

# A Journey Through the Opportunity of Low Resourced Natural Language Processing — An African Lens



UNIVERSITEIT VAN PRETORIA  
UNIVERSITY OF PRETORIA  
YUNIBESITHI YA PRETORIA



Data Science for Social Impact

Vukosi Marivate and David Adelani



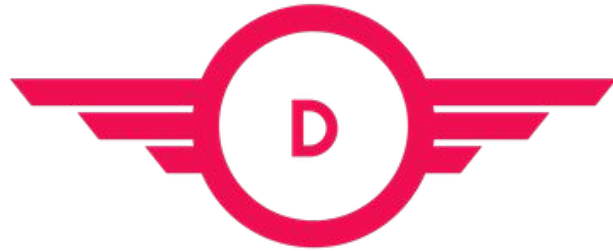
**SIC**

Saarland Informatics  
Campus

*Motho ke Motho ka Batho*  
*(Setswana - tn,tsn)*  
A Person Is a Person  
Because of Others



# Who I Am - Dr. Vukosi Marivate @vukosi



Data Science for Social Impact



DEEP LEARNING  
INDABA



ABSA UP Chair of Data Science at the University of Pretoria  
Co-Founder Deep Learning Indaba and Masakhane NLP  
PI for Data Science and Social Impact

<https://dsfsi.github.io/>

# Who I Am - David Ifeoluwa Adelani

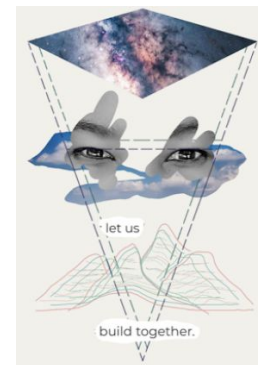
PhD Student (@davlanade)

Spoken Language Systems Group, Saarland University,  
Saarland Informatics Campus, Saarbrücken, Germany.

Active Member of  
Masakhane NLP



**SIC** Saarland Informatics  
Campus



# Synthesis of an Idea - *Hundzula*

hundzula [xitsonga]

**verb:**

- to change

**noun:**

- the act or process through which something  
becomes different



# STOKING THE FIRE



# Low Resource Natural Language Processing - Languages

Many Speakers, Not Many Resources

Data

- Speech
- Text
- Etc.

Tools for language (digital dictionaries, grammar tools etc.)



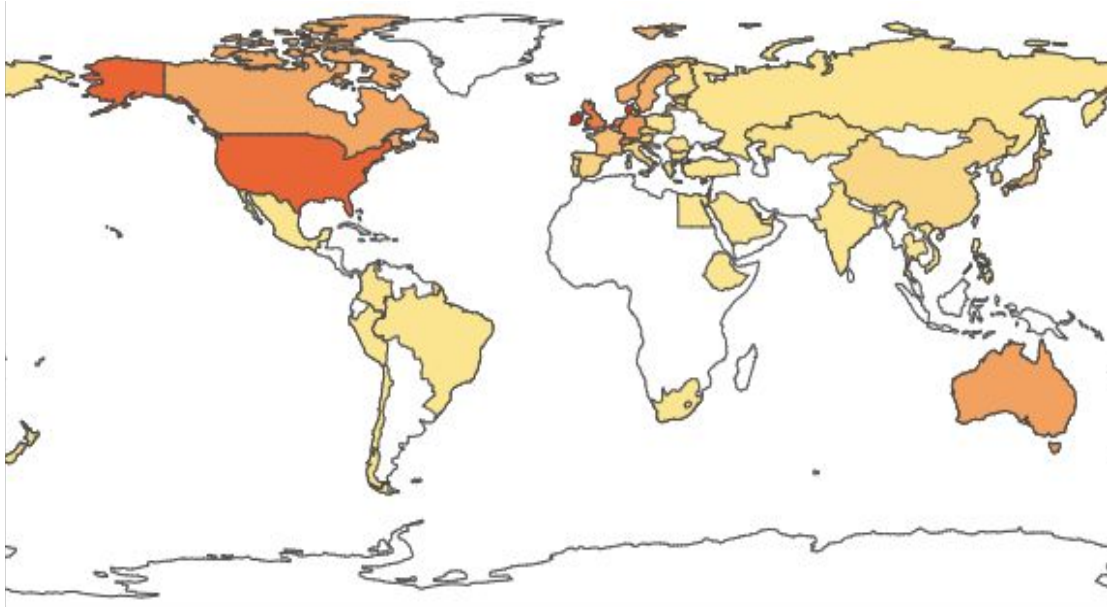
# Low Resource Natural Language Processing - Domain

