낮은 언어자원의 한계를 극복하여 이해도가 높은 답변이 가능하게 하는 티그리냐 질문-답변 데이터셋 및 언어 모델 구축

연구책임자 박종철 | 소속학과 전산학부 | 참여연구원 Fitsum Gaim, 양원석, 박한철 | 연구실 홈페이지 http://nlpcl.kaist.ac.kr

우리 전산학부의 Fitsum Gaim, 양원석, 박한철 (지도교수: 박종철) 연구팀이 2023년 7월 9일~13일 토론토에서 열린 ACL 2023 에서 Outstanding Paper Award를 수상했다. 연구팀의 획기적인 논문인 "Question-Answering in a Low-resourced Language: Benchmark Dataset and Models for Tigrinya"는 동아프리카 에리트레아와 에티오피아에서 사용되는 저자원 언어인 티그리냐를 다룬다. 이 연구팀은 티그리냐 언어를 위한 질문-답변 데이터셋을 세계 최초로 구축하고 티그리냐로 작성된 문서를 읽고 답할 수 있는 언어모델을 만들었다. 이 상은 자연언어처리 분야 세계 최고인 본 학술대회에 제출된 연구 중 상위 1%에게만 주어지는 의미 있는 상이다.

연구배경

자연언어처리는 최근 초거대 언어모델(Large Language Model)의 전세계적인 경쟁적 개발에 이어 ChatGPT와 같은 생성형 질의 응답 시스템 등의 비약적인 발전으로 인공지능의 최종 목표인 AGI, 즉 일반인공지능에 빠르게 다가가는 중요한 연구 분야로 간주되고 있다. 다만 이와 같은 연구의 수준은 영어권 등 언어 자원이 풍부한 경우와 달리 저자원 언어에 대해서는 상대적으로 매우 취약하여 기존의 언어 문화적 불균형이 다가오는 미래에 훨씬 더 심화될 수 있다는 심각한 문제를 가지고 있다.

연구내용

본 연구팀은 대표적인 저자원 언어권으로 알려져 있는 아프리카 언어, 특히 동아프리카 언어 중 티그리냐 언어에 대한 질문-답변 데이터셋을 세계 최초로 구축하고 티그리냐로 작성된 문서를 읽고 답할 수 있는 언어모델을 개발하였다.

기대효과

본 연구팀이 개발한 방법론, 데이터수집 방법, 주석 툴, 그리고 모델은 언어 자원이 부족한 언어들에 대한 중요한 선례로 활용될 것으로 기대된다. 특히 본 연구는 최근 심각해지고 있는 디지털 격차를 해소하기 위해 언어적으로 다양하고, 역사적으로도 혜택을 받지 못했던 커뮤니티에도 대등한 연구가 가능한 디지털 표현 방법을 제공하였다는 의미가 있다.

Reasoning Type	Example	Frequency
Synonymy	Question: ኣብ ርሳስ ዘሎ ንመጽሓፊ ዝጠቅም ማዕድን እንታይ ይበሃል? [What is the mineral in pencils useful for writing?] Context: ...ኣብ ውሽጢ ርሳስ ዝርከብ ንጽሕፈት ዝሕግዝ እምኒ ወይ ማዕድን ግራፋይት ተባሂሉ ይጽዋዕ።... [The stone or mineral inside a pencil that is used for writing is called graphite.]	35.1%
World knowledge	Question: ግራፋይት ኣበየኖም ሃገራት ብዝያዳ ይርከብ? [In which countries is Graphite found?] Context: ...ግራፋይት ኣብ መላእ ዓለም ዳርጋ ብምዕሩይ ዝርጋሐ ዝርከብ። ብዝያዳ ግን ኣብ ቻይና፡ ህንዲ፡ ሰሜን ኮርያ፡ ሜክሲኮ፡ ብራዚል፡ ቼክ ሪፓብሊክን ቱርክን ይዝውተር።... [Graphite is almost evenly distributed worldwide. But it is most common in China, India, North Korea, Mexico, Brazil, Czech Republic and Turkey.]	11.1%
Syntactic/ Morphological variation	Question: ኣብ ርሳስ ዘሎ ንመጽሓፊ ዝጠቅም ማዕድን እንታይ ይበሃል? [What is the mineral in pencils useful for writing?] Context: ...ኣብ ውሽጢ ርሳስ ዝርከብ ንጽሕፈት ዝሕግዝ እምኒ ወይ ማዕድን ግራፋይት ተባሂሉ ይጽዋዕ።... [The stone or mineral inside a pencil that is used for writing is called graphite.]	71.4%
Multi-sentence reasoning	Question: ግራፋይት ኣብ ፈኩስ ሓመድ ዝርከብ ምኽኑ እንታይ ጠባይ የሰዕበሉ? [What characteristics does graphite inherit from existing on the surface of light soil?] Context: ...ግራፋይት ካብ ካልኦት ዓይነታት ማዕድን ዝፈልዮ ነገር እንተሎ: ኣብ ተራር እምኒ ወይ ከውሒ, ኣይኮነን ዝርከብ። ኣብ ፈኩስ ሓመድ ብቑጽላታት ተጸፍጺፉ ይርከብ። በዚ ድማዩ ግራፋይት ተረኺማሽን መጭላቑን ጠባይ ዘርኢ።... [One thing that distinguishes graphite from other minerals is that it is not found on hard stone or rock. It is found laid as solid layers on the surface of light soil. That's why graphite exhibits a malleable and slippery character.]	8.5%

Dataset

웹홍보

연구성과 Outstanding Paper Award, ACL 2023 (Toronto, Canada)
Keynote Speech, EMNLP 2023 (Singapore)

연구비 지원 IITP 스타랩 사업(2018~현재, 연 3억)