Cross-Lingual Adaptation For Named Entity Recognition in Runyankore

Our Corpus

nyn Runyankore [TARGET] Bantu
lug Luganda Bantu
kin Kinyarwanda Bantu
swa Swahili Bantu
luo Luo (Dholuo) Nilotic
twi Twi Kwa
nya Nyanja (Chichewa) Bantu
sna Shona Bantu
bbj Ghomala Grassfields Bantu
ibo Igbo Volta–Niger
tsn Tswana (Setswana) Bantu
yor Yoruba Volta–Niger
pcm Nigerian Pidgin (Naija)
hau Hausa Chadic
zul isiZulu Bantu
fon Fon Kwa
ewe Ewe Kwa
xho isiXhosa Bantu
mos Mossi (Mooré) Gur
bam Bambara Mande
wol Wolof Atlantic–Congo

Tools and Resources

Hugging Face

XLM-RoBERTa, Multilingual BERT (mBERT), AFROXLM-R

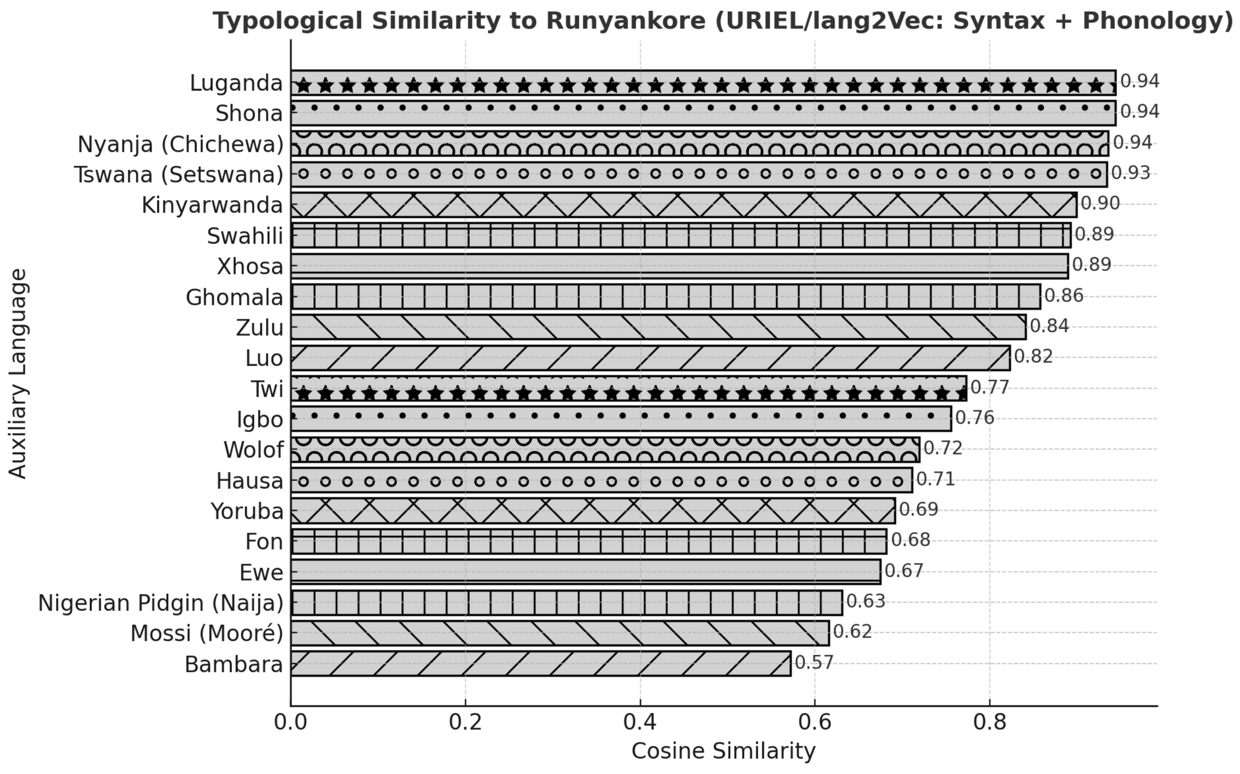
ANNOTATION TOOL : doccano

DATASETS:
Auxiliary languages: MasakhaNER 2.0

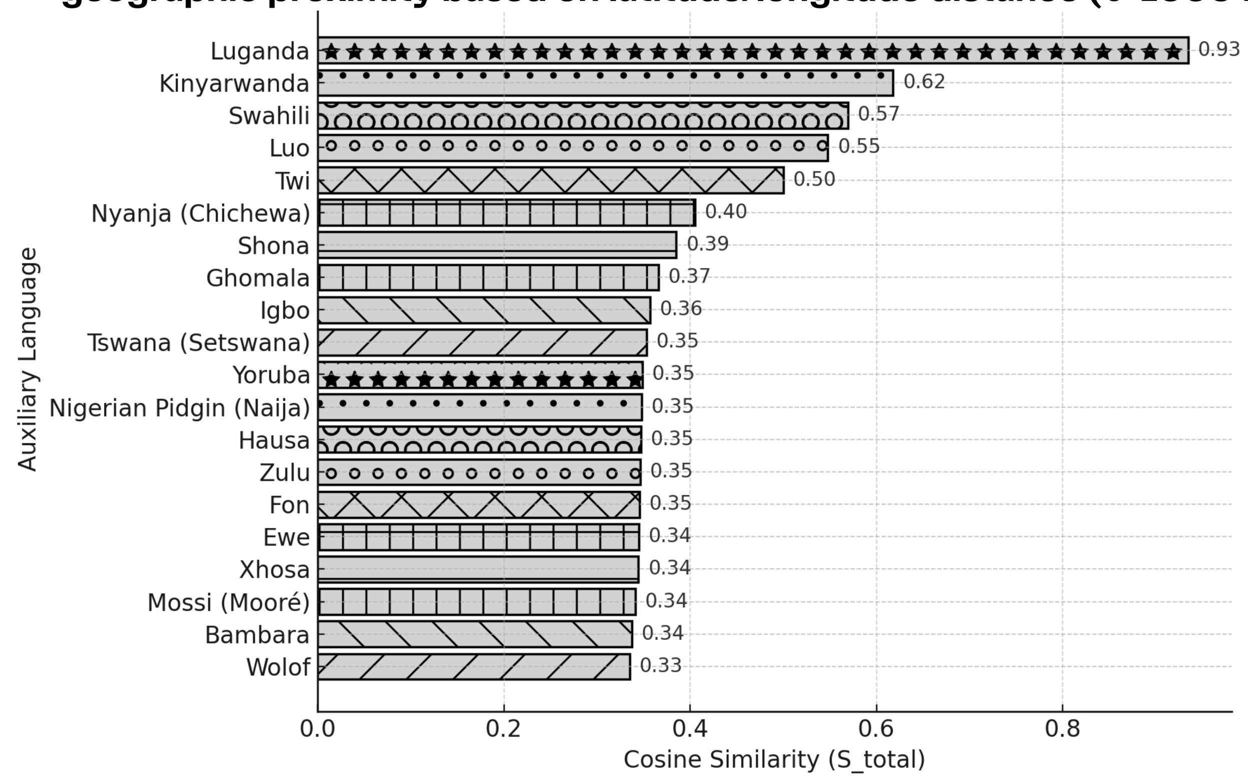
Runyankore datasets:
1. Sunbird African Language Technology (SALT)
2. Multilingual Parallel Text Corpora for East African Languages (MPTC)

TYPOLOGY-BASED METHODOLOGIES

Overall typological similarity to Runyankore computed from URIEL/lang2Vec typological feature vectors combining syntax and phonology representations.

