AFRIDOC-MT: Document-level MT Corpus for African Languages

Jesujoba O. Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David I. Adelani, Clement O. Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo Putini, David O. Ademuyiwa, Andrew Caines, Dietrich Klakow

Masakhane, Saarland Uni, DFKI, Inria, McGill, Mila, NII, Imperial College, ...

Motivation

- Document-level MT is crucial especially in LLM era
- And existing models (LLMs) can handle long context
- But current African language MT datasets are sentence-level
- **Question:** How well do existing MT models perform on document-level translation for African languages? 🤔

Research Objectives/contributions

1. Create a document-level parallel corpus for African languages (Amharic, Hausa, Swahili, Yorùbá, Zulu)
2. Benchmark existing MT models and LLMs on African languages for document-level translation
3. We identify key translation issues: off-targets, repetition, under-generation

AFRIDOC-MT Corpus

- Document-level dataset for two domains (*health* & *tech*)
- **Data sources**: Scraped English articles from WHO (*health*) and Techpoint Africa (*tech*)
- Articles were translated by **expert human translators** into five African languages:
  - Amharic, Hausa, Swahili, Yorùbá, Zulu
  - 334 and 271 articles from *health* and *tech*
- Languages selected based on factors such as **geographical representation**, **speaking population**, and **web coverage**

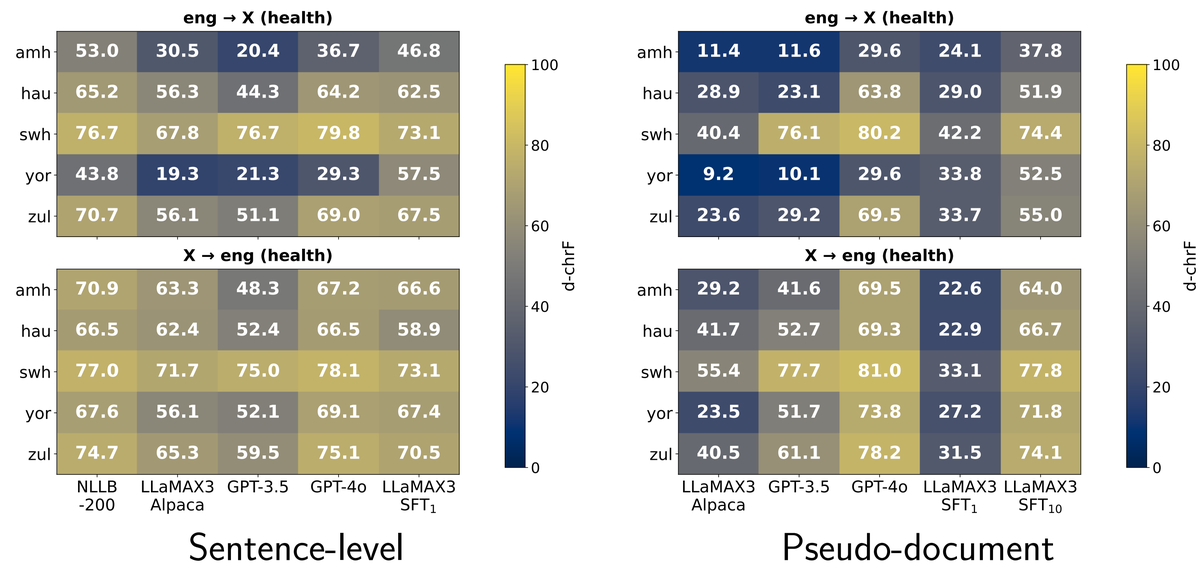
Domain	Train	Dev.	Test	Min/Max/Avg
Number of documents				
health	240	33	61	2/151/29.9
tech	187	25	59	8/247/36.9
Number of sentences				
health	7041	977	1982	-
tech	7048	970	1982	-

- **Quality check:** manual with quality estimation (AfriCOMET)

