Improving Swahili Verb Parsing with Synthetic Training Data

Alessio Tosolini, Ben Kosa, Mike Violette

Background

Low resource languages are languages with little training data available.

- These languages tend to be underrepresented in Natural Language Processing (NLP).
- Many NLP architectures, like GPT-3, require large amounts of data to model language well.
- Modern high resource languages coincidentally tend to have lower morphological complexity.

Swahili → Low resource language + High morphological complexity

Not modeled well by modern NLP systems:
hawakupenda → Google Translate → **"they don't like you"**
hawakupenda → Human Expert → **"they didn't like it"**

This can be seen in the task of Morphological Segmentation:

- SOTA for high resource languages
  → **Transformer-based** | high accuracy, flexible
- SOTA for low resource languages
  → **Rule-based** | variable accuracy, unflexible

SwaRegex:

- Rule-based morphological parser for Swahili verbs (Muthee 2022).
- Only programmed to parse derivationally simple verbs.

Task: Morphological Segmentation

Our goal is to create a model that is able to segment Swahili verbs into its surface morphemes.

(1)	unamwona	*surface form*
	u-na-mw-on-a	*surface morphemes*
	u-na-m-on-a	*underlying morphemes*
	2SG.SBJ-PRS-3SG.OBJ-see-FV	*gloss*
	'You see him/her'	*free translation*

Task: Morphological Segmentation

Given the *surface form* of an inflected verb, a model must return the space-separated *surface morphemes.*

(2)	"unamwona"	*input*
	"u na mw on a"	*correct output*

* We gloss the final vowel (FV) of Swahili verbs as a separate morpheme. Some analyses of Swahili verbal morphology do not consider the final vowel as a separate morpheme, instead grouping it with the verb root. For our purposes and for consistency with SwaRegex's analysis, we consider it a standalone morpheme.

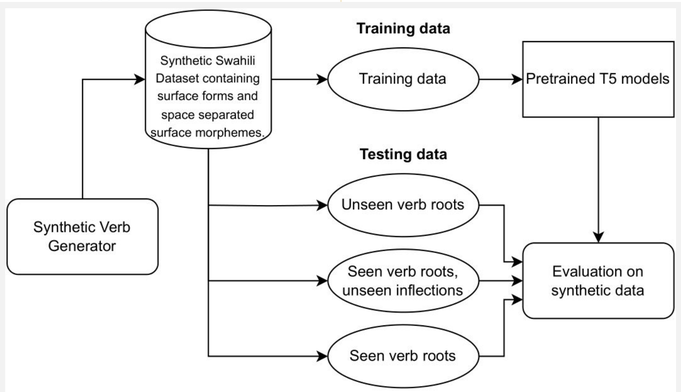
Methodology

We fine tuned the T5-small transformer provided by HuggingFace on 1,000,000 training verbs.

Prompt: "Translate Swahili to morphological tokens: {surface_form}"

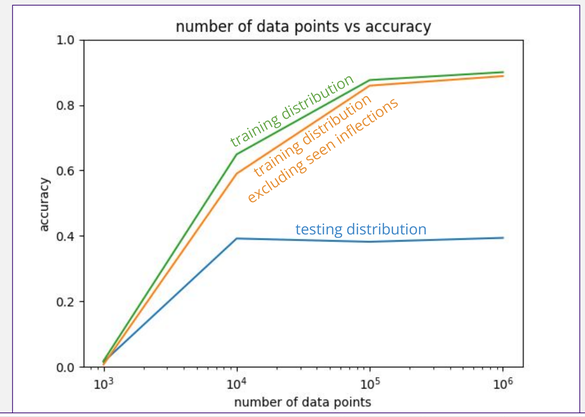
Synthetic Verb Generator: We made a stochastic generator of Swahili verbs, using a set of 500 training verb roots and 50 testing verb roots. All inflectional and derivational verb slots were used.

We calculated test accuracy on verbs from the:
- training distribution (seen verb roots),
- training distribution excluding seen inflections,
- testing distribution (unseen verb roots).



Visualization of Training Methodology

Results



Test Set Size	1k	10k	100k	1m
Accuracy on Train Distribution	0.018	0.649	0.876	0.9
Accuracy on Unique Train Distribution	0.007	0.590	0.859	0.888
Accuracy on Unique Test Distribution	0.015	0.392	0.382	0.394

- SwaRegex achieved parsed 10.4% of a testing set of 1,000 verbs, although it was not designed for derivationally complex verbs.
- Our model achieved over 88% parsing accuracy for verbs with seen roots and over 40% parsing accuracy for verbs with unseen roots.
- Accuracy plateaus around 100k training verbs.

Implications

- Generating synthetic training data can be a valid method for improving low resource parsing tasks.
- Transformer architectures perform significantly better when testing on verbs whose morphemes it has trained on.

Citations

Batsuren, K., Bella, G., Arora, A., Martinovic, V., Gorman, K., Žabokrtský, Z., … Vylomova, E. (2022, July). The SIGMORPHON 2022 Shared Task on Morpheme Segmentation. *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 103–116. doi:10.18653/v1/2022.sigmorphon-1.11

Muthee, M. G., Makau, M., & Amos, O. (2022). SwaRegex: a lexical transducer for the morphological segmentation of swahili verbs. *African Journal of Science, Technology and Social Sciences*, *1*(2), 77–84. doi:10.58506/ajstss.v1i2.119

Shikali, S. C., & Refuoe, M. (2019, November). *Language modeling data for Swahili*. doi:10.5281/zenodo.3553423