Low Resource NLP also covers Domains or NLP Tasks that do not have enough resources

- Health/Medical NLP Data
- Government Interactions
- Legal

We take a view on both, but do it through looking at African Languages.

# Mapping the World



Normalized paper count by country at the 2018 NLP conferences ([Caines, 2019](#))

# Defining the challenges to Low Resource Languages

Low availability of resources (Data, Tools, et

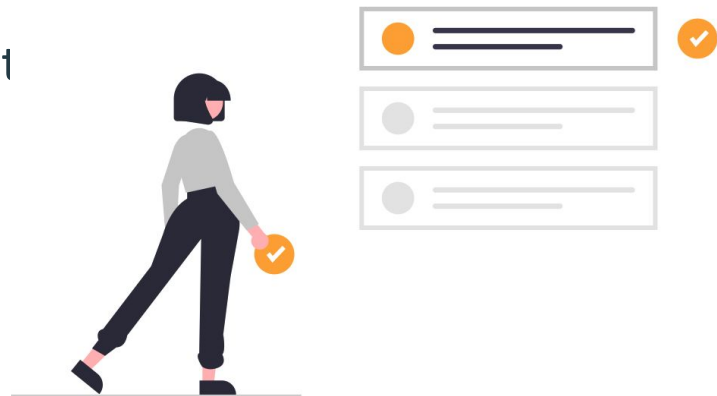
Discoverability

Reproducibility

Focus

Benchmarks

Scale and Complexity



**A Focus on Neural Machine Translation for African Languages**

**Laura Martinus**

Explore / Johannesburg, South Africa  
laura@explore-ai.net

**Jade Abbott**

Retro Rabbit / Johannesburg, South Africa  
ja@retrorabbit.co.za

# Defining the challenges to Low Resource Languages

Low availability of resources (Data, Tools, et

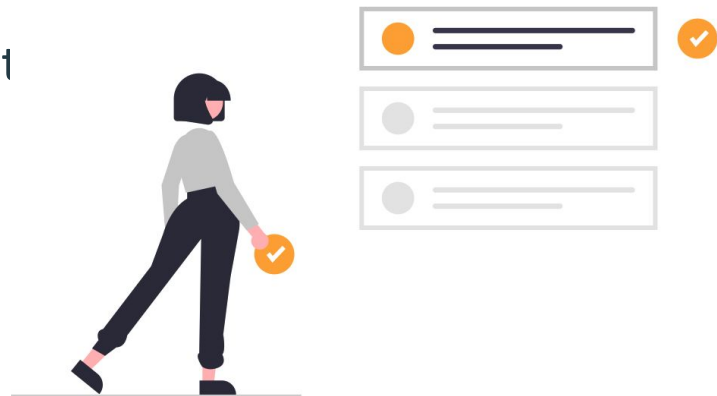
Discoverability

Reproducibility

Focus

Benchmarks

Scale and Complexity



**A Focus on Neural Machine Translation for African Languages**

**Laura Martinus**

Explore / Johannesburg, South Africa  
laura@explore-ai.net

**Jade Abbott**

Retro Rabbit / Johannesburg, South Africa  
ja@retrorabbit.co.za

# Defining the challenges to Low Resource NLP

The Left-Behinds

The Scraping-Bys

The Hopefuls

The Rising Stars

The Underdogs

The Winners

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Table 1: Number of languages, number of speakers, and percentage of total languages for each language class.

