 것이다. 이 접근 방식을 사용하면 개별 단어의 정확성과 긴 구문의 유창성에 대해 보다 균형 잡힌 통찰력을 가질 수 있

18) MT_Basic의 사전 공정 작업, 즉 light PE가 추가된 사항과 관련하여 범용의 MT로 활용할 수 있는가에 대한 논의의 여지가 발생할 수 있다. 이러한 점을 인식하여 Levenshtein Distance 방식을 적용하여 가공되지 않은 MT와 최소한의 사전 작업이 이루어진 MT_Basic에 대한 MTPE와의 편집 거리(editing distance)를 각각 측정하였다. 그 결과 MT와 MT_Basic의 평균 Levenshtein Distance는 각각 약 65.71과 65.67로 추출되었다. 0.04의 차이는 MT와 MT_Basic의 편집 거리가 매우 유사하므로 참조번역인 MTPE와 비교할 때 비슷한 수준의 편집이 요구됨을 나타낸다. 따라서 본고에서 적용하는 최소한의 light PE가 추가된 MT_Basic을 범용의 MT로 상정하여 TQA의 비교 대상 테스팅 데이터로 활용하는 데 대한 논의의 여지는 충분히 상쇄되었다고 할 수 있다.

19) 본 연구의 계량적인 분석은 ChatGPT의 Data Analysis 툴을 활용하였다.

다. 다음은 MTPE를 기준으로 분석한 평균 및 누적 N-gram BLEU 스코어 측정값이다.

〈표 2〉 MT_Basic과 ChatGPT의 N-gram 기반 BLEU 스코어

Score Type		MT Basic	ChatGPT
Average	1-gram	0.655	0.559
	2-gram	0.549	0.426
	3-gram	0.466	0.331
	4-gram	0.399	0.258
Cumulative	1-gram	0.678	0.579
	2-gram	0.571	0.444
	3-gram	0.488	0.348
	4-gram	0.421	0.276

위의 <표 2>는 다양한 N-gram 수준에 걸쳐 평균 및 누적 BLEU 스코어를 모두 요약하여 제시하고 있으며, 범용의 번역 엔진 MT-Basic과 생성형 AI ChatGPT의 번역 품질에 대한 포괄적인 개요를 알 수 있다. 스코어는 MT_Basic이 일반적으로 모든 수준에서 ChatGPT에 비해 BLEU 스코어가 더 높게 나왔다. 각각의 측정값은 다음과 같이 해석할 수 있다.

- 1-gram: MTPE에 비해 MT_Basic 및 ChatGPT는 모두 상대적으로 높은 수준의 단어 대 단어 정확도를 나타낸다. MT_Basic의 스코어가 더 높으며, 이는 단어 수준에서 MTPE와 부합도가 높다는 것을 의미한다.
- 2-gram: MT_Basic 및 ChatGPT 모두 1-gram 스코어보다 감소하였다. 두 단어 조합을 일치시키는 것이 좀 더 어렵기 때문에 이렇게 감소된 결과치가 도출되었을 것이다. MT_Basic은 계속 더 높은 스코어를 얻었는데, 이는 ChatGPT에 비해 두 단어 시퀀스의 정확도가 더 높다는 의미이다.
- 3-gram과 4-gram: 앞의 두 평가기준보다 더 긴 단어 시퀀스를 정확하게 일치시키는 난이도가 반영되어 더 낮은 스코어가 나왔다. MT_Basic 및 ChatGPT 모두 낮은 스코어를 나타내지만, MT_Basic은 ChatGPT 대비 더 높은 스코어를 유지하여 MTPE와의 부합도가 좀 더 높다는 의미이다.

지금까지 N-gram 기반의 BLEU 스코어를 통하여 비교 대상 MT_Basic와

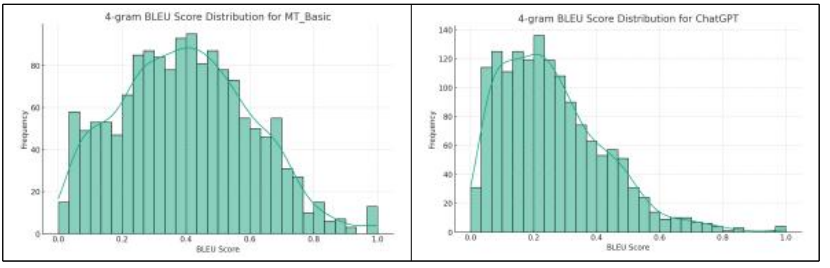
ChatGPT의 번역 품질에 대하여 포괄적인 의미를 살펴보았다. 이를 기반으로 세부적인 특성을 파악하고자 4-gram 기반으로 논의를 이어가기로 한다.

3.1.2 4-gram 기반의 BLEU 스코어

3.1.2.1 전체 세그먼트 분석

위의 <표 2>에서 제시한 BLEU 스코어를 기반으로 논의한 비교 대상 MT_Basic와 ChatGPT의 번역 품질을 단어 차원보다는 시퀀스 위주의 번역 품질평가를 위하여 4-gram 위주로 보다 구체적으로 접근해 보기로 한다. 우선 세그먼트별 번역 품질평가도 중요하므로 전체 세그먼트의 빈도를 반영한 BLEU 스코어를 위의 <표 2>에서 제시한 스코어 중 4-gram 위주로 시각화하면 다음과 같다.

〈그림 5〉 MT_Basic과 ChatGPT의 4-gram BLEU 스코어 분포

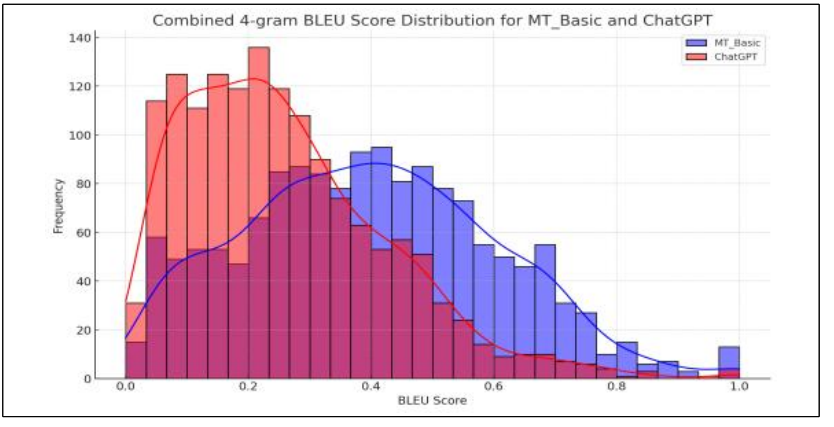


먼저 위의 <그림 5>의 히스토그램에서 가로축은 테스팅 데이터의 참조번역 MTPE를 기준으로 MT_Basic과 ChatGPT의 세그먼트 부합도를 반영한 4-gram BLEU 스코어의 값을 표시하며, 세로축은 이에 해당하는 세그먼트의 빈도(frequency)를 표시한다. 가로축의 BLEU 스코어가 1.0에 가까울수록 MTPE 참조번역과의 유사도가 높다는 것을 의미한다. 그리고 세로축의 빈도가 높을수록 해당 BLEU 스코어의 데이터가 차지하는 분량이 많다는 의미이다. BLEU 스코어 1.0에 가까운 구간에서 빈도의 밀집도가 높으면 참조번역과의 유사도가 높으므로 번역 품질이 높다는 의미이다. 뿐만 아니라 그래프의 분포에서 우편향 혹은 좌편향이 심할수록 번역 품질의 가변성이 높다는 의미로 해석할 수 있다.

반대로 BLEU 스코어 0.0에 가까운 구간에서 빈도의 밀집도가 높으면 참조번역과의 유사도가 낮으므로 번역 품질이 낮다는 의미이다. 이를 종합하면 가로축은 번역 품질의 우수성을 세로축은 번역 품질의 가변성을 나타낸다. 빈도의 높낮이 격차가 클수록 번역 품질의 가변성이 높은 것이므로 분포도가 우편향에 가까울수록 번역의 품질이 좋은 것이다. 한편, 특정 BLEU 스코어 구간대의 밀집도는 참조번역 MTPE와 해당 세그먼트 사이의 유사성의 정도를 반영하므로 특정 세그먼트(들)의 번역 특성을 판단할 때 유효하다.

위에서 살펴본 바와 같이 4-gram BLEU 스코어를 기반으로 데이터를 시각화하면 데이터 세트 전체에서 번역 품질의 일관성과 범위를 이해하는 데 도움이 된다. 뿐만 아니라 번역 품질평가와 관련하여 스코어의 분포와 분산에 대한 통찰력을 제공하므로 3.1.1의 평균 및 누적 BLEU 스코어 분석에서 다루지 못하는 영역을 보완하는 효과도 있다. 그러므로 이번에는 다시 참조번역 MTPE 기준 MT_Basic과 ChatGPT의 번역 품질의 차이를 보다 명시적으로 살펴보고자 <그림 5>의 두 그래프를 하나로 통합하여 논의를 계속 이어가기로 한다.

