Improving the Mkulima Repository Content: Utilizing Theses, Dissertations, and LLMs for Agricultural Knowledge Dissemination in Kiswahili

Joseph P. Telemala
Sokoine University of Agriculture – SUA

ETD, Conference, Livingstone, Zambia.

November 5, 2024

Overview

Unlocking Knowledge Across Languages

- "For many, research is written in a language they don't speak or in terms they don't understand — an uncracked code to invaluable information."

EDT in Tanzania

- Over 95% Swahili-speakers.
- Theses and Dissertations are in English and with academic jargon.
- But small-holder farmers are not interested in TD!
    - They don't care if TD are in English or Kiswahili But if TD can't benefit end users, why produce knowledge at such level?
- We need a way to help the knowledge in TD reach end users.

SUA repositories

- So, in this study we use Sokoine University of Agriculture (SUA's) case.
  - It has an institutional repository (called SUAIRe), with TD and research articles and papers, many in agriculture or related fields.
  - It has a **Mkulima** repository – a repo specific for Swahili documents.
    - Content sources - volunteer writers.
  - TD are a forgotten source for the Mkulima repository.
    - How? Manual translation of ETDs – (almost) impossible!! or machine translation – (may be) possible

LLMs for MT

- The literature shows great success of Neural MT, especially those that use transformers (Zimerman and Wolf, 2023; Rahali and Akhloufi, 2023)
- Transformer-based models (Vaswani et al., 2017) improve translation accuracy by handling long-range dependencies, outperforming older models like RNNs and CNNs.
- There several models on the Hugging Face's Transformers library, with models like MarianMT and mBERT enabling quality translations for multiple languages, including low-resource ones like Kiswahili.

Our Objective

1. To evaluate the accuracy and fluency of Kiswahili translations of theses and dissertation abstracts generated by the LLM-based MT model through human assessment.

2. To analyze common translation errors and challenges encountered by the LLM-based MT model in translating agricultural research abstracts.

Data Sources

- 2020 – 2024 electronic theses and dissertations available in the SUAIRe whose topics are related to agriculture.
- Some pre-processing to remove some information and irrelevant metadata for translation such as names, publication year, author names, images, etc to ensure consistent formatting.