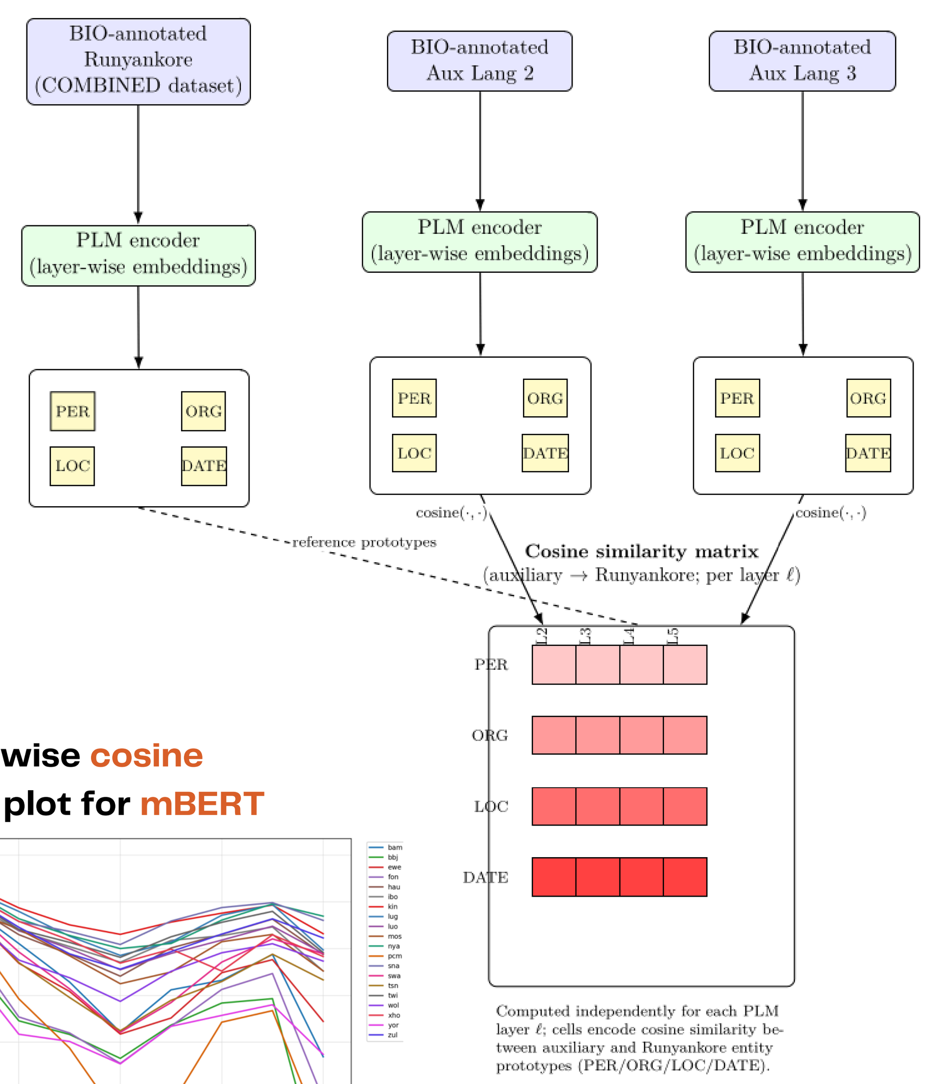


Overall typological similarity to Runyankore (S_total) computed from LinguaMeta's script match, shared country/region locale overlap, and geographic proximity based on latitude/longitude distance (τ=1000 KM).