〈그림 6〉 MT_Basic과 ChatGPT의 4-gram BLEU 스코어 통합 분포



<그림 6>은 <그림 5>를 통합하여 MT_Basic(보라색) 및 ChatGPT(주황색)에 대한 4-gram BLEU 스코어 분포를 중심으로 시각화한 그래프이다. 이 통합된 그래프는 BLEU 스코어 분포 측면에서 비교 대상 MT_Basic과 ChatGPT의

번역 품질의 특성을 보다 직접적으로 비교할 수 있는 통찰력을 제공한다. 가령, 분포의 중복 범위(보라색과 주황색이 겹쳐서 자주색으로 나타나는 부분)와 차이나는 범위(보라색과 주황색이 그대로 유지되는 부분)가 시각적으로 명백하게 드러나는데, 이는 MT-Basic과 ChatGPT가 각각 상대적인 일관성과 변동성이 있음을 보여준다. 각 분포의 최고점과 분포의 범위는 스코어의 중심성과 변동성에 대한 명확한 통찰력을 제공한다. 이를 통해서 테스팅 데이터 전체에서 MT_Basic과 ChatGPT의 상대적인 번역 품질을 평가할 수 있다.

특히 ChatGPT의 분포는 MT_Basic과 비교하여 기울기의 크기와 모양에서 좌편향 쏠림현상이 나타나는데 이는 번역 품질의 신뢰성과도 상관관계가 있다. 또한 세로축의 빈도의 높낮이 또한 MT_Basic과 비교하여 편차가 크게 나타나는데 이는 본 연구에서 적용한 번역 품질평가 툴인 BLEU 스코어와 테스팅 데이터의 신뢰도 측면에서 일관성과 타당성이 있음을 의미한다. 한편, 히스토그램의 분포에서 MT_Basic은 ChatGPT의 분포와 비교하여 기울기의 경사도 및 모양이 완만하고 넓게 펼쳐져 있다. 이는 MT_Basic의 번역 품질이 ChatGPT에 비하여 편차가 적고 균일성이 높다는 의미이다. 하지만 빈도 측면에서 MT_Basic이 ChatGPT에 비하여 BLEU 스코어의 중간대에 주로 분포되어 있는데 이는 보다 적극적인 해석이 필요하다. 왜냐하면 사용자 입장에서 MT_Basic과 ChatGPT를 활용한 번역 경험을 통하여 일종의 선입견 혹은 선호도에 영향을 미칠 수 있기 때문이다. 이에 관해서는 3.2절의 'MT_Basic과 ChatGPT의 번역 품질평가 결과 분석' 부문에서 보다 자세히 논의하기로 한다.

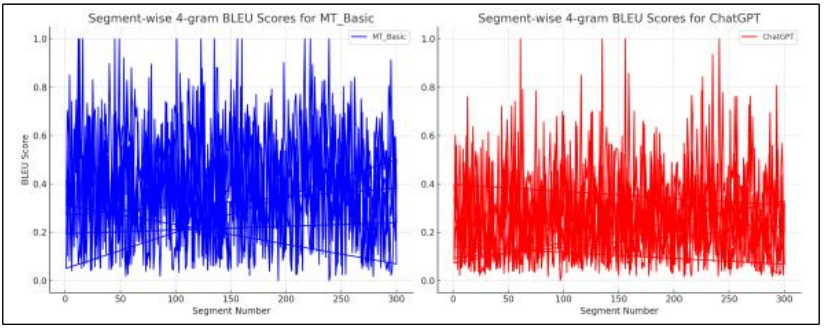
다시 히스토그램 분석으로 돌아가서 공통부분(공집합)에서 벗어나는 차집합에 해당하는 부분(보라색과 주황색이 그대로 유지되는 부분)을 살펴보면 우편향과 좌편향의 양상이 보다 명시적으로 드러난다. 빈도를 의미하는 세로축의 값 또한 MT_Basic에 비하여 ChatGPT의 편차 값이 크게 드러나는데 이는 ChatGPT가 세그먼트별 번역 품질의 편차가 보다 크다는 의미이다. <그림 6>은 MT_Basic과 ChatGPT의 번역 품질의 우수성과 신뢰성 측면에서 매우 명시적인 차이를 효과적으로 제시하며 비교집단 간의 번역 품질에 대한 전체적인 통찰력을 제공하는 점에서 본 연구에서 시사하는 바가 매우 크다.

3.1.2.2 개별 세그먼트 분석

개별 세그먼트 또는 문장에 대한 BLEU 스코어를 분석하면 데이터 세트의 여러 부분에서 번역 품질의 변화를 확인할 수 있다. 이는 번역 품질이 세그먼트별로 크게 다른 경우 특히 두드러진 양상을 드러내기 때문이다. 다시 말해서, BLEU 스코어에 대한 개별 세그먼트 분석은 테스팅 데이터의 각 세그먼트에 대한 BLEU 스코어를 살펴보고 패턴 또는 이상값을 식별할 수 있으므로 개별 문장 또는 세그먼트 수준에서 번역 품질의 가변성과 일관성에 대한 통찰력을 제공할 수 있다.

세그먼트별 분석을 하려면 다음의 세 가지 작업을 수행해야 하는데, 먼저 세그먼트별 BLEU 스코어 시각화 작업을 통해 각 세그먼트의 BLEU 스코어를 표시하는 MT_Basic 및 ChatGPT에 대한 플롯을 만든다. 이 분석은 세그먼트별로 스코어가 어떻게 다른지 확인하는 데 도움이 된다. 둘째, 특이점 식별 작업을 위하여 BLEU 스코어의 편차가 특히 높거나 낮은 세그먼트를 찾는데, 이는 번역의 품질이 매우 좋거나 그렇지 않음을 판단하는 데 도움이 된다. 셋째, 번역 품질의 일관성을 파악하기 위해 스코어 변동성 분석 작업을 수행하여, 세그먼트 전반에 걸쳐 스코어의 범위와 분포를 평가할 수 있다. 이를 기반으로 보다 포괄적인 번역 품질을 개관하고자 4-gram 스코어를 중심으로 MT_Basic 및 ChatGPT에 대한 세그먼트별 BLEU 스코어를 시각화하면 다음과 같다.

〈그림 7〉 MT-Basic과 ChatGPT의 개별 세그먼트 BLEU 스코어 분포



MT_Basic(파란색) 및 ChatGPT(빨간색) 번역에 대한 세그먼트별 4-gram

BLEU 스코어를 시각화한 위의 그림에서 각 플롯에는 테스팅 데이터 전체의 개별 세그먼트에 대한 BLEU 스코어가 표시되었다. 가로축은 테스팅 데이터의 세그먼트 번호이며 세로축은 해당 세그먼트에 대한 BLEU 스코어이다. 좌우 양측의 MT_Basic과 ChatGPT의 플롯 그래프는 개별 세그먼트에 대한 BLEU 스코어의 변동성과 추세를 표시한다. 그래프의 높이가 높을수록, 즉 BLEU 스코어가 높을수록 번역 품질이 양호하며, 반대로 높이가 낮을수록 번역 품질이 낮다는 의미이다.

전체적으로는 MT_Basic은 ChatGPT 대비 BLEU 스코어가 상단 위주로 분포되어 있으며, ChatGPT는 하단에 집중되어 있다. 때문에 파란색과 붉은색의 격차를 시각적으로 확인할 수 있다. 앞의 <그림 5>와 <그림 6>과 마찬가지로 개별 세그먼트 분석을 적용해도 BLEU 통계치의 일관성이 있음을 알 수 있다. 이와 같이 다각적인 분석을 적용했을 때에도 동일한 번역 품질평가 결과를 제시하는 점은 번역 품질의 균질성과 가변성, 평가 툴과 테스팅 데이터의 신뢰도와 타당성 측면에서 유효성이 있음을 다시 한 번 입증해 준다고 할 수 있다. 그리고 분석 결과의 활용성 측면에서 개별 세그먼트별 추세와 변동성이 제시하는 번역의 일관성과 품질에 대한 통찰력을 기반으로 세그먼트 전반에 걸쳐 번역 품질의 가변성과 일관성, 그리고 편차가 드러나는 특정 세그먼트를 찾을 때에도 용이하다. 그러므로 이를 통해서 해당 세그먼트의 번역을 보다 자세히 직관할 수 있으므로 대량의 데이터에서 번역 품질을 관리할 때에도 유용하게 활용할 수 있다.

