Question-Answering in a low-resourced language: benchmark dataset and models for Tigrinya

Principal Investigator
Jong Cheol Park

Department
School of Computing

Co-Investigators
Fitsum Gaim, Wonsuk Yang, Hancheol Park

Hompage
http://nlpcl.kaist.ac.kr

Fitsum Gaim, Wonsuk Yang, and Hancheol Park, under the guidance of Professor Jong C. Park from the School of Computing at KAIST, received the Outstanding Paper Award at the Association for Computational Linguistics (ACL) 2023 Conference held in Toronto, Canada, from July 9th to July 14th, 2023. Their groundbreaking work, entitled "Question-Answering in a Low-resourced Language: Benchmark Dataset and Models for Tigrinya," focuses on the under-studied East-African language Tigrinya, which is spoken in the countries Eritrea and Ethiopia. The team created the first question-answering dataset for the language and developed language models that can answer questions by reading Tigrinya documents. This recognition was awarded to only the top 1% of all submissions to the conference, which is the most prestigious conference in the field.

Background

Thanks to the recent breathtaking achievements of generative question-answering systems based on large language models (such as ChatGPT) currently in active use and development all over the world, natural language processing (NLP) is considered as one of the most critical research disciplines that may allow us to achieve the holy grail of artificial intelligence: artificial general intelligence. However, the maturity levels of NLP research studies worldwide remain quite disparate: studies are severely hindered by the scarcity of resources for most languages (in stark contrast to the ample resources for English), and this is threatening to deepen the already severe linguistic and cultural imbalance in the world.

Description

The research team constructed – for the first time – question-answering datasets for Tigrinya, a language spoken in East Africa notorious for its under-resourced nature. The team also developed language models that can read documents in Tigrinya and answer related questions.

Implications

This research, which encompasses methodologies, data collection techniques, annotation tools, and language models developed by the research team, will serve as a beacon to guide related research studies focused on other under-resourced languages. In particular, the present research has remarkable significance in providing digital representations that allow communities rich in linguistic diversity yet historically underprivileged to overcome the worsening digital divide and to conduct research without concern of language-related barriers.

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 ኣበየኖት ሃገራት ብዝያዳ ይርከብ? [In which countries is Graphite found?] Context: ...ግራፋይት ኣብ መላእ ዓለም ዳርጋ ብምዕሩይ ዝርጋሐ'ዩ 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 ምዃኑ እንታይ ጠባይ የስዕብሉ? [What characteristics does graphite inherit from existing on the surface of light soil?] Context: ...ግራፋይት ካብ ካልኦት ዓይነታት ማዕድን ዝፈልዮ ነገር እንተሎ፡ ኣብ ተሪር እምኒ ወይ ከውሒ ኣይኮነን ዝርከብ። ኣብ ፈኩስ ሓመድ ብቐጻላታት ተጸፍጺፉ ይርከብ። በዚ ድማ'ዩ ግራፋይት ተረኹማሽን 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

Research outcomes
Outstanding Paper Award, ACL 2023 (Toronto, Canada)
Welcome Keynote Speech, EMNLP 2023 (Singapore)

Research funding
Government Funding: IITP Starlab Research Project (2018~present)