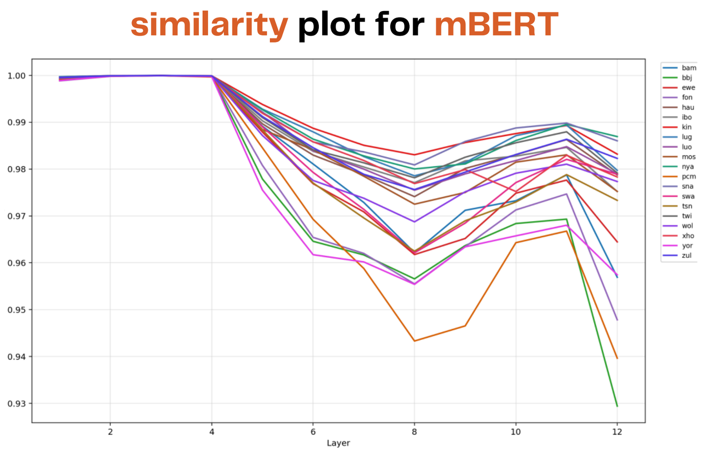


EMBEDDING-BASED METHODOLOGIES

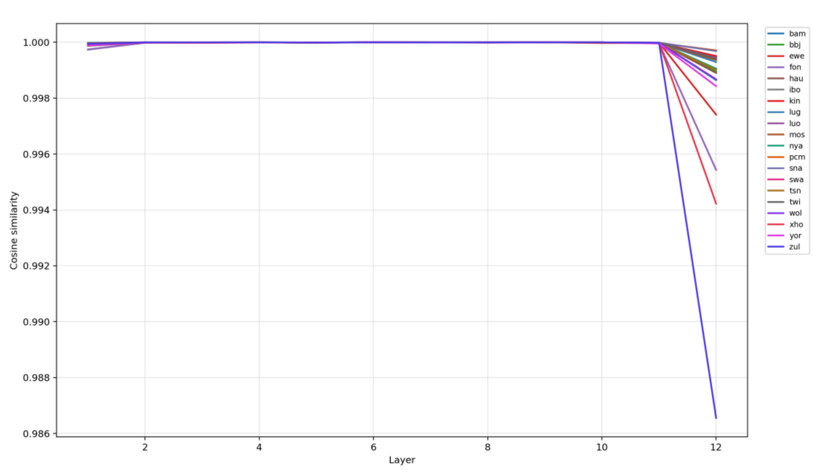
Cosine Distance



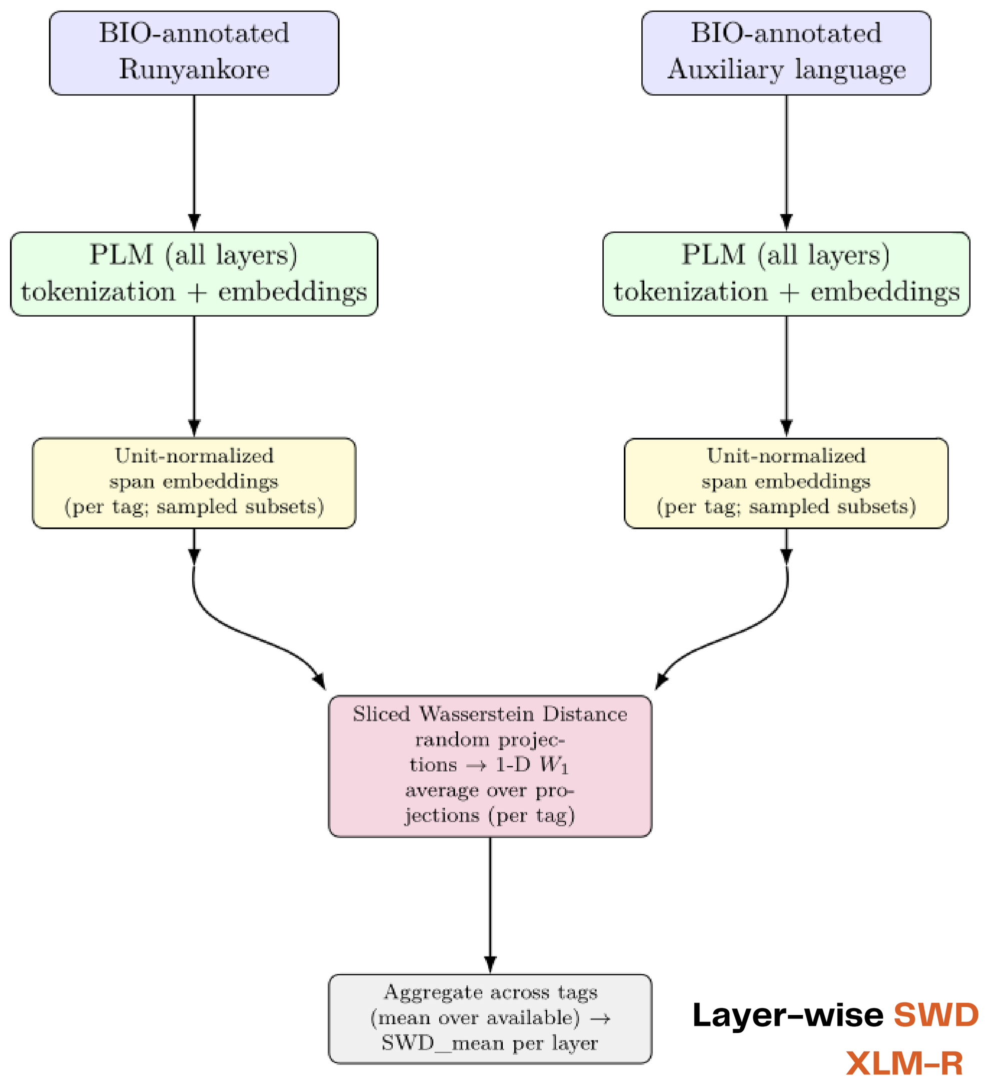
Layer-wise cosine similarity plot for mBERT