3.2 MT_Basic과 ChatGPT의 번역 품질평가 결과 분석

3.2.1 BLEU 스코어 통계 기반

앞의 3.1절 분석에서 도출한 MT_Basic과 ChatGPT 번역 품질평가 결과 비교 대상 간의 BLEU 스코어의 차이가 있음을 발견하였다. 이에 관하여 본 장에서는 앞서 논의한 내용을 바탕으로 번역 품질평가 결과를 우선 다음과 같이 요약할 수 있다.

- N-gram BLEU 스코어의 일관성: 우선 MT_Basic은 전반적으로 ChatGPT에 비해 모든 N-gram 수준에서 더 높은 BLEU 스코어가 추출되었다. 이는

MT_Basic이 개별 단어 선택과 긴 구문 구조 측면에서 참조 MTPE 번역과 일치도가 높음을 의미한다. ChatGPT의 BLEU 스코어는 MT_Basic에 비해 낮지만 여전히 상당한 수준의 번역 정확도를 나타낸다. 특히 N-gram이 높을수록 스코어가 낮게 나타났다. 이처럼 비교 대상 간의 BLEU 스코어 차이가 발생한 원인은 MTPE 번역과 비교하여 MT_Basic과 ChatGPT 모두 단어 수준 보다는 문장 구조나 구문 선택 영역에서 차이가 날 수 있기 때문이다.

- 4-gram 스코어(유창성)의 일관성: N-gram 분석에서 MT_Basic과 ChatGPT 모두 N-gram 크기가 증가함에 따라 BLEU 스코어가 감소하는 경향이 있다. 이 패턴은 N-gram 스코어가 높을수록 보다 긴 단어 시퀀스에 초점을 맞춰 일치도를 분석하기 때문이다. 1-gram 스코어는 개별 단어의 정확성에 중점을 두는 반면, 4-gram 스코어는 긴 구문의 유창성과 일관성을 반영한다. 그리고 누적 BLEU 스코어는 1-gram에서 4-gram 일치의 조합을 고려하여 번역 품질에 대한 전체적인 개관을 할 수 있다. MT_Basic은 누적 스코어에서 지속적으로 ChatGPT 번역을 능가하여 개별 N-gram 분석에서 볼 수 있는 추세에 대한 해석을 강화하는 효과가 있다. BLEU 스코어 분포, 특히 4-gram 분석을 통해 번역 품질의 일관성과 범위에 대한 통찰력을 얻을 수 있다. MT_Basic은 일반적으로 더 집중된 분포를 보여주며, 이는 보다 일관된 번역 품질을 나타낸다. ChatGPT 번역은 분포 범위가 좀 더 넓은데 이는 변동성이 크다는 의미이다.
- 세그먼트별 BLEU 스코어의 일관성: 세그먼트별 분석은 데이터 세트의 여러 부분에 걸쳐 번역 품질의 가변성을 보여준다. MT_Basic 및 ChatGPT 모두 세그먼트별로 BLEU 스코어의 변동성을 보여주며, 이는 일부 문장이 다른 문장보다 더 효과적으로 번역되었음을 나타낸다. 이러한 변동성은 ST의 복잡성, 특정 언어 구조를 처리하는 번역 모델의 효율성 또는 영역별 용어로 인해 발생할 수 있다.

이를 종합하면, MT_Basic은 BLEU 스코어 지표 측면에서 참조번역인 MTPE와 일치도 및 상관성이 높게 나타난다. 그러나 BLEU 스코어는 번역 품질평가의 모든 측면을 반영하지 않는 한계가 있다. 특히 의미의 정확성, 문체

또는 가독성과 같은 측면을 제대로 포착하지 못하기 때문이다. 그럼에도 분석 결과는 비교 대상 번역 MT_Basic과 ChatGPT의 강점 및 약점을 파악하는데도 도움이 된다. 가령, MT_Basic은 단어 및 구문 사용 측면에서 MTPE와 부합도가 높지만 ChatGPT는 특정 상황에서 보다 자연스럽게 번역하거나 관용적인 표현이나 언어의 경우 보다 나은 번역을 제공할 수 있다. 결국 이러한 통계는 번역 품질에 대한 정량적인 통찰력을 제공하지만, 정성적인 평가와 함께 번역 작업의 특정 맥락과 요구사항을 고려하여 해석할 필요가 있다.

3.2.2 MT와 생성형 AI 특성 기반

이어서 ChatGPT 번역과 MT_Basic 간에 관찰된 BLEU 스코어 차이가 발생한 원인을 다음과 같이 MT와 생성형 AI 특성을 기반으로 관련된 측면을 고려하면 다음과 같다.

- 번역 방법론의 차이: LLM을 기반으로 하는 ChatGPT는 기존의 기계번역 시스템과 비교하여 다른 번역 방법론을 사용한다. 광범위한 분야의 대량의 데이터를 기반으로 사전 학습을 통하여 텍스트를 이해하고 생성하는 성능을 강화한다. 때문에 MT-Basic에 적용된 번역 엔진과 비교하여 특정 언어의 뉘앙스를 처리하는 방법에 차이가 있을 수 있다.
- 훈련용 데이터의 품질 및 유형: ChatGPT와 같은 생성형 AI 모델의 성능은 훈련된 데이터의 다양성과 품질에 크게 좌우된다. ChatGPT의 훈련용 데이터에 다양한 언어 스타일과 컨텍스트가 포함된 경우 해당 번역은 보다 포괄적인 이해를 반영할 수 있지만 특정 유형의 텍스트에 최적화될 수 있는 MT-Basic의 기계번역 엔진과 비교하면 정확성 측면에서 차이가 발생할 수 있다.
- 문맥 및 뉘앙스 처리: ChatGPT의 번역은 고급 언어 이해 기능을 통해 문맥, 관용적 표현 및 문화적 뉘앙스를 더 잘 이해할 수 있다. 그러나 이것이 항상 표층적인 층위에서 텍스트의 유사성을 주로 측정하는 BLEU 스코어에 충분히 반영되는 것은 아니다.
- 모델별 최적화 및 목표: ChatGPT 및 MT-Basic에서 적용한 학습용 데이터는 각기 목표 및 최적화 기준이 다를 수 있다. MT-Basic은 번역의 직접적인 정확성을 위해 최적화될 수 있지만 ChatGPT에서 적용하는 LLM은 유창

성, 일관성 및 문맥적 적절성을 목표로 할 수 있으며 잠재적으로 번역이 위의 분석에서 적용한 참조번역, 즉 인간 번역사가 수행한 MTPE와의 부합도에서 차이가 발생할 수 있다.

- 고유한 제한 사항: 기계번역 플랫폼은 각각 고유한 제한사항이 있다. 특히 ChatGPT와 같은 생성형 AI는 LLM을 기반으로 훈련된 딥러닝 모델을 사용하여 새로운 콘텐츠를 생성하는 일종의 인공지능 기술이므로 번역에 특화된 플랫폼이 아니라는 명제를 전제해야 한다. 따라서 ChatGPT는 고급 기능에도 불구하고 도메인별 텍스트나 희소/소수 언어 구성으로 인해 번역과 관련해서는 여전히 어려움을 겪을 수 있다. 다시 말해서 LLM에 동원된 리소스의 규모에 따라 고자원과 저자원 언어군을 활용하는 상황에서 번역 품질의 차이를 발생시킬 수 있다. 앞서 선행연구에서 언급한 바와 같이 저자원 언어로의 번역평가를 위해 연구용으로 구축한 M2M100 모델(스페인어에서 원주민 언어로의 번역을 위해 구축한 다국어 번역 모델)의 번역 결과가 ChatGPT보다 양호한 경우도 있다(Stap and Ali 2023: 164). 하지만 본 연구에서 적용한 한국어는 규모 면에서 스페인어 원주민 언어와 비교할 수 있는 언어군이 아니므로 본 연구 결과와 매칭하는 것은 다소 무리가 있을 것 같다. 따라서 다시 고유한 제한사항과 관련하여 논의를 정리하면, MT-Basic은 번역에 특화된 범용의 기계번역 플랫폼/엔진이므로 특정 번역 작업에서 보다 일관성 있는 번역 품질 서비스를 제공할 수 있으며, 특히 특정 언어 쌍에 맞추어 조율하는 경우 더욱 번역 품질이 높게 나타날 수 있다.
- 기술의 진화 및 업데이트: ChatGPT는 생성형 AI의 발전 속도에 비례하여 업데이트 및 성능 향상 작업이 지속적으로 빈번하게 이루어지고 있다. 때문에 번역 성능도 함께 진화할 수 있다. 이와 대조적으로 일부 기계번역 플랫폼이나 엔진은 기계번역의 성능 개선 작업에 소극적인 경우 현재 상태에서 정체될 수 있다. 이는 번역 품질의 차이를 발생시키는 원인이 될 수 있다.
- 사용자 경험: BLEU 스코어의 차이는 기계번역 플랫폼이나 엔진을 사용하는 사용자의 경험(UX, user experience)도 반영한다. 사용자의 번역 작업 시 특정한 요구 사항(텍스트의 장르, 사용목적, 문체, 언어사용역 등)에 따라 번역 플랫폼 별로 번역의 품질 결과가 달리 나타날 수 있다. 이는 ChatGPT의 경우도 예외가 아니다. 앞서 수차례 언급한 바와 같이 ChatGPT는 번역에

특화된 플랫폼이 아님에도 불구하고 이미 글로벌 사용자들은 번역 용도로 이를 적극 활용하고 있기 때문이다. 따라서 번역을 수행한 사용자의 경험 혹은 경험 공유를 통해 상황에 따라 선별적으로 번역 플랫폼을 선택하여 사용할 수 있다. 주관성을 완전히 배제할 수는 없지만 사용자의 경험 또한 번역 품질의 차이를 판단하는 주요 요소에 포함할 수 있다.