**The State and Fate of Linguistic Diversity and Inclusion in the NLP World**

Pratik Joshi\* Sebastin Santy\* Amar Budhiraja\*  
Kalika Bali Monojit Choudhury  
Microsoft Research, India  
{t-prjos, t-sesan, amar.budhiraja, kalikab, monojitc}@microsoft.com

# Defining the challenges to Low Resource NLP

The Left-Behinds

The Scraping-Bys

The Hopefuls

The Rising Stars

The Underdogs

The Winners

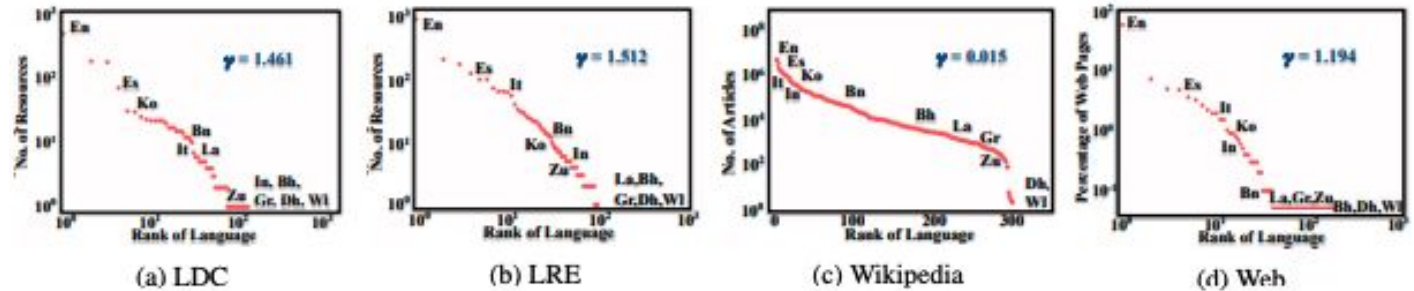


Figure 3: Plots of different available resources for different languages. Languages to the far right do not have a representation in the resource category. Languages annotated are: Class 0-Dahalo (Dh), Wallisian(Wl); Class 1-Bhojpuri (Bh), Greenlandic (Gr); Class 2-Lao (La), Zulu (Zu); Class 3- Bengali (Bn), Indonesian (In); Class 4- Korean (Ko), Italian (It); Class 5- English (En), Spanish (Es).

# Our History Shapes our Current and Future

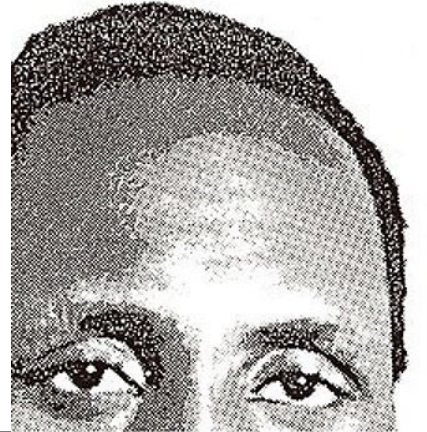
Many countries in the Global South affected by a history of colonialism

The effects of colonialism also touch on Language and how we ourselves experience it.

Ultimately, the resources we have today, affected by this history.

- Many available resources produced by Christian Missionaries.

**NGŪGĪ**  
**WA THIONG'O**  
*Decolonising the Mind*  
THE POLITICS OF LANGUAGE  
IN AFRICAN LITERATURE



# Our History Shapes our Current and Future

Ultimately, the resources we have today, affected by this history.

- Many available resources produced by Christian Missionaries.
- A focus on English as the language.
- Translation tools affected recording and shaping of local languages.

How do we change it?