Benchmark Experiments

- **Encoder-decoder Models:** Toucan (1.2B), NLLB-200 (1.3/3B), MADLAD-400 (3B), Aya (13B)
- **Decoder-only Models:** Gemma2-IT (9B), Llama3.1-IT (8B), LlaMAX3-Alpaca (8B), GPT-3.5, GPT-4o
- **Evaluation Setup:** sentence-level, and pseudo-documents
- In the **pseudo-document** setup, documents are divided into chunks of ten sentences and concatenated for evaluation
- **Evaluation Metrics:** s-BLEU, s-chrF, d-BLEU, **d-chrF**, **GPT-4o as a judge**, **human evaluation**, qualitative analysis
- **Fine-tuning:** NLLB-200 (sentence-level) and LLaMAX3, and Llama3.1 for both evaluation setups
- **LlaMAX3-SFT$_1$** and **LlaMAX3- SFT$_{10}$** are LLaMAX3 fine-tuned on sentences and pseudo-documents respectively

Benchmark Results

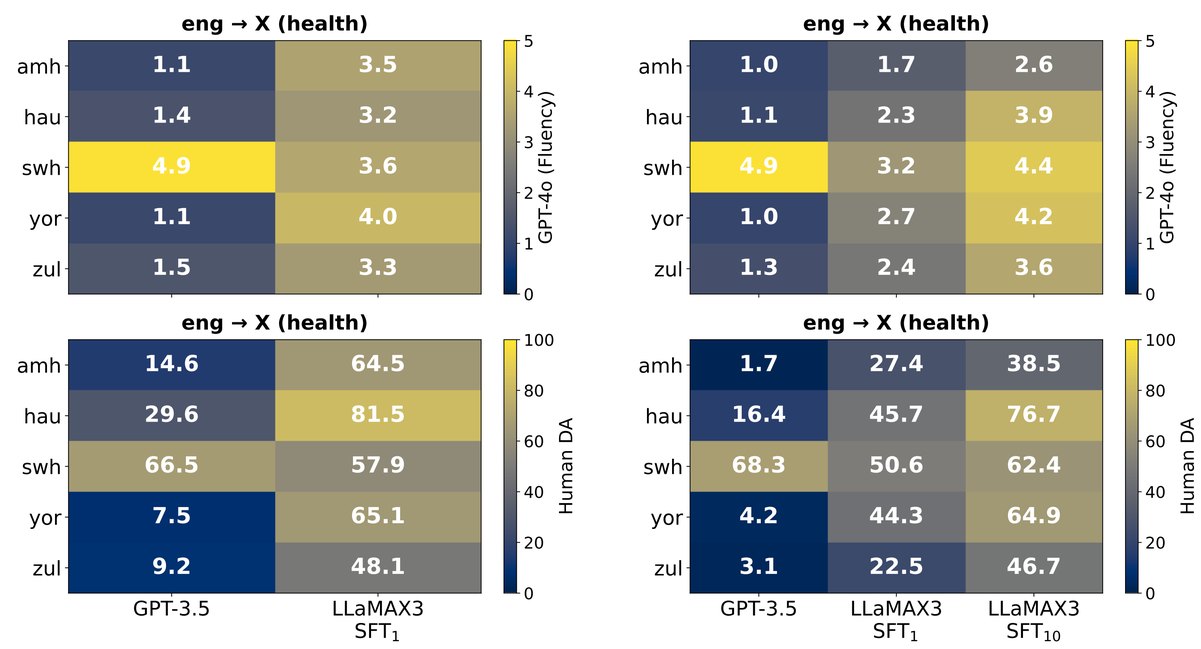


Sentence-level Pseudo-document

Findings

1. GPT-4o outperforms other decoder-only counterparts
2. Pseudo-document translation is worse than sentence-level translation when translating into African languages
3. Translation into Amharic (Ge'ez script) and Yorùbá (with diacritics) is particularly difficult
4. LLMs trained on longer documents are better for long document translation

Assessing Outputs: Humans and GPT



Sentence-level Pseudo-document

Findings

1. Human DA and GPT-4o show some agreement
2. Sentence-level translations are highly rated
3. Models frequently under/over-generate, repeat phrases, or stray off-target (length generalization problem)

Conclusion

1. We present AFRIDOC-MT and benchmark NMT/LLMs
2. Long document translation into African languages still needs attention

See our paper for many more results!