위의 논의사항을 요약하면 번역 품질평가에서 ChatGPT와 MT-Basic의 스코어 차이는 상호 간의 다양한 강점과 약점을 포괄한다. 따라서 특정 번역 요구사항에 적합한 플랫폼 선택의 중요성과 지속적으로 기계번역 기술의 향상 및 연구가 이루어질 필요가 있다.

3.2.3 생성형 AI 번역 품질에 대한 심층 분석

서론에서 언급했던 ChatGPT를 사용하는 주요 목적에 번역이 포함되어 있는데, BLEU 스코어 측정 결과는 이와 반대로 범용의 기계번역 MT_Basic이 더 높게 나타나는 일관성을 보였다. 이는 통계의 오류인지 아니면 사용자의 오해인지 좀 더 다른 관점에서 진단해 볼 필요가 있다. 유추 가능한 원인을 차례대로 제시하면 다음과 같다.

- MT의 프리에디팅(pre-editing) 여부: 먼저 번역 품질평가용 테스팅 데이트 구축 과정에 대하여 살펴볼 필요가 있다. 한영 말뭉치 구축 작업 당시 범용의 기계번역 엔진에 ST를 입력하면 번역 누락, 추가, 미스매칭, 구두점 인식 오류, 동음이의어의 번역 오류 등이 발견된다. 때문에 말뭉치의 품질관리 차원에서 1차 MT에 대한 소극적인 MTPE를 거친 수정된 MT 버전을 적용하여 코퍼스를 구축하였다. 때문에 ChatGPT와 MT_Basic의 BLEU 스코어의 차이가 날 수 있다.
- 착시효과: MT_Basic과 ChatGPT의 번역 출력 결과물은 가시적으로 기능어(function word) 대비 내용어(content word) 번역의 차이가 드러나는 정도가 다르다. 앞의 말뭉치 구축 과정에서 언급한 바와 같이 범용의 MT는 다양한 구두점 인식 오류부터 내용 누락이나 추가 등 번역 결과물의 품질에 대한 즉각적인 판단이 가능한 사례가 좀 더 가시적으로 빈번하게 드러난다. 하지만 ChatGPT의 번역 결과물은 최소한 기능어 중심의 오류는 거의 찾아 볼

수 없으며, 자연스러운 어법과 유창성을 바탕으로 완성도가 높아 보이는 가시적인 효과가 있다. 때문에 ChatGPT 사용자는 출력된 번역 결과물에 1차적인 만족감을 얻을 수밖에 없는 것이다.

- 번역 품질의 균질성: ChatGPT는 어마어마한 규모의 LLM 기반의 훈련을 통해 문장에서 다음에 오는 단어나 표현을 예측하는 수행 능력이 탁월하기 때문에 대부분의 도메인에서 번역 품질이 균등한 편이다. 그리고 번역 시 자연스러운 표현과 어법을 구사하기 때문에 참조번역 MTPE를 기준으로 계량적으로 측정하는 BLEU 스코어에는 이러한 측면이 충분히 반영되지 않는다. 때문에 사용자 입장에서는 BLEU 스코어와는 반대로 ChatGPT 번역의 품질이 높다고 생각할 수 있다. 하지만 범용의 MT는 번역에 특화되어 있기 때문에 도메인에 따라서 번역 품질의 균질성이 달라질 수 있다. 이러한 측면은 사용자의 경험과도 직결되므로 개인별 선입견 및 선호도에 영향을 미칠 수 있다. 때문에 번역 사용자들은 범용의 MT보다 ChatGPT의 번역 품질을 높이 평가할 수 있다.
- 유창성의 차이: 실질적인 번역 품질의 차이 정도와 관련해서는 ChatGPT와 범용의 MT의 번역 결과물은 대체로 문장과 표현의 완결성, 그리고 자연스러운 어법을 적용한 유창성 측면에서 격차가 있으며 이는 사용자의 호감도나 선호도에 영향을 미칠 수 있다. 때문에 사용자가 판단하는 번역 품질은 실질적인 번역의 품질과는 다소 차이가 있을 수 있으므로 이 또한 고려대상에 포함해야 한다.

이외에도 다양한 원인을 유추해 볼 수 있으나 종합적으로 판단하면 ChatGPT는 일반적인 측면에서 논의되는 소위 환각현상(hallucination) 이슈가 번역에서도 발생할 수 있다는 점이다. 다시 말해서 ChatGPT는 실질적인 번역의 품질과 무관하게 상대적으로 가시적인 오류가 드러나지 않는 번역 결과물을 출력하므로 사용자에게 일종의 착시 혹은 환각 효과를 가져 올 수 있다. 그러므로 번역 플랫폼을 사용하는 입장에서 이러한 착시나 환각 효과를 벗어나 보다 나은 번역 결과물을 얻고자 한다면 생성형 AI 플랫폼과 범용의 번역 엔진 혹은 플랫폼을 동시에 구동하며 크로스체크(cross check) 방식으로 사용하는 방법을 고려할 수 있다. 가령, 먼저 ChatGPT에서 ST에 대한 번역을 수행한 1차 결과

물을 토대로 범용의 MT에서 재가공하거나 반대로 범용의 MT 출력물을 기반으로 ChatGPT를 활용하여 수정 버전을 출력하는 방식으로 상호 보완하여 활용하는 것이 좋을 것이다.

3.3 테스팅 데이터 및 평가방법의 유효성

지금까지 'AI 학습용 말뭉치'를 활용한 과정 및 결과를 기반으로 범용의 MT와 ChatGPT와 같은 생성형 AI 언어 모델의 번역 품질평가용 테스팅 데이터로서 유효성이 있는지 논의하였다. 따라서 본 연구에서 활용한 테스팅 데이터는 다음과 같은 측면에서 유효성을 지닌다고 할 수 있다.

- 병렬 코퍼스 구조: 해당 말뭉치는 번역 품질평가에 필수적인 한국어와 영어의 ST와 TT 병렬 세그먼트 구조로 구축되어 있다.
- 품질평가 기준 제공: 번역평가 시 기준으로 삼는 참조 번역 데이터가 반드시 필요한데, 테스팅 데이터에 포함된 MTPE는 고품질의 참조 표준을 제공한다. 이를 통해 MT 및 ChatGPT 결과물을 MTPE와 직접 비교할 수 있다. 나아가 MTPE는 번역 품질을 객관적으로 평가하는 데 매우 중요한 요소이며, 번역 서비스의 성능 개선을 위한 지표로 활용될 수 있다.
- 다양한 도메인의 콘텐츠: 해당 테스팅 데이터는 다양한 분야와 세그먼트를 포함하고 있어, 번역 플랫폼에 제공하는 번역 서비스의 성능을 다각적으로 평가할 수 있는 기반이 된다. 각 분야별 특성과 어휘, 문법적 구조를 고려한 테스트는 번역 서비스의 다양한 사용 사례에 대한 이해를 돕는다.
- 정량적인 분석 가능: 위의 번역 품질평가에서 측정한 바와 같이 해당 말뭉치 데이터는 BLEU 스코어와 같은 측정항목을 사용하여 정량적인 평가 결과를 도출하였다. 이는 비교대상 번역 플랫폼은 물론 여타의 생성형 AI 번역 결과물을 MTPE 참조 번역과 비교한 일치도에 대한 객관적인 데이터를 제공한다. BLEU N-gram 분석을 통한 정량적인 분석 결과 제공은 각 번역의 품질을 구체적이고 정량적으로 평가할 수 있는 근거를 마련한다. 이는 번역 서비스의 성능을 비교하고 분석하는 데 매우 유용하다.
- 세그먼트별 분석: 본 연구에서 활용한 테스팅 데이터는 세그먼트별 분석에서도 효용적인 평가를 수행하였으며, 이를 통해 비교 대상 번역 시스템의

특정 강점과 약점을 식별하는 데 도움이 되었다. 그리고 각 시스템의 장점과 취약점을 보다 세밀하게 이해할 수 있도록 활용되었다.

