KinyaBERT: a Morphology-aware Kinyarwanda Language Model

Dejan Dichoski
and Josè Chacon

BERT for low-resource languages

“Bidirectional Encoder Representations from Transformers is a family of language models introduced in 2018 by researchers at Google”. - Wikipedia

- Most of BERT's evaluations have been conducted on high-resource languages.
  - This has obscured its applicability on low-resource languages.
- BERT-like models use sub-word tokenization algorithms, such as BPE.
  - These models are not optimal for morphologically rich languages, even given a morphological analyzer.

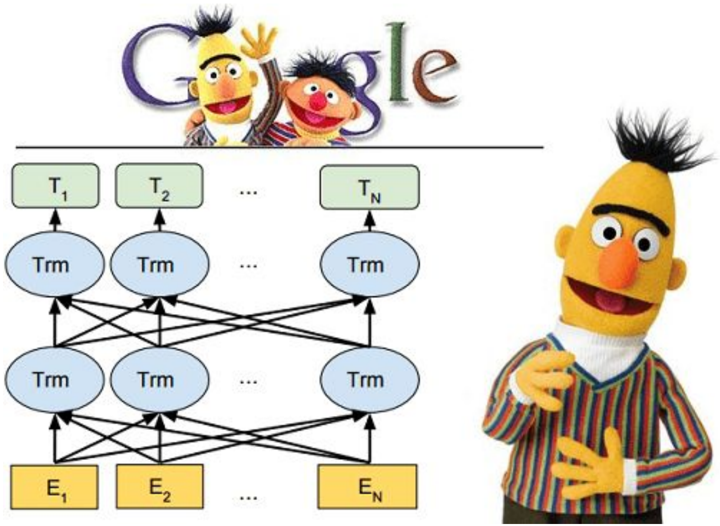


Image source: [GOOGLE BERT. (Bidirectional Encoder Representation… | by Anish Thapaliya | Medium]

Kinyarwanda
A morphologically rich language

Morphological segmentation of the word ‘ntuzamwibeshyeho’

ntuzamwibeshyeho							
nti-	**-u-**	**-za-**	**-mu-**	**-ii-**	**-beshy-**	**-e-**	**-ho**
Negation	**Subject** (2nd pers/sing.)	**Tense** (future)	**Direct Object** (1st class/human/sing.)	**Reflexive** (wrt. subj)	**Stem**	**Aspect** (imperative)	**Prep.**
not	**you**	**will**	**him/her**	**self**	**lie**	**(imperative)**	**about**

Morpheme-by-morpheme translation of the word — *’Never lie to yourself about him/her’*