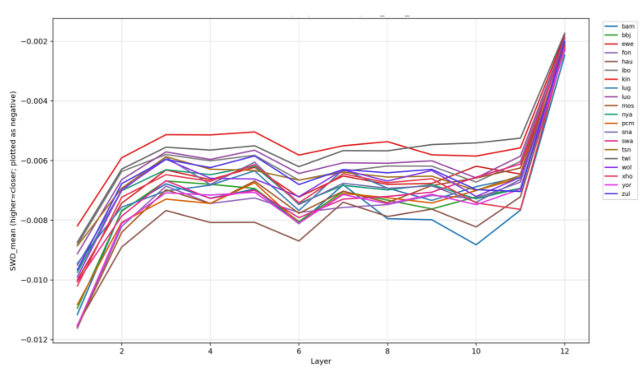
Layer-wise cosine similarity plot for XLM-R



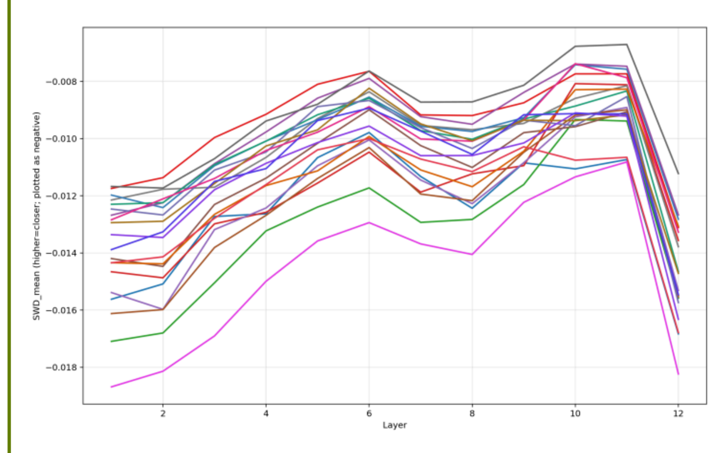
Distributional Comparison via Sliced Wasserstein Distance (SWD)



Layer-wise SWD plot for XLM-R



Layer-wise SWD plot for mBERT



Introduction / Motivation

African languages remain underrepresented in NLP, with most multilingual models performing best on high-resource languages. **Runyankore**, a major Bantu language of **Uganda**, lacks any NER dataset, excluding it from key benchmarks like MasakhaNER 2.0.

This work develops the first annotated Runyankore NER dataset and investigates cross-lingual transfer from related African languages. Using both typological features (URIEL/Lang2Vec, LinguaMeta) and embedding-based similarity measures (cosine distance, Sliced Wasserstein Distance (SWD)) on the DATE, PERSON, LOCATION, and ORGANIZATION entities, the study identifies which linguistic relationships best support transfer learning.