***missionary linguists have played a particular role in the construction and invention of languages around the world. Of particular concern here are the ways in which language use, and understandings of language use, have been-and still are-profoundly affected by missionary projects.***



# Our Personal Histories with Language

In many African countries, historically local languages looked down upon as a consequence of colonialism.

Look at the Irish situation with the British. The humiliation of Native Americans, how their language was denigrated. ***In Africa, of course, we were forbidden to speak our mother tongues.*** Japan imposed its language on the Koreans. So wherever you look at modern colonialism, ***the acquisition of the language of the colonizer was based on the death of the languages of the colonized.*** So it is a war zone. - ***Ngũgĩ wa Thiong'o***



Politics World Culture Events Shop

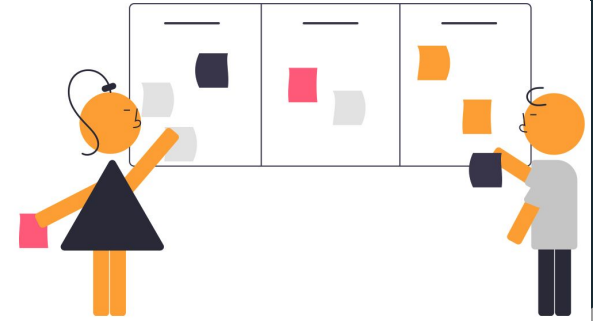
AFRICA EDITORIAL APRIL 2, 2018, ISSUE

## Language Is a 'War Zone': A Conversation With Ngũgĩ wa Thiong'o

The Kenyan author discusses colonialism and abandoning English to write in his native Kikuyu.

By Rohit Inani 

# A challenge worth tackling



## Māori are trying to save their language from Big Tech

Te Hiku Media gathered huge swathes of Māori language data. Corporates are now trying to get the rights to it

Represents US

Represents Culture

Carries Indigineous Culture

A blind spot within Big Tech

- Not just a tech problem to solve.

Participation is key 



GETTY IMAGES / TIM GRAHAM / CONTRIBUTOR

# Non action can be life changing



the facebook files

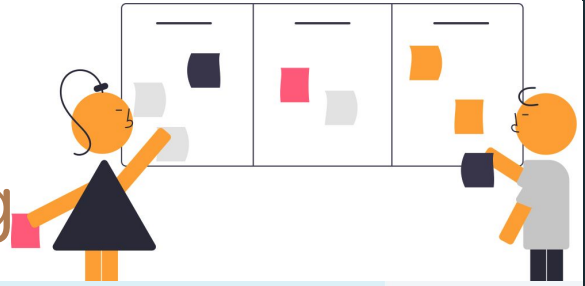
## Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show.

Employees raised alarms about how the site is used in developing countries, where its user base is already huge and expanding

The developing world already has hundreds of millions more Facebook users than the U.S.— more than **90%** of monthly users are now outside the U.S. and Canada. With growth

Facebook employees and contractors spent more than 3.2 million hours searching labeling or, in some cases, taking down information the company concluded was false or misleading, the documents show. Only **13% of those hours were spent** working on content from outside the U.S. The company spent almost three times as many hours outside the U.S. working on “brand safety,” such as making sure ads don’t appear alongside content advertisers may find objectionable.

# Non action can be life changing



Power asymmetries continue.

Who are the perceived users of the systems?

Rule Based and Statistical Translation has a history in law enforcement.



## Google Says Google Translate Can't Replace Human Translators. Immigration Officials Have Used It to Vet Refugees.

Documents shared with ProPublica show that immigration officials have been told to vet refugees' social media posts using Google Translate. Language experts caution even students against using the service.