Real meaning — *’Never underestimate him/her’*

*Antoine Nzeyimana (2020) - Morphological disambiguation from stemming data

Contributions

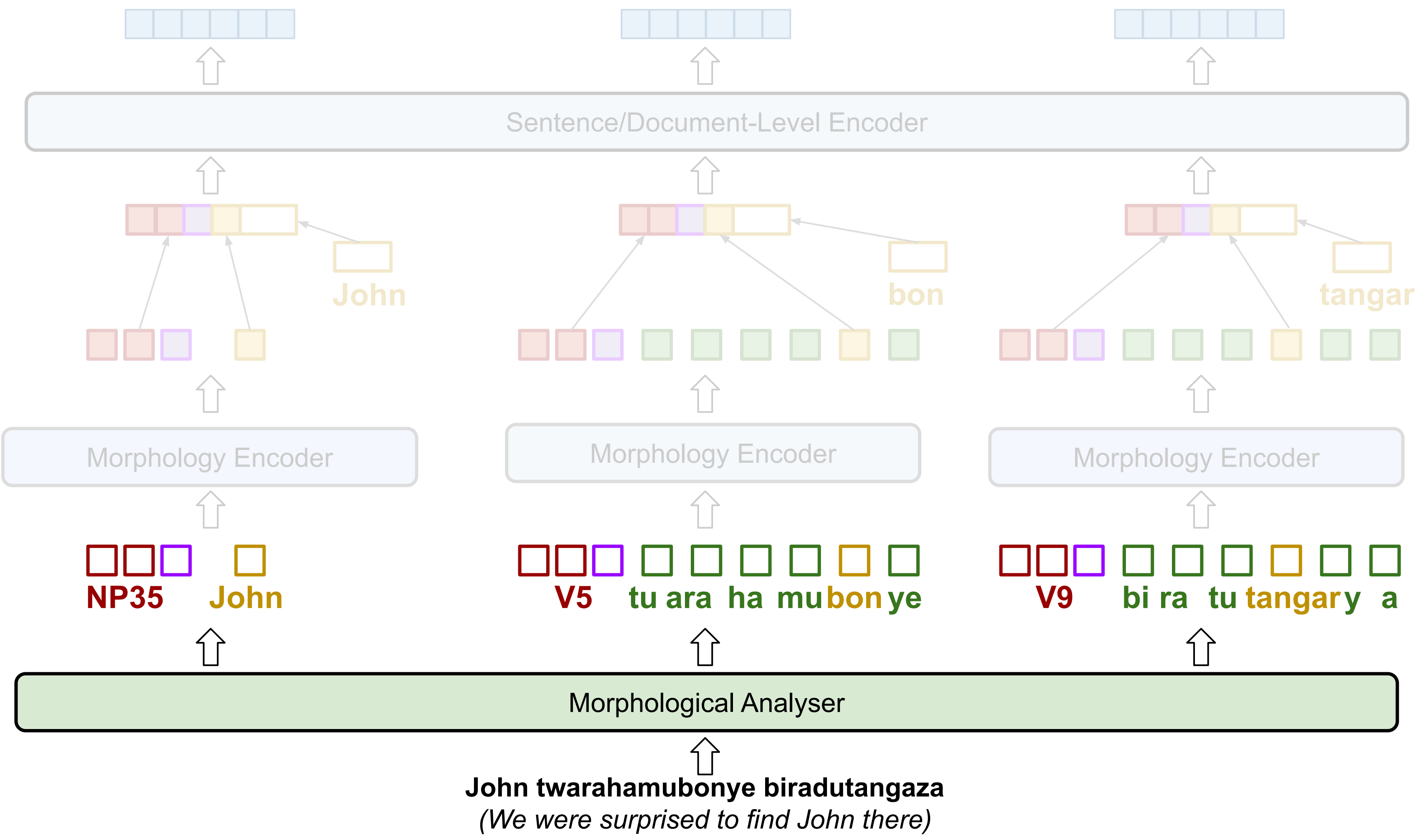
Architecture:

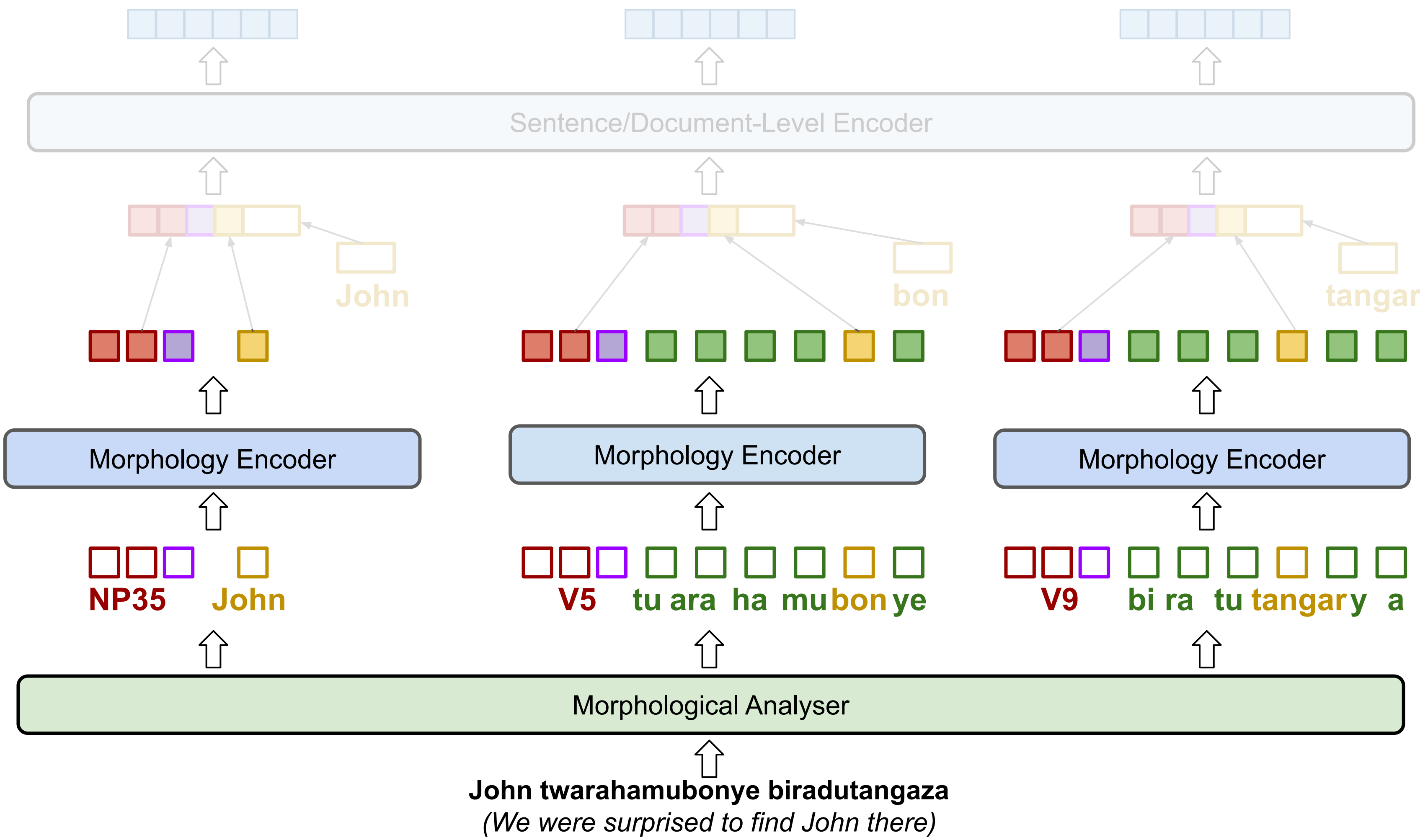
- A BERT architecture designed specifically for morphologically rich languages, like the Kinyarwanda language.

Model evaluation:

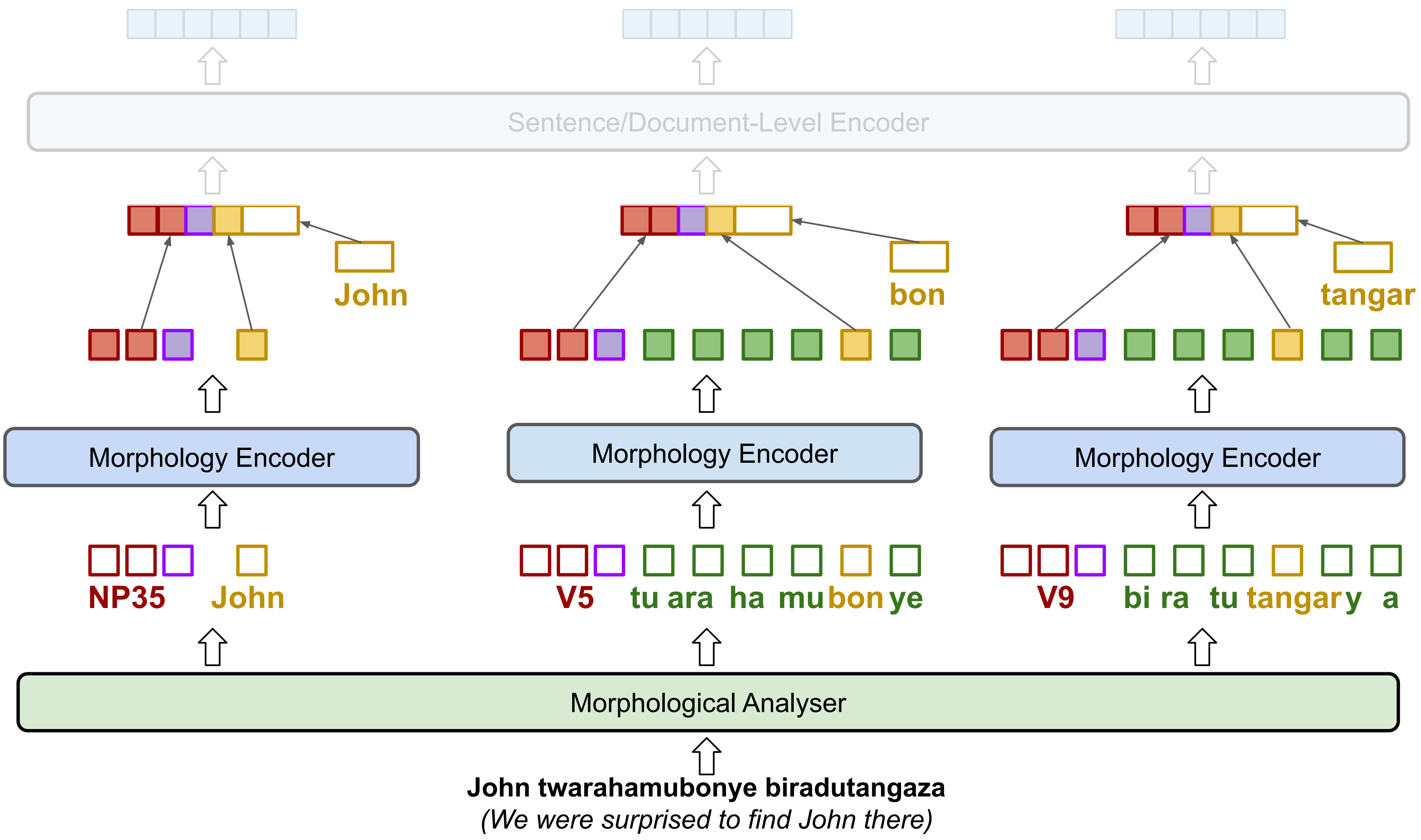
- A machine-translated subset of the GLUE benchmark and an author-generated news categorization dataset.
- Benchmark for future studies on Kinyarwanda language understanding.