- 기계번역 엔진/플랫폼별 성능 비교: 본 테스팅 데이터를 활용하여 범용의 MT와 ChatGPT 간의 번역 품질을 비교하였으며, 이는 다른 종류의 번역 플랫폼이나 엔진의 번역 품질평가에도 유용한 기준을 제공한다. 이는 물론 이번 평가를 통해 얻은 통찰력은 비교 대상 번역 플랫폼의 번역 품질을 개선하는 데 도움이 될 수 있다. 오류나 불일치가 발생하는 지점을 이해하면 이러한 시스템을 추가로 개발하고 학습시킬 수 있기 때문이다.
- 실제 적용 가능성: 해당 테스팅 데이터는 번역 플랫폼/엔진 개발자, 언어학자, 번역가 등이 활용하여 번역 품질을 평가하고 개선 방안을 모색하는 데 실질적인 도움이 될 수 있다.

하지만 테스팅 데이터로서의 유효성을 검증할 때 다음의 제한사항 또한 반드시 고려할 필요가 있다.

- 정성적 분석의 한계: BLEU와 같은 정량적 지표는 평가지표로서 가치가 있지만 문화적인 뉘앙스, 관용적 표현 또는 문체, 언어사용역 등과 같은 번역 품질의 모든 측면을 충분히 반영하지는 못한다. 따라서 인간 번역가의 정성적 평가도 중요하다.
- 도메인별 성능의 불균형성: 말뭉치 데이터가 특정 도메인으로 제한되는 경우 평가는 다른 도메인이나 텍스트 유형의 번역 평가에는 유효하지 않을 수 있다.
- 맥락 및 의미의 정확성: BLEU 및 유사 측정항목은 주로 표층적인 수준의 텍스트 유사성에 중점을 두며 맥락이나 의미의 정확성을 제대로 반영하지 못할 수 있다.
- 참조번역의 번역 품질에 대한 신뢰성: MTPE를 번역 품질의 평가기준으로 사용했으나, 이는 주관적인 평가의 영향을 받을 수 있으며, 평가자들 간의 견해의 차이가 있을 수 있다. 즉, MTPE 자체의 번역 품질 균질성에 격차가 있을 수 있다.
- 참조번역의 추출 기반에 대한 다양성: 번역 품질 평가 테스팅 데이터를 설계할 때 참조번역으로 활용하는 MTPE의 추출 기반에 대한 고려도 필요하다[20].

지금까지 논의한 내용을 요약하면, 본 연구에서 번역 품질 테스트용 데이터로 활용한 AI 학습용 병렬 말뭉치는 MT 및 ChatGPT의 번역 품질을 평가하는 데 유용한 리소스라 할 수 있다. 비교 대상 번역 플랫폼에서 제공하는 번역 품질을 평가하고 비교할 수 있는 체계적이고 포괄적인 방법을 제공하여 정량적인 통찰력과 제한적이지만 정성적인 통찰력을 모두 제공한다. 따라서 본 연구에서 적용한 'AI 학습용 말뭉치' 테스팅 데이터는 번역 품질 테스트용으로 충분한 효용 가치가 있으며, 번역 서비스의 성능 평가 및 개선을 위한 중요한 자료로 활용될 수 있다. 하지만, 분야별, 언어별 특성과 텍스트의 다양성을 보다 포괄적으로 반영하기 위한 추가 데이터 구축의 필요성도 고려해야 한다. 그리고 향후 연구에서는 다양한 참조번역을 활용한 번역 품질평가도 시도할 필요가 있다.

4. AI Hub의 '인공지능 학습용 말뭉치' 활용방안

4.1 활용사례

AI Hub의 '인공지능 학습용 말뭉치' 데이터를 활용한 학술적인 연구 성과를 파악하고자 한국학술지인용색인(KCI)에서 키워드 검색어 'AI Hub 혹은 AI 허브 데이터'를 적용하여 추출한 연구 성과는 2023년 11월 5일 현재 39건[21]으로 나타났다. 연도별 발표 건수는 2019년 1편, 2020년 3편, 2021년 3편, 2022년 13편, 2023년 11월 9일 현재 19편이다. 주제별 논문 수는 공학(27편, 69.23%), 자연과학과 사회과학 각 4편(10.25%), 복합학(3편, 7.69%), 예술체육과 인문학 각 1편(2.56%)을 차지한다.

---

20) 본 연구는 범용의 MT를 기반으로 한 MTPE 한 종을 참조번역으로 활용하였으나, ChatGPT를 기반으로 한 MTPE를 참조번역으로 활용하지는 않았다. 이 경우 번역 품질평가에 대한 해석이 달라질 수는 있다. 하지만 번역 품질평가 테스팅 데이트를 적용할 때 다양한 참조번역을 제공하는 점에 대해서는 평가 수행의 효율성과 결과물의 해석 측면에서 좀 더 논의가 필요할 것 같다.

21) KCI 키워드 검색에서 해당 국책사업 시행 이전 발표 논문(문석재 외 2008) 1건이 포함되었으나 논의의 범주를 벗어나므로 제외하였다. 따라서 실제 연구사례는 39건에 해당한다.

KCI 논문 중에서 본 연구 주제와 직접적인 상관성을 지닌 연구 성과는 다음의 두 편을 꼽을 수 있다. 첫 번째 사례는 한영 기계번역 관련 IWSLT의 테스트 세트를 활용하여 공공 한영 병렬 말뭉치를 이용한 기계번역 성능 향상에 관하여 논의한 연구(박찬준, 임희석 2020)가 있다. 이 연구에서는 공공에게 개방된 AI Hub 공개 데이터와 사용 빈도가 높은 한영 병렬 데이터 OpenSubtitles와 성능 비교를 통해 데이터의 품질을 검증하였다. 그리고 두 번째 사례는 문장 간 유사도와 문장 속성을 활용한 한영 영한 번역 병렬 말뭉치 품질 예측 모델(고원희, 최진혁, 최규동 2023)에 관한 연구가 있다. 여기서는 한영 영한 번역 병렬 말뭉치의 품질평가 모델을 제안하고, BLEU와 자체 개발한 번역 병렬 말뭉치 품질평가 모델 TwiQE를 활용하여 AI Hub에 공개된 한영 및 영한 번역 병렬 말뭉치 각 2종에 대한 품질평가를 시도하였다. 그리고 이를 바탕으로 말뭉치의 품질 예측 모델 구축에 활용하였다.

두 연구에서 주목할 점은 무엇보다도 공개된 혹은 자체 개발한 테스트 셋을 활용하여 AI Hub에 공개된 병렬 말뭉치 데이터에 대한 품질 재검증 절차를 진행했다는 점이다. 이러한 과정을 통하여 다시 한 번 NIA에서 구축한 한영 및 영한 번역 병렬 말뭉치 데이터의 품질에 대한 보다 객관적인 신뢰성을 재확인한 점이다. 이는 다시 말해서 NIA의 번역 병렬 말뭉치 데이터를 테스팅 데이터로 활용할 수 있는 가능성을 더욱 확장하였다고 할 수 있다. 이 점은 본 연구에서 NIA의 한영 병렬 말뭉치 데이터를 테스팅 데이터로 활용할 수 있는 실제 사례를 보여줌으로써 서로 다른 세 편의 논문 사이의 긴밀한 상호관련성을 찾을 수 있다.

그리고 고려대학교 자연어처리 연구실(NLP&AI Lab)에서는 AI 학습용 데이터를 활용하여 국내외 자연어 분야의 논문, 특허, 기술이전 등 다수의 연구를 수행하고 있다. 그리고 과제 발굴 수요를 조사하여 관련 기관에 '한국어 상식 추론 평가 데이터 셋'(KommonGen) 구축의 필요성을 제안하여 채택하는 방식으로 과제 수주율을 높여서 역동적으로 다양한 과제를 수행하고 있다. 이외에도 산학협력을 포함한 연구 및 과제 수행 결과는 새로운 데이터 셋 구축과 활용 가능성을 확장함으로써 데이터 및 AI 생태계의 선순환 과정을 구축하고 있다(과기부와 NIA 2022: 58-59).

그 외 다양한 활용 사례가 있지만 본고의 주제 및 논의의 범주를 벗어나므

로 여기서는 더 이상 언급하지 않기로 한다. 종합적으로 실질적인 활용사례를 살펴본 결과 AI 학습용 데이터 플랫폼, AI Hub에 개방된 다양한 분야의 데이터를 보다 적극적으로 활용할 수 있는 방안을 시급하게 모색할 필요가 있다.