<https://thegradient.pub/machine-translation-shifts-power/>

by Yeganeh Torbati, Sept. 26, 2019, 11:37 a.m. EDT

# Language is more than symbols

Language as communication and as culture are then products of each other. ***Communication creates culture: culture is a means of communication. Language carries culture, and culture carries, particularly through orature and literature, the entire body of values by which we come to perceive ourselves and our place in the world.*** How people perceive themselves and affects how they look at their culture, at their places politics and at the social production of wealth, at their entire relationship to nature and to other beings. Language is thus inseparable from ourselves as a community of human beings with a specific form and character, a specific history, a specific relationship to the world — Decolonising the Mind (16)

**NGŪGĪ**  
**WA THIONG'O**  
*Decolonising the Mind*  
THE POLITICS OF LANGUAGE  
IN AFRICAN LITERATURE



What has changed  
over the last 5 years  
that makes this an  
exciting time



# Data + Compute + People

We have had the rise of the TRIFECTA

# More Data

Lacuna Fund

Masakhane Data

AI4D Data

Better Curation

zenodo Search Upload Communities vima@vima.co.za

## African Natural Language Processing (AfricaNLP)

Recent uploads

Search African Natural Language Processing (AfricaNLP) View

December 1, 2020 (0.2) Dataset Open Access View

### Swahili : News Classification Dataset

Davis David;

Swahili is spoken by 100-150 million people across East Africa. In Tanzania, it is one of two national languages (the other is English) and it is the official language of instruction in all schools. News in Swahili is an important part of the media sphere in Tanzania. News contributes to education,

Uploaded on September 17, 2021

1 more version(s) exist for this record

July 31, 2021 (1.0) Dataset Open Access View

### South African Disinformation [Fake News] Website Data - 2020

Harm, de Wet; Marivate, Vukosi;

See publication: Is it Fake? News Disinformation Detection on South African News Websites We used, as sources, investigations by the news websites MyBroadband (<https://mybroadband.co.za/forum/threads/list-of-known-fake-news-sites-in-south-africa-and-beyond.879854/>) and News24 (<https://exposed>).

New upload

### African Natural Language Processing (AfricaNLP)

This is a project that aims to curate African Natural Language Processing projects and data that is uploaded on Zenodo. Please fill free to tag this community when uploading.

**Curated by:**  
vukosi

**Curation policy:**  
We are very much open to any submissions but may ask for added documentation to projects when needed.

**Created:**  
February 15, 2020

**Harvesting API:**  
[OAI-PMH Interface](#)



# Compute

Better Online Compute

Lowering Barrier to Entry

Collaboration

Open Sharing

kaggle™



HUGGING FACE



Google  
colab



GitHub



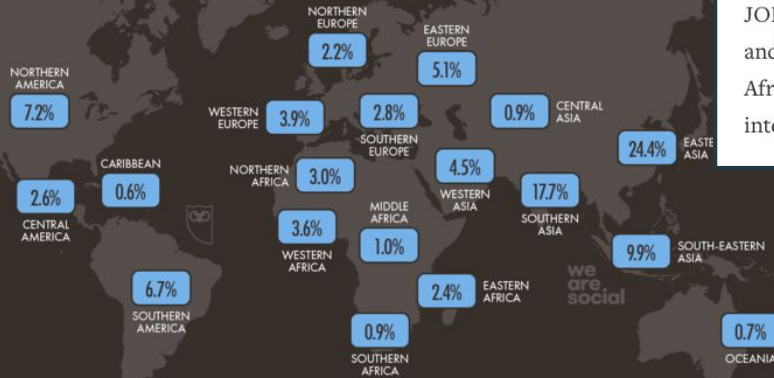
open source

# Exploding Social Media Communities

JAN  
2021

## SHARE OF GLOBAL INTERNET USERS BY REGION

THE NUMBER OF INTERNET USERS IN EACH REGION AS A PERCENTAGE OF THE TOTAL NUMBER OF GLOBAL INTERNET USERS



## Young Africans go online to preserve local languages, fight COVID-19

By Kim Harrisberg, Kristi Eaton

7 MIN READ



JOHANNESBURG/TULSA (Thomson Reuters Foundation) - The words “facemask” and “hand sanitiser” are now familiar the world over, but for isiZulu speakers in South Africa those terms did not exist a year ago, until a group of volunteers took to the internet to create them.