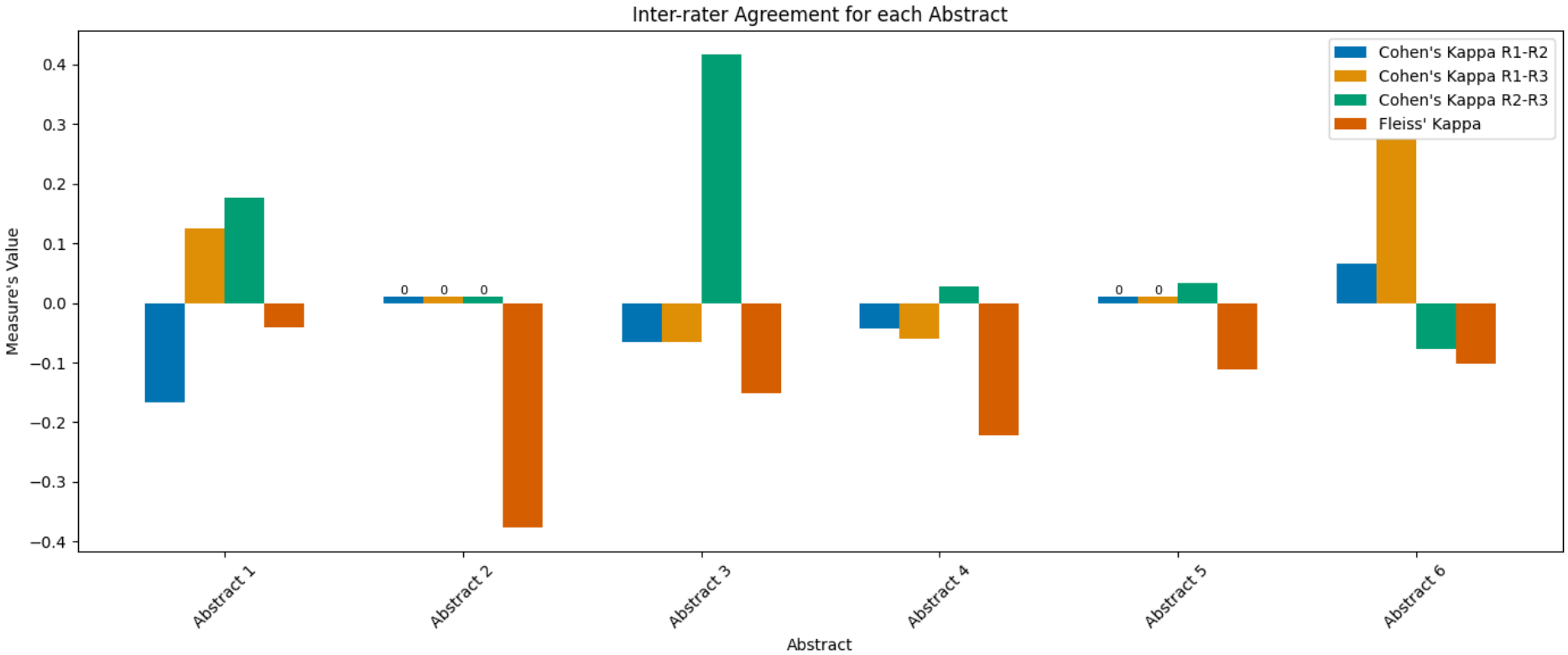
Machine Translation Model

- We used the MarianMT model (Helsinki-NLP/opus-mt-en-sw), developed by OPUS MT that is specific to the English-Kiswahili pair.

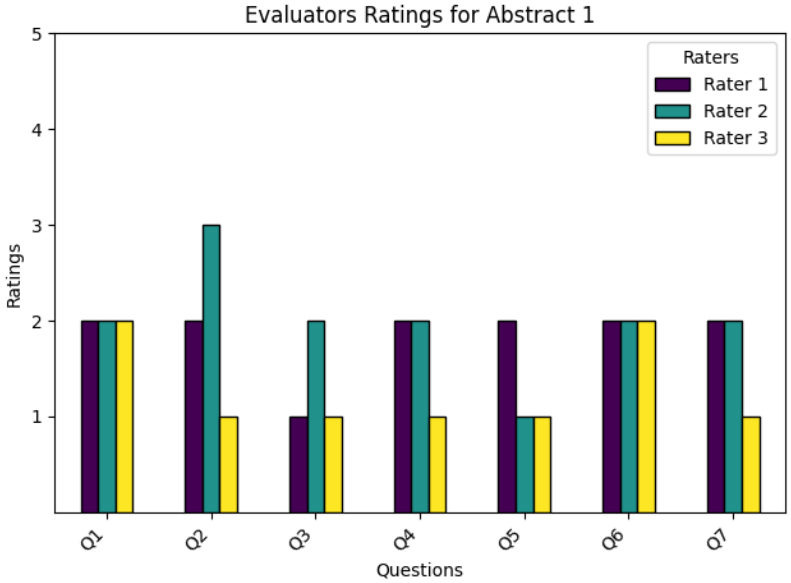
Evaluation and Error Analysis

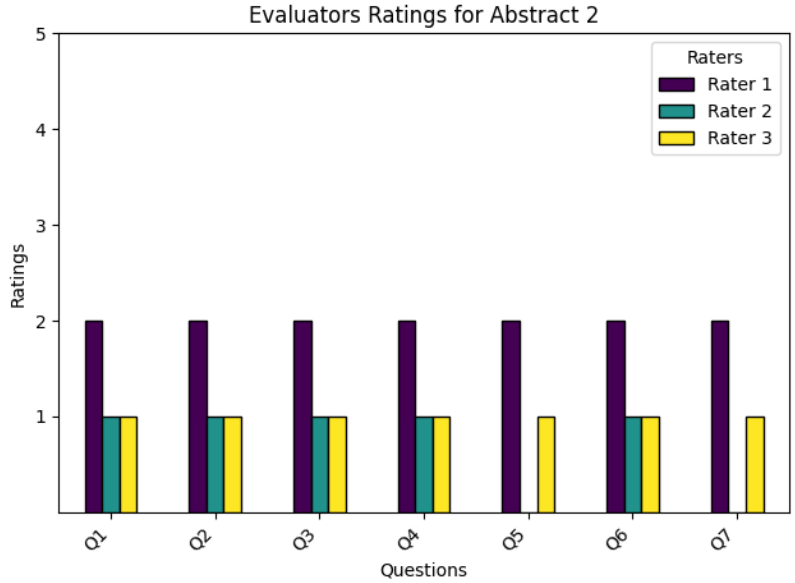
- We used human evaluators (Master's of Information and Knowledge Management students).
- Evaluators were presented with six pairs of abstracts, consisting of the original English version and its Swahili translation, one pair at a time.
- For each pair, evaluators were asked a series of Likert-scale questions designed to assess various aspects of the translation and capture specific themes related to translation quality, including comprehension, accuracy, grammar, and naturalness.
- To ensure that there was consistency and reliability of evaluators' ratings among the evaluators, we used Cohen's Kappa and Fleiss' Kappa inter-rater metrics.