Dataset Creation

The Runyankore NER corpus was annotated with a semi-automated approach, where XLM-R was trained with Luganda NER data, and used to identify initial entity suggestions in the Runyankore dataset (Zero-shot transfer), following MasakhaNER 2.0 guidelines, covering PERSON, ORG, LOC, DATE, and O for non-entity types. Over 5,000 sentences were double-annotated and cross-checked for quality assurance.

Example (BIO format):

Besigye B-PER
akeetaba O
omu O
mirimo O
y'eby'obutegyeki O
omuri O
Uganda B-LOC
omu O
mwaka B-DATE
gwa I-DATE
rukumi I-DATE
rwenda I-DATE
nshanju I-DATE
. O

Entity distribution across SALT, MPTC, and COMBINED datasets. Columns B and I represent Beginning and Inside tags, respectively.

Dataset	Entity Type	Train		Dev		Test	
		B	I	B	I	B	I
SALT	DATE	637	993	328	440	333	486
	LOC	1179	263	586	149	620	142
	ORG	261	392	136	215	136	191
	PER	198	106	86	43	113	55
MPTC	DATE	204	195	106	94	104	102
	LOC	191	48	100	36	94	33
	ORG	47	86	18	30	25	41
	PER	2	1	1	0	1	1
COMBINED	DATE	841	1188	434	534	437	588
	LOC	1370	311	686	185	714	175
	ORG	308	478	154	245	161	232
	PER	200	107	87	43	114	56

Methodology

This study combines data creation, language similarity analysis, and cross-lingual NER experiments to evaluate transfer into Runyankore.

- **Dataset Development: A new Runyankore NER corpus was annotated using MasakhaNER 2.0 guidelines, covering Person, Organization, Location, and Date entities, initially identified in raw Runyankore data by applying a zero-shot transfer from an XLM-R PLM trained on Luganda NER data, then manually verified and corrected by a human.**
- **Model Evaluation: Three multilingual pre-trained models: mBERT, XLM-R, and AfroXLM-R, were tested under zero-shot and cross-lingual setups.**
- **Typological Similarity: Computed from URIEL/lang2Vec and LinguaMeta features, focusing on syntax, phonology, and geographic proximity to identify linguistically related source languages.**
- **Embedding-Based Similarity: Calculated from contextual embeddings using cosine distance and the Sliced Wasserstein Distance (SWD) to capture representational closeness across languages.**
- **Transfer Comparison: The two similarity strategies: typological vs embedding-based guided auxiliary language selection to determine which better predicts cross-lingual transfer effectiveness.**

Together, these techniques provide a structured framework for analyzing how linguistic and representation-level similarities influence the success of transfer learning for low-resource African languages.

Research Objectives

This study investigates two key questions:
(1) How do multilingual model pretraining choices influence NER performance in Runyankore?
(2) Which language similarity strategy, typological or embedding-based, more effectively guides auxiliary language selection for cross-lingual transfer?

Findings indicate that embedding-based similarity shows a modest but consistent correlation with improved transfer, particularly among closely related Bantu languages such as Luganda and Kinyarwanda.
Overall, embedding-driven selection achieved slight yet measurable gains over typology-based approaches.

Conclusion & Future Work

This study introduced the first Runyankore NER dataset and demonstrated the advantages of embedding-based auxiliary language selection for cross-lingual transfer over typological-based selection.

Future work will extend to other Runyakitara languages: Rukiga, Runyoro, and Rutoro, and integrate findings into MasakhaNER pipelines to strengthen African-language representation in multilingual NLP.

Author: Prosper ARINEITWE ASIIMWE arnari002@myuct.ac.za
Msc. Computer Science (Artificial Intelligence)
Supervisors: Dr. Jan BUYS jan.buys@uct.ac.za
Dr. Francois MEYER francois.meyer@uct.ac.za

University of Cape Town · iYunivesithi yaseKapa · Universiteit van Kaapstad