“Young Africans Go Online to Preserve Local Languages, Fight COVID-19.” Reuters, 1 Apr. 2021. [www.reuters.com, https://www.reuters.com/article/us-africa-internet-youth-trfn-idUSKBN2BO49E](https://www.reuters.com/article/us-africa-internet-youth-trfn-idUSKBN2BO49E) .

# Emergence of Grassroots AI



CERN

1954



Santa Fe Institute

1984

UBUNTU

MD4SG



Delta Analytics



Women in Machine Learning



WiMLDS



BAI



Natural Language Processing



DEEP LEARNING INDABA



Neuroscience Imbizo



Data Science Africa



MLT



Queer in AI



Data Science Nigeria



LatinX AI



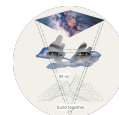
DisAbility in AI



Indigenous AI



FOR.AI



Masakhane



North Africans in NLP



GhanaNLP



Radical AI Network



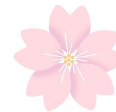
ML Collective



Turkic Interlingua



Americas NLP



Big Science



SISONKE BIOTIK



2013

2017

2019

2020

2021

# Finding Each Other

## DEEP LEARNING FOR LOW-RESOURCE NLP

The second edition of DeepLo will be colocated with [EMNLP 2019](#) in Hong Kong.

The workshop will bring together experts in deep learning and natural language processing whose research focuses on learning with scarce data. Specifically, it will provide attendees with an overview of existing approaches from various disciplines, and enable them to distill principles that can be more generally applicable. It will also discuss the main challenges arising in this setting and outline potential directions for future progress. The target audience consists of researchers and practitioners in related areas.



## LREC Conferences

The International Conference on Language Resources and Evaluation is organised by ELRA biennially with the support of institutions and organisations involved in HLT.

LREC Conferences bring together a large number of people working and interested in HLT.

LREC 1998 LREC 2000 LREC 2002 LREC 2004 LREC 2006 LREC 2008 LREC 2010  
LREC 2012 LREC 2014 LREC 2016 LREC 2018 LREC 2020

# NLP@Deep Learning Indaba

The Natural Language Processing session at the [Deep Learning Indaba 2019](#) takes place on Thursday, 29 August 2019. It is organized by [Herman Kamper](#), [Sebastian Ruder](#), and [Yukosi Marivate](#). We are proud to have an amazing set of speakers for this year!

### Speakers



## The 3rd Workshop on Technologies for MT of Low Resource Languages (LoResMT 2020)

AACL-IJCNLP, China, December 4, 2020

<https://sites.google.com/view/loresmt-2020>

@ AAAL-IJCNLP 2020 (<http://aacl2020.org/>)

# Code Mixing and Switching

Ever Evolving Language

## Code-Mixing in Social Media Text

### The Last Language Identification Frontier?

Amitava Das\* — Björn Gambäck\*\*

\* NITT University, Neemrana, Rajasthan 301705, India  
amitava.santu@gmail.com

\*\* Norwegian University of Science and Technology, 7491 Trondheim, Norway  
gamback@idi.ntnu.no

---

*ABSTRACT. Automatic understanding of noisy social media text is one of the prime present-day research areas. Most research has so far concentrated on English texts; however, more than half of the users are writing in other languages, making language identification a prerequisite for comprehensive processing of social media text. Though language identification has been considered an almost solved problem in other applications, language detectors fail in the social media context due to phenomena such as code-mixing, code-switching, lexical borrowings, Anglicisms, and phonetic typing. This paper reports an initial study to understand the characteristics of code-mixing in the social media context and presents a system developed to automatically detect language boundaries in code-mixed social media text, here exemplified by Facebook messages in mixed English-Bengali and English-Hindi.*

## Focus on LR NLP and LRL

- Work on large language models has mostly been on dominant online languages.
- Efficacy is limited to those who converse in these languages.
- How do we deal with LRL and also domains without much data?
- Initial attempts at Multilingual Models, with their limitations.