4.2 활용 확대 방안

지금까지 논의한 'AI 학습용 말뭉치' 데이터의 번역 품질 테스팅 데이터로서의 유효성 검증을 기반으로 테스팅 데이터의 활용성을 확장할 수 있는 방안을 번역 품질평가, 다양한 번역 플랫폼의 고도화, 21세기 번역(학) 생태계의 재편에 대한 대비 등의 가치 제고를 목표로 다양한 측면에서 다음과 같이 고려할 수 있다.

- 다양한 언어 쌍으로 확장: 본 연구에서는 한영 번역에만 국한하였으나, 다양한 언어 쌍으로 데이터 세트를 확장하여 광범위한 언어에 대한 번역 서비스의 품질을 평가할 수 있다. 가능하다면 동일한 ST를 기반으로 다양한 언어 쌍(현재 한영, 한일, 한중 구축 완료[22])의 코퍼스를 구축할 필요가 있다. 이러한 다중 언어 확장은 다양한 언어 쌍에 대한 번역 시스템을 평가하고 개선하기 위한 보다 포괄적인 기반을 제공할 수 있기 때문이다.
- 도메인과 장르의 다양성 확장: 말뭉치 구축 대상 분야 및 주제 범위를 확장할 필요가 있다. 다양한 맥락에서 번역 품질을 평가하려면 보다 다양한 도구를 구축할 필요가 있는데, 이를 위해 광범위한 분야로 도메인을 확장하고 자연어와 구어체를 포함한 다양한 장르의 사용역을 반영한 텍스트를 포함할 필요가 있다.
- 주석 및 메타데이터 추가: 언어 주석(예: 품사 태그, 구문 구조) 및 메타데이터(예: 도메인, 텍스트 소스, 난이도)를 사용하여 말뭉치의 성능을 강화한다. 이러한 추가 기능을 활용하면 보다 상세한 언어 분석과 번역 학습용 모델의 고도화가 용이해진다.
- 인간 번역 평가 스코어 통합: BLEU와 같은 자동화된 측정항목과 함께 전문 번역가의 평가를 포함한다. 유창성, 적절성, 관용적인 어법 등에 대한 평가와 미묘한 의미의 차이를 반영하는 심층적인 번역 품질평가를 통해 세부적

22) <표 1> 참조.

인 내용까지 정밀하게 파악할 수 있는 통찰력을 제공할 수 있다.

- 번역 알고리즘 개선: 분석 결과를 바탕으로 기계 번역 알고리즘의 개선 사항을 도출하고, 이를 통해 번역의 정확성과 자연스러움, 그리고 유창성을 개선할 수 있다. 이를 위하여 말뭉치의 하위 세트를 번역 모델의 벤치마크로 표준화하여 비교 대상 번역 플랫폼의 성능 향상을 위한 표준 데이터로 활용한다. 또한 이를 정기적으로 업데이트하여 시간 경과에 따른 다양한 번역 기술의 진행 상황을 추적하고 비교할 수 있다.
- 번역 메모리 시스템에 접목: 코퍼스는 유사하거나 반복적인 텍스트 세그먼트에 대한 MTPE 참조 번역을 제공하여 인간 번역사를 지원하는 번역 메모리 시스템에 활용될 수 있다.
- 교육 및 훈련 도구로 활용: 번역가 교육 및 언어 학습 도구로 활용하여, 실제 예시와 번역의 품질평가를 통해 언어 학습자와 전문 번역가의 역량을 강화할 수 있다. 뿐만 아니라 코퍼스는 교차 언어 연구, 비교 언어학, 새로운 번역 알고리즘 개발과 같은 분야의 학술 연구를 위한 풍부한 데이터 세트로 접목할 수 있다.
- NLP 도구와의 통합: 코퍼스를 사용하여 특히 한국어 및 영어(다양한 언어쌍)에 대한 언어 모델, 감정 분석, 텍스트 분류 시스템과 같은 자연어 처리(NLP) 도구를 훈련하고 테스트할 수 있다.
- 자동화된 평가 도구 개발: 번역 품질평가를 자동화하는 소프트웨어 도구를 개발하여, 보다 빠르고 효율적으로 번역 서비스의 성능을 평가할 수 있다. 이러한 도구는 연속적인 성능 모니터링과 신속한 피드백 제공에 기여할 수 있다.
- 다문화 커뮤니케이션 지원: 다양한 언어와 문화적 배경을 가진 사용자 간의 효과적인 커뮤니케이션을 지원하기 위해, 번역 품질 테스팅 데이터를 활용하여 더 나은 번역 서비스를 제공할 수 있다.
- 지속적인 업데이트 및 품질 관리: 정기적으로 새로운 텍스트로 말뭉치를 업데이트하고 품질 검사를 수행하여 관련성과 정확성을 지속적으로 보장한다. 이를 위하여 사용자로부터의 실시간 피드백을 수집하고, 이를 기반으로 테스팅 데이터 세트를 지속적으로 업데이트하여 번역 서비스의 품질을 지속적으로 향상시킬 수 있다.

- 저자원 언어군 테스팅 데이터 적극 활용 필요: LLM 기반의 생성형 AI 플랫폼 서비스가 속출하는 상황에서 LLM에 활용되는 데이터 규모의 극단적인 차이로 인한 번역 품질 서비스의 불균형성을 인식해야 한다. 이를 보완하기 위해서는 상대적인 격차가 큰 저자원 언어 군에서 생성한 테스팅 데이터를 적극적으로 활용할 필요가 있다. 본 연구에서 적용한 AI 학습용 이중어 병렬 코퍼스와 같은 객관적인 신뢰성을 확보한 데이터를 기반으로 해당 언어권의 번역전문가들이 주도하는 번역 품질평가를 수행하면 비교 대상 번역 플랫폼의 서비스 품질을 효율적으로 개선할 수 있기 때문이다.

이처럼 ‘AI 학습용 말뭉치’ 데이터의 번역 품질 테스팅 데이터로서 활용성을 확장하려면 다양한 언어쌍의 병렬 말뭉치 구축, 도메인과 장르의 다양성 확장, 주석 및 메타데이터 추가, 인간 번역 평가 스코어 통합, 번역 알고리즘 개선, 번역 메모리 시스템에 접목, 교육 및 훈련 도구로서의 활용, NLP 도구와의 통합, 다문화 커뮤니케이션 지원, 지속적인 업데이트 및 품질관리, 데이터 규모의 극단적인 차이로 인한 번역 품질 서비스의 불균형성을 보완하기 위한 저자원 언어군의 테스팅 데이터 적극 활용 등 다양한 방안을 고려해 볼 수 있다. 이러한 확장 방안들은 토대로 번역 품질 테스팅 데이터의 활용 범위를 넓히고, 다양한 분야에서 그 가치를 극대화할 수 있는 기회가 적극적으로 확장되기를 기대한다.

5. 결론 및 제언

지금까지 AI Hub에서 공개하는 ‘AI 학습용 말뭉치’가 ChatGPT가 제공하는 번역 서비스의 품질평가용 테스트 데이터로서 활용 가능성과 유용성에 대한 검증을 시도해 보았다. 이를 통해 논의의 도입부에서 언급한 핵심 논제를 다음과 같이 요약할 수 있다. 먼저 첫 번째 논제 “ChatGPT는 번역 분야에서 어느 정도의 품질과 정확성을 갖고 있는가?”와 두 번째 논제인 “BLEU 평가지표를 적용한 번역 품질평가 결과 해석은 어떻게 해야 하는가”에 관해서 함께 논의하기로 한다. BLEU 스코어를 적용하여 번역 품질평가를 진행한 결과 비교 대상인

범용의 MT는 생성형 AI ChatGPT에 비하여 BLEU 스코어 지표 측면에서 참조번역인 MTPE와 일치도 및 상관성이 높게 나타났다. ChatGPT의 주요 사용 목적에 번역이 포함되어 있는 사실에 비추어 BLEU 스코어 측정 결과는 이와 반대로 범용의 기계번역 MT의 품질이 높게 나타났다. 물론 BLEU 스코어는 번역 품질평가의 모든 측면을 반영하지는 못하는 한계가 있다. 의미의 정확성, 문체 또는 가독성과 같은 측면을 제대로 포착하지 못하기 때문이다. 그럼에도 이와 같은 평가 결과가 나온 데는 1) 번역 품질평가용 테스팅 데이트에서 추출한 원시 말뭉치 데이터의 MT는 1차 사전공정 작업이 이루어진 버전에서 출발한 점, 2) MT-Basic와 ChatGPT의 번역 출력 결과물의 기능어 대비 내용어의 가시적인 차이, 3) 번역 품질의 균질성의 차이를 유발하는 요인, 4) 사용자의 선호도나 호감도의 격차를 유발하는 사용자 경험 요인으로 인하여 ChatGPT는 일반적인 측면에서 논의되는 이슈인 소위 환각 현상이 번역에서도 발생할 수 있다는 점이다.