KinyaBERT model architecture

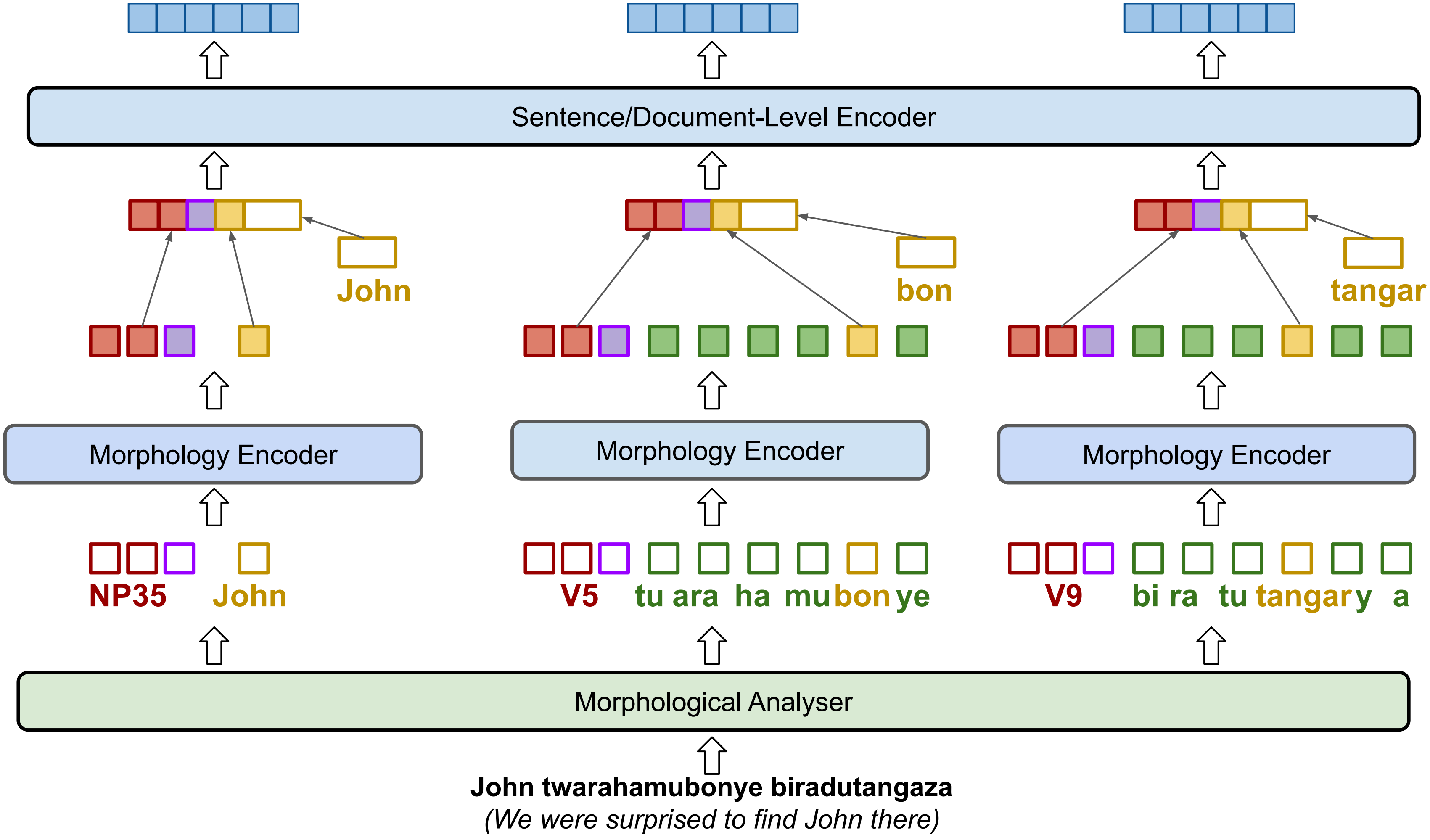

Sentence/Document-Level Encoder
John
bon
tangar
Morphology Encoder
Morphology Encoder
Morphology Encoder
NP35
John
V5
tu ara ha mu bon ye
V9
bi ra tu tangar y a
Morphological Analyser
John twarahamubonye biradutangaza
(We were surprised to find John there)



Sentence/Document-Level Encoder
John
bon
tangar
Morphology Encoder
Morphology Encoder
Morphology Encoder
NP35
John
V5
tu ara ha mu bon ye
V9
bi ra tu tangar y a
Morphological Analyser
John twarahamubonye biradutangaza
(We were surprised to find John there)



Sentence/Document-Level Encoder
John
bon
tangar
Morphology Encoder
Morphology Encoder
Morphology Encoder
NP35
John
V5
tu ara ha mu bon ye
V9
bi ra tu tangar y a
Morphological Analyser
John twarahamubonye biradutangaza
(We were surprised to find John there)


Sentence/Document-Level Encoder
John
bon
tangar
Morphology Encoder
Morphology Encoder
Morphology Encoder
NP35 John
V5 tu ara ha mu bon ye
V9 bi ra tu tangar y a
Morphological Analyser
John twarahamubonye biradutangaza
(We were surprised to find John there)

Model pre-training

- Objective: Utilize a masked language model to predict stems and their associated affixes for all tokens.
- Affix prediction task = a multi-label classification problem, tackled using two methods:
  1. **Affix Distribution Regression (ADR) - KinyaBERT$_{\text{ADR}}$**
     - Use the *Kullback–Leibler (KL) divergence loss* function to predict the N-length affix distribution vector.
  2. **Affix Set Classification (ASC) - KinyaBERT$_{\text{ASC}}$**
     - Use *cross entropy loss* to predict the affix set associated with the target word.

Evaluation Tasks and Results

Machine-translated GLUE Benchmark

- Collection of nine natural language understanding tasks.
- Used Google Translate API to translate a subset of the GLUE benchmark into Kinyarwanda.