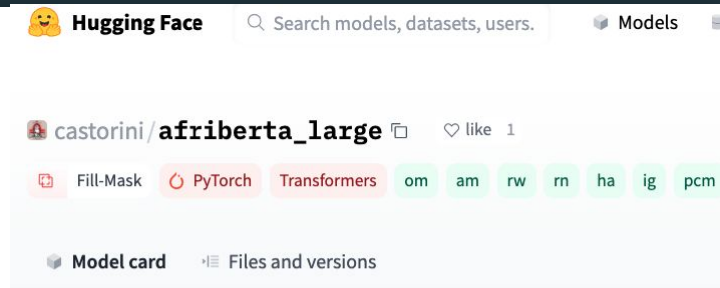
# Language Models

## Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-Resource Languages

**Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin**

David R. Cheriton School of Computer Science  
University of Waterloo

{kelechi.ogueji, yuxin.zhu, jimmylin}@uwaterloo.ca



The screenshot shows the Hugging Face interface for the model 'castorini/afriberta\_large'. At the top, there is a search bar and a 'Models' tab. Below the model name, there are tags for 'Fill-Mask', 'PyTorch', and 'Transformers', along with language tags: 'om', 'am', 'rw', 'rn', 'ha', 'ig', and 'pcm'. There are also buttons for 'Model card' and 'Files and versions'.

afriberta\_large

### Model description

AfriBERTa large is a pretrained multilingual language model with around 126 million parameters. The model has 10 layers, 6 attention heads, 768 hidden units and 3072 feed forward size. The model was pretrained on 11 African languages namely - Afaan Oromoo (also called Oromo), Amharic, Gahuza (a mixed language containing Kinyarwanda and Kirundi), Hausa, Igbo, Nigerian Pidgin, Somali, Swahili, Tigrinya and Yorùbá. The model has been shown to obtain competitive downstream performances on text classification and Named Entity Recognition on several African languages, including those it was not pretrained on.



# Representation

A bigger push to have the internet represent all of us.

## INFORMATION ACCESSIBILITY

AI could make African languages more accessible with machine translation — but people need to make it happen



 (Image: Adobe Stock)

# Data + Compute + People

We have had the rise of the TRIFECTA

# Challenges and Opportunities Modern Approaches in Lower Resource in ML



# Sections of the tutorial

- Representations - word embeddings, multilingual LMs
- Low resource Machine Translation
- Few-shot learning for Low-resource languages

# Representations - word embeddings, multilingual LMs



# How to represent words in Texts?



Set of documents e.g Wikipedia

**Nairobi** is the capital of Kenya.

Nairobi

Nai

ro

bi

n

a

i

r

o

b

What is the **sentence representation**?

How about **representation per word**?

**Too many words**, how about **sub-word representation**?

Why not **Character representation**?

# Discrete Representations

- A word is a **sequence of symbols**,
  - No notion of similarity
  - What is the similarity score between “lift” and “elevator”?
- Words are often **represented as indices** in vocabulary or
- **One-hot vector**, still **no notion of similarity**

- **Continuous representation** is needed to capture notion of similarity

<b>A</b>	1 0 0 0 ... 0 0 ... 0 ...
<b>A1</b>	0 1 0 0 ... 0 0 ... 0 ...
...	....
...	....
<b>Elevator</b>	0 1 0 0 ... 1 0 ... 0 ...
...	....
<b>lift</b>	0 1 0 0 ... 0 0 ... 1 ...
...	....

$$v(\text{" lift "})^T v(\text{" elevator "}) = 0$$

# Continuous Representations

- A **dense vector** of certain dimension (e.g 300d) that capture **syntactic and semantic relationships**.
- Capturing **similarity between words** - > **do well on many semantic-related tasks**
- **How do we learn them?**