세 번째 핵심 논제인 “번역 품질평가용 데이팅 데이터 선정기준과 유효성”과 관련하여 본 연구에서 테스트용 데이터로 활용한 ‘AI 학습용 말뭉치’는 목적성과 타당성, 객관적인 신뢰성을 확보한 데이터이며, 규모의 적정성을 갖춘 데이터 구축을 기준으로 선정되었다. 그리고 해당 말뭉치 데이터는 참조번역으로 활용할 수 있는 MTPE를 포함하는 병렬 코퍼스이며, 다양한 도메인의 콘텐츠를 포함하며, 정량적인 분석, 세그먼트별 분석, 기계번역 엔진/플랫폼별 성능 비교, 그리고 실제 적용 가능성 측면에서 번역 품질평가 테스팅 데이터로서 충분한 효용 가치가 있다. 그리고 번역 서비스의 성능 평가 및 개선을 위한 중요한 자료로 활용될 수 있다. 하지만, 분야별, 언어별 특성과 텍스트의 다양성을 보다 포괄적으로 반영하기 위한 추가 데이터 구축의 필요성은 고려해야 한다.

네 번째 핵심 논제, “번역 품질평가용 테스팅 데이터의 품질 향상 방안과 보다 적극적인 활용방법”과 관련하여 다양한 언어 쌍으로 구축 범위 확장, 도메인과 장르의 다양성 확장, 주석 및 메타데이터 추가, 인간 평가 스코어 통합, 번역 알고리즘 개선, 번역 메모리 시스템에 접목, 교육 및 훈련 도구로 활용, NLP 도구와의 통합, 지속적인 업데이트 및 품질관리, 저자원 언어군의 테스팅 데이터 적극 활용 등의 방안을 모색할 수 있다.

지금까지 논의한 내용을 종합하면, 본 연구는 ChatGPT와 MT의 번역 품질

을 N-gram 기반의 BLEU 스코어와 4-gram 기반의 전체 및 세그먼트별 분석 방식으로 평가를 진행하였다. 평가 결과, 비교 대상 MT_Basic와 ChatGPT 모두 참조번역 MTPE 표준과의 높은 유사성을 보였으나, N-gram 적용 범위와 세그먼트에 따라 번역 품질에 일정한 차이가 있음을 확인하였다. N-gram 분석과 4-gram 기반의 세그먼트 분석을 통해, 범용의 MT와 생성형 AI ChatGPT 번역이 MTPE를 기준과의 부합도 및 번역의 균질성과 특성 등 정량적인 측면에서 번역 품질에 대한 통찰력을 제공하였다. 마지막으로 분석 결과에서 일관적으로 제시된 MT_Basic와 ChatGPT의 BLEU 스코어의 격차가 발생하는 원인으로 특히 플랫폼 사용자의 경험에 따라 ChatGPT의 번역 품질 판단에서도 환각 이슈가 발생할 수 있다는 근원적인 이유를 제시한 점은 본 연구의 특별한 성과로 판단된다.

따라서 AI Hub에서 공개하는 한영 ‘AI 학습용 말뭉치’ 데이터를 테스팅 데이터로 활용한 본 연구는 기계번역 분야에서 번역 품질평가의 중요성을 강조하며 효과적인 평가 방법론을 제시했다. 그리고 번역 품질에 대한 심층적인 분석은 번역 서비스의 개선을 위한 중요한 기초 자료를 제공한다. 따라서 본 연구는 번역 플랫폼 서비스의 개발자, 사용자, 언어학자에게 유용한 통찰력을 제공하며, 기계 번역의 질적 향상을 추구하는 데 일조할 것이다. 그럼에도 본 연구의 제한점은 번역 품질평가 방식을 BLEU 스코어를 기반으로 분석하는 데 그쳤다는 점이다. 향후 연구에서는 보다 다양한 언어 쌍과 문화적 배경을 포함한 데이터 세트를 활용한 보다 다각적인 번역 품질 평가방식을 적용할 필요가 있다. 이를 통해 광범위한 언어적 다양성에 대한 번역 플랫폼 서비스의 적응력을 평가할 수 있을 것이다.

마지막으로 제안하고 싶은 점은 LLM에 활용되는 데이터 규모의 극단적인 차이로 인한 번역 품질 서비스의 불균형성을 보완하기 위한 대안으로 규모의 격차가 상대적으로 큰 저자원 언어 군에서 생성한 테스팅 데이터를 적극적으로 활용할 필요가 있다. 다시 말해서 빅테크 기업들이 개발하는 각종 생성형 AI 플랫폼과 번역 엔진의 번역 품질평가를 위해서는 평가의 주체가 한국어를 모국어로 사용하는 연구자나 기관이 수행하는 것은 물론이며 반드시 한국어와 영어 기반의 테스팅 데이터 세트를 적극 활용할 것을 제안한다. 이를 통하여 번역 품질 테스팅 데이터로서의 유효성과 효용성를 더욱 확장할 수 있고, 이는 또한 인

공지능과 생성형 AI 기반으로 급격하게 재편되는 21세기 번역(학) 생태계의 선순환적인 시스템 구축과도 직결되기 때문이다. 이러한 선순환적인 시스템 가동의 역동적인 추진을 위한 인프라 구축에 AI 말뭉치 구축 사업 등 국가적인 규모의 대형 프로젝트를 주관하는 과기부와 NIA의 보다 적극적이며 지속적인 지원이 필요하다. 이러한 필요성은 궁극적으로는 우리나라가 AI 세계 3위 목표[23]에 도달하고 디지털 초격차 시대의 본격적인 진입을 위한 AI 경쟁력의 가속도를 더욱 높이기 위한 중요 전략으로서의 기반을 더욱 견고하게 다지기 위함이다.

참고문헌

고원희, 최진혁, 최규동 (2023)「문장 간 유사도와 문장 속성을 활용한 한영 영한 번역 병렬 말뭉치 품질 예측 모델」,『정보과학회 컴퓨팅의 실제 논문지』29(9): 410-417.

과학기술정보통신부, NIA, TTA (2023a)「(제1권) 품질관리 가이드라인 v3.0」.

과학기술정보통신부, NIA, TTA (2023b)「(제2권) 구축 안내서 v3.0_산출물 작성 가이드」.

과학기술정보통신부, NIA, TTA (2023c)「(요약본) 품질관리가이드라인 v3.0」.

과학기술정보통신부 (2023. 4. 14.)「초거대 인공지능(AI) 경쟁력 강화 방안. 관계부처 합동」.

과학기술정보통신부 (2021. 01. 19.)「인공지능(AI) 학습용 데이터 구축·활용 고도화 방안. 관계부처 합동 및 내부자료」.

곽은주, 김동미 (2022)「인간과 기계번역의 공존을 위한 담론: 행위자 네트워크 이론을 중심으로」,『통번역교육연구』20(3): 5-32.

23) 영국의 Tortois Media가 2019년부터 The Global AI Index라는 명칭으로 국가별 역량을 평가하며, 2023년에는 62개국을 대상으로 111개 지표를 적용하였다. AI 역량을 구현(Implementation), 혁신(Innovation), 투자(Investment) 부문으로 구분하고, 이를 다시 인재, 인프라, 운영 환경, 연구, 개발, 정부 전략, 상업적 벤처 항목으로 구분하여 평가한다(KISTEP 브리프 87).

김경훈 (2020) 「인공지능 국가경쟁력의 현재와 미래」, 『Weekly Column』 2020(9), 서울시산학연협력포럼.

문석재, 정계동, 강석중, 최영근 (2008) 「저장-프로시저 기반의 비즈니스 프로세스 상호운용을 위한 XMDR Hub 프레임워크」, 『한국정보통신학회논문지』 12(12): 2207-2218.

박수정, 최은실 (2023) 「챗GPT의 아이러니 번역 활용 가능성 고찰」, 『번역학연구』 24(2): 131-160.

박찬준, 임희석 (2020) 「공공 한영 병렬 말뭉치를 이용한 기계번역 성능 향상 연구」, 『디지털융복합연구』 18(6): 271-277.

이선화 (2023) 「챗GPT를 적용한 번역수업 실천 사례 연구: 학부생 번역 과제를 중심으로」, 『번역학연구』 24(3): 351-379.

이유정 (2023) 「현대시 인공지능(AI) 번역의 오류 양상 연구: ChatGPT-3.5를 활용한 김소월 시 번역결과물을 중심으로」, 『문화와융합』 45(10): 97-110.

이창수 (2023) 「챗GPT 출현 이후 기계 번역과 인간 번역 간의 번역 문체 차이 변화 연구」, 『번역학연구』 24(3): 539-561.

전현주 (2017) 「4차 산업혁명과 한국의 번역산업 현황 및 통번역 교육의 미래」, 『통번역교육연구』 15(3): 235-261.