Task:	**MRPC**	**QNLI**	**RTE**	**SST-2**	**STS-B**	**WNLI**
#Train examples:	**3.4K**	**104.7K**	**2.5K**	**67.4K**	**5.8K**	**0.6K**
Translation score:	**2.7/4.0**	**2.9/4.0**	**3.0/4.0**	**2.7/4.0**	**3.1/4.0**	**2.9/4.0**
Model	**Validation Set**					
XLM-R	84.2/78.3±0.8/1.0	79.0±0.3	58.4±3.2	78.7±0.6	77.7/77.8±0.7/0.6	55.4±2.0
BERT$_{BPE}$	83.3/76.6±0.8/1.4	81.9±0.2	59.2±1.5	80.1±0.4	75.6/75.7±7.8/7.3	55.4±1.9
BERT$_{MORPHO}$	84.3/77.4±0.6/1.1	81.6±0.2	59.2±1.5	81.6±0.5	76.8/77.0±0.8/0.7	54.2±2.5
KinyaBERT$_{ADR}$	**87.1/82.1**±0.5/0.7	81.6±0.1	61.8±1.4	81.8±0.6	79.6/79.5±0.4/0.3	54.5±2.2
KinyaBERT$_{ASC}$	86.6/81.3±0.5/0.7	**82.3**±0.3	**64.3**±1.4	**82.4**±0.5	**80.0/79.9**±0.5/0.5	**56.2**±0.8
Model	**Test Set**					
XLM-R	82.6/76.0±0.6/0.6	78.1±0.3	56.4±3.2	76.3±0.4	69.5/68.9±1.0/1.1	63.7±3.9
BERT$_{BPE}$	82.8/76.2±0.6/0.8	81.1±0.3	55.6±2.8	79.1±0.4	68.9/67.8±1.8/1.7	63.4±4.1
BERT$_{MORPHO}$	82.7/75.4±0.8/1.3	80.8±0.4	56.7±1.0	80.7±0.5	68.9/67.8±1.5/1.3	<u>65.0</u>±0.3
KinyaBERT$_{ADR}$	84.4/**78.7**±0.5/0.6	81.2±0.3	58.1±1.1	80.9±0.5	73.2/72.0±0.4/0.3	<u>65.1</u>±0.0
KinyaBERT$_{ASC}$	**84.6**/78.4±0.2/0.3	**82.2**±0.6	**58.8**±0.7	**81.4**±0.6	**74.5/73.5**±0.2/0.2	<u>65.0</u>±0.2

KinyaBERTASC achieved a 4.3% better average score than the strongest baseline.

Named entity recognition (NER)

The task requires predicting 4 entity types annotated by native speakers for Kinyarwanda:

1. Persons (PER)
2. Locations (LOC)
3. Organizations (ORG)
4. Date and time (DATE).

Task:	**NER**	
#Train examples:	**2.1K**	
Model	**Validation Set**	**Test Set**
XLM-R	80.3±1.0	71.8±1.5
BERT_{BPE}	83.4±0.9	74.8±0.8
BERT_{MORPHO}	83.2±0.9	72.8±0.9
KinyaBERT_{ADR}	**87.1**±0.8	**77.2**±1.0
KinyaBERT_{ASC}	86.2±0.4	76.3±0.5

KinyaBERTADR achieves best performance, about 3.2% better average F1 score than the strongest baseline.

News categorization

This is a document classification task (12 categories), using data collected from seven major news websites that regularly publish in Kinyarwanda.

Task:	**NEWS**	
#Train examples:	**18.0K**	
Model	**Validation Set**	**Test Set**
XLM-R	83.8±0.3	84.0±0.2
BERT$_{BPE}$	87.6±0.4	**88.3**±0.3
BERT$_{MORPHO}$	86.9±0.4	86.9±0.3
KinyaBERT$_{ADR}$	**88.8**±0.3	88.0±0.3
KinyaBERT$_{ASC}$	88.4±0.3	88.0±0.2

Differing performances between validation and test sets.

Conclusion

- The proposed BERT architecture has shown to be capable of representing morphological compositionality.

- Experiments conducted on Kinyarwanda language, revealed performance improvement on downstream NLP tasks, affirming the effectiveness of morphology-aware language models.

Future work

- Further research into morphologically-aware language models is needed.

- Character-aware language models have been presented as an alternative to the current subword tokenization techniques.
  - A comparison between this approach and morphology-aware models remains an open area for research.

Thanks for your... 🥁

Attention

Transformer image source: "Attention Is All You Need" paper