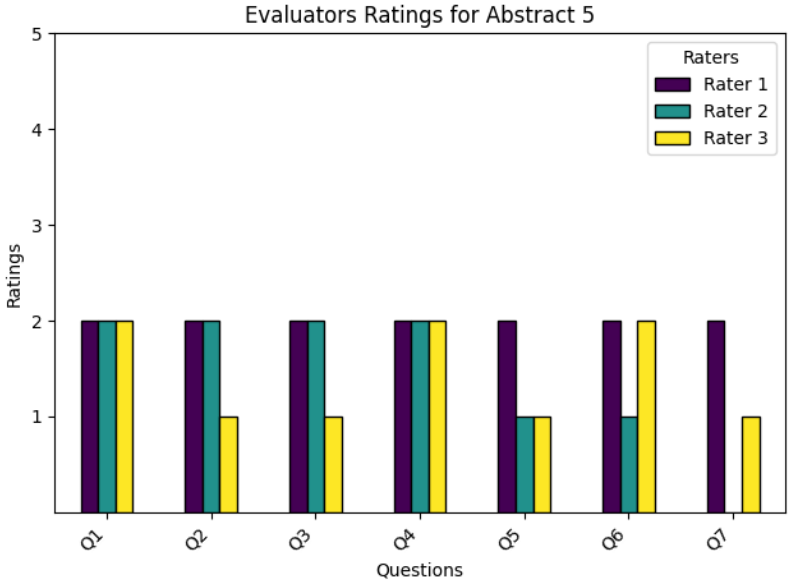
Inter-rater agreement scores

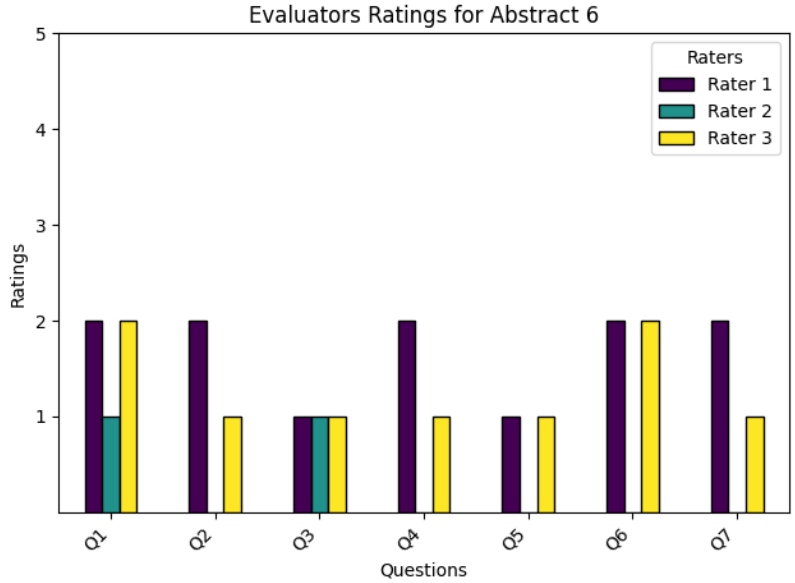


Evaluation Scores









Frame Title

Category	Original text	Translated text	Error Description
Terminology	"Rodents belong to the order Rodentia" "One Health" approach	"Rodents ni ya jamii ya Rodntina" "njia moja ya mawasiliano"	"Rodentia" is a scientific term that should remain unchanged. "One Health" is a specific health approach and is mistranslated as "njia moja ya mawasiliano" (one way of communication), which is not accurate.
Grammar & Syntax	"They were anaesthetized using Isoflurane."	"Hizo ziliundwa kwa kutumia Isoflurane."	The verb "ziliundwa" (were formed) is incorrect; should be "walilevya" /"walileweshwa..." (were anesthetized).
Coherence	"Rodent-borne diseases are transmitted either directly or indirectly..."	"Maradhi yanayoenezwa hupitishwa ama moja kwa moja ama..."	The translation does not clearly communicate the indirect transmission modes.

Discussion

- Identified issues leading to low evaluator ratings.
    - Mistranslations, grammatical errors, disrupted logical flow, domain-specific terminology (esp. agric and scientific jargon).
- There is a need for for curated Swahili-English dataset focusing on agricultural terms (how to incorporate it with LLMs? – research question).
    - In the mean time, there is still a huge need for humans in the loop when dealing with domain-specific MT.

Conclusion

- This study explored potential of LLMs for machine translation to overcome language barriers in Tanzanian agriculture.
- Translation ETD (even at abstract level) to Kiswahili is essential for reaching local, Swahili-speaking farmers, by enhancing access to agricultural research.
- OPUS-MT model shows potential but struggles with agricultural jargon and scientific names.
  - Highlights need for human oversight to ensure translation accuracy and cultural relevance.

Asante


SOKOINE UNIVERSITY OF AGRICULTURE
1984
ARDHI NI HAZINA


NDLTD