전현주 (2020) 「인간과 기계번역의 공존 패러다임 모색: PBL 기반의 AI 번역 툴 활용 번역수업 운영 프로세서를 중심으로」, 『통번역교육연구』 18(4): 59-96.

전현주 (2022) 「인공지능 번역플랫폼 기반 번역가의 직명 및 직무기술의 분화에 관한 연구」, 『통번역학연구』 26(1): 167-193.

지윤주, 이상빈, 이선우 (2023) 「학부번역전공자의 챗GPT 관련 인식과 챗GPT 번역 및 포스트에디팅 실험 연구」, 『통번역학연구』 27(3): 203-226.

최효은, 이준호, 이청호 (2023) 「자동화된 기계학습(AutoML)을 활용한 특허 특화 번역엔진의 영한번역 성능 평가」, 『번역학연구』 24(2): 101-130.

KISTEP 혁신정보분석센터 (2023) 「KISTEP 브리프」 87.

Adelani, David, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie

Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf ... Sam Manthalu (2022) 'A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation', in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3053-3070, Seattle: Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli and Armand Joulin (2021) 'Beyond English-centric Multilingual Machine Translation', *Journal of Machine Learning Research* 22(107): 1-48.

Bang, Yejin, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu and Pascale Fung (2023) 'A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity', *arXiv preprint arXiv:2302.04023*.

David Stap, Ali Araabi (2023) 'ChatGPT Is Not a Good Indigenous Translator', in *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, 163-167, Toronto: Association for Computational Linguistics.

Geng, Fang and Hu Jian (2023) 'A New Direction for Post-editing with Artificial Intelligence: A Study Based on ChatGPT[J]', *Foreign Languages in China* 20(3): 41-47.

Gao, Yuan, Ruili Wang and Feng Hou (2023) 'Unleashing the Power of ChatGPT for Translation: An Empirical Study', *arXiv preprint arXiv:2304.02182*.

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu

(2023) ‘Is ChatGPT a Good Translator? A Preliminary Study’, *arXiv preprint arXiv:2301.08745*.

Rudolph, Jürgen, Samson Tan and Shannon Tan (2023) ‘ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education?’, *Journal of Applied Learning and Teaching* 6(1): 342-360.

Kurzweil, Ray (2005) *The Singularity Is Near*, New York: Viking Books.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang and Zhaopeng Tu (2023) ‘Is ChatGPT a Good Translator? Yes with GPT-4 as the Engine’, *ArXiv:2301.08745 [cs]*.

Wu, Jinrui (2023) ‘A Comparative Analysis of Chinese-English Translation Quality Based on ChatGPT: A Case Study of Chinese Characteristic Words’, *Journal of Social Science Humanities and Literature* 6(5): 53-58. DOI: https://doi.org/10.53469/jsshl.2023.06(05).08

Yilmaz, Erdem Dogukan, Ivana Naumovska and Vikas A. Aggarwal (2023) ‘AI-Driven Labor Substitution: Evidence from Google Translate and ChatGPT’, *INSEAD Working Paper No. 2023/24/EFE*, 1-59.

<인터넷 자료>

삼성SDS 데이터 분석서비스팀 (2023. 4. 26.)「ChatGPT 기술분석 백서: 1부 ChatGPT란」, 2023년 11월 7일 검색.
https://www.samsungsds.com/kr/insights/chatgpt_whitepaper1.html

양정애 (2023. 4. 12.)「챗GPT 이용 경험 및 인식 조사」,『웹진 Media Issue』 9(3), 언론진흥재단. 2023년 11월 7일 검색.
https://www.kpf.or.kr/front/research/selfDetail.do?seq=595547

한국학술인용정보 (2023)「검색어_AI 허브 데이터」, 2023년 11월 5일 검색.
https://www.kci.go.kr/kciportal/po/search/poArtiSearList.kci
https://www.kci.go.kr/kciportal/po/search/poCitaReportWithArticles.kci

AI Hub (2023. 10. 31.)「2022년 인공지능 학습용 데이터 정식개방 안내」,『AI Hub』, 2023년 11월 9일 검색.
https://aihub.or.kr/aihubnewse//noticview.do?pageIndex=1&nttSn=10248&cu

rrMenu=132&topMenu=103&searchCondition=&searchKeyword=

AI Hub (2023. 7. 31.)「AI 허브 데이터 활용을 위한 기계 번역앱 구축과 번역기 평가」,『AI Hub』, 2023년 11월 9일 검색. https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71593

AI Hub (2023. 3. 7.)「공지사항」,『AI Hub』. 2023년 11월 5일 검색.「[2023.2.공개] 인공지능 학습용 데이터 품질관리 가이드라인 및 구축 안내서 v3.0 발간」,『AI Hub』, 2023년 11월 2일 검색. https://aihub.or.kr/aihubnews/qlityguidance/view.do?currMenu=131&topMenu=103&nttSn=10125

OpenAI (2022.11.30.)「Latest News: Introducing ChatGPT」,『Open AI』. 2023년 11월 5일 검색. https://openai.com/blog?page=3

KDI경제정보센터 (2020)「2025년까지 AI 학습용 데이터 1,300종 추가 구축」,『나라경제』 11, 2023년 10월 23일 검색. https://eiec.kdi.re.kr/publish/naraView.do?fcode=00002000040000100001&cidx=13071

NamePepper (2023. 11. 5.)「Number of ChatGPT Users and Key Stats (2023)」, 2023년 11월 9일 검색. https://www.namepepper.com/chatgpt-users

NIA (2021. 12. 1.)『디지털뉴딜 브리프』 1(1), 2023년 10월 23일 검색. https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=39485&bcIdx=24010&parentSeq=24010

REUTERS (2023. 2. 10.)「Microsoft Co-founder Bill Gates: ChatGPT 'Will Change Our World'」, 2023년 11월 9일 검색. https://www.reuters.com/technology/microsoft-co-founder-bill-gates-chatgpt-will-change-our-world-2023-02-10/

Tortoise Media (2021)「The Global AI Index」, 2023년 10월 20일 검색. https://www.tortoisemedia.com/intelligence/global-ai/

[Abstract]

A Study of the Usability of the AI Learning Corpus Data Provided by AI Hub: Focusing on ChatGPT's TQA

Eun-Joo Kwak, Jaehoon Noh, Mijin Park & Hyunju Chun
(Sejong University, WISE ST Global, IITA, Shinhan University)

This study attempted to verify the usability and applicability of the 'AI Learning Corpus' released by AI Hub as test data for quality assessment of the translation service provided by ChatGPT. For this purpose, the translation quality of ChatGPT and MT was evaluated using N-gram-based and segment-wise BLEU score analysis methods. As a result of the Translation Quality Assessment (TQA), the general-purpose MT, which is the subject of comparison, showed higher consistency and correlation with MTPE, a reference translation, in terms of the BLEU score index compared to the generative AI ChatGPT. In light of the fact that ChatGPT's main purpose of use includes translation, the BLEU score measurement results showed, on the contrary, that the quality of general-purpose machine translation MT was high. The fundamental reason for the gap between MT_Basic and ChatGPT's BLEU scores consistently presented in the analysis results is that hallucination issues may occur in ChatGPT's translation quality judgment, especially depending on the platform user's experience. It is judged to be a special achievement of the research.

Keywords: ChatGPT, AI Hub, AI learning corpus, TQA testing data, BLEU score, hallucination

주제어: 챗GPT, AI Hub, AI 학습용 말뭉치, 번역 품질평가용 테스팅 데이터, BLEU 스코어, 환각효과

곽은주(1저자, https://orcid.org/0009-0001-8402-1129)
세종대학교 영어영문학과 교수
ejkwak@sejong.ac.kr
관심 분야: 번역 품질평가, 생성형 AI 프롬프트 활용 인증 평가, TQA 테스팅 데이터 평가

노재훈(공동저자)
㈜와이즈에스티글로벌 CEO
jnoh@wisest.co.kr
관심 분야: 생성형 AI 플랫폼 활용 엔진 설계, 번역품질평가시스템 구축, AI 학습용말뭉치 구축

박미진(공동저자)
(사)국제통역번역협회
(IITA, International Interpretation & Translation Association) 사무총장
jinnypark116@gmail.com
관심 분야: 생성형 AI 플랫폼 활용성 검증 및 인증 평가, AI 학습용 말뭉치 구축

전현주(교신저자, https://orcid.org/0000-0003-4131-3348)
신한대학교 국제어학과 부교수
transju@shinhan.ac.kr; wisepearl33@gmail.com
관심 분야: 번역품질평가, 생성형 AI 프롬프트 QA검수 및 리라이팅 & 인증평가, ISO 인증평가

논문 투고: 2023년 11월 15일
1차 심사 완료: 2023년 12월 1일
2차 심사 완료: 2023년 12월 14일
게재 확정: 2023년 12월 18일