## Distributional Hypothesis

*"words that are used and occur in the **same contexts** tend to purport similar meanings" (Harris, 1954)*

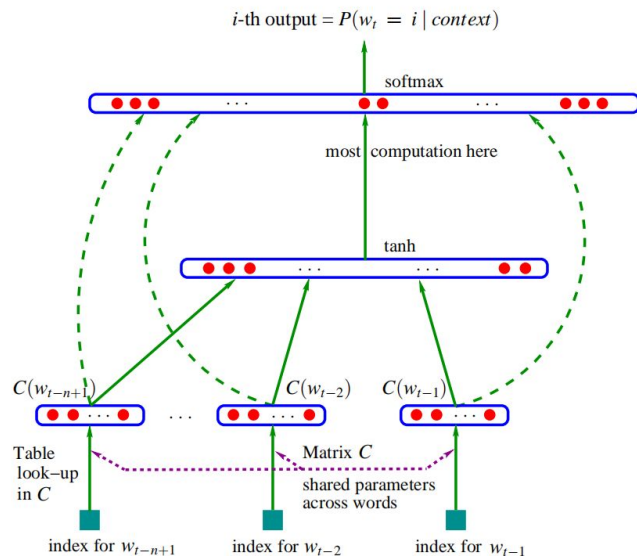
*"You shall know a **word** by the **company it keeps**" (Firth, J. R. 1957)*



# Word Representations (1): FFNN

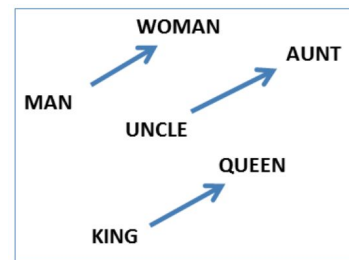
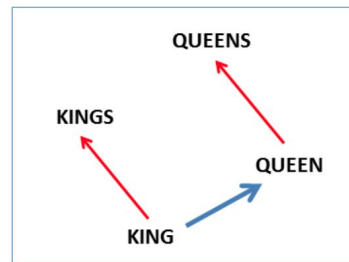
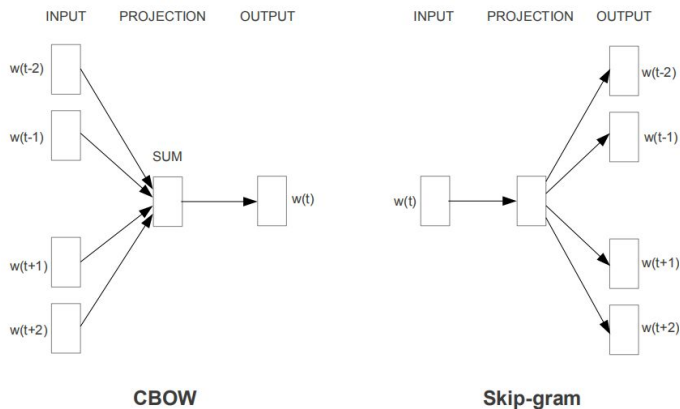
- Bengio et al (2003) proposed **feed forward neural network** for learning word representations
- $P(\text{word} \mid \text{context})$
- **Context** contains **n tokens from the left**
- Training is **computationally expensive**

- The race for the best architecture search begins
  - *Which architecture can best leverage the contexts of words?*



# Word Representations (2): Word2Vec

- Mikolov et al (2013) introduced better architectures capturing distributional hypothesis
  - Make use of *left and right context*
  - Trained on **1 billion tokens**



# Word Representations (3): GloVe

- Word2Vec makes use of only **local (or surrounding) contexts** to learn word vectors
- Trained on **6 billion tokens**
- Pennington et al. (2014) make use of **global co-occurrence counts for context**
  - while maintaining performance on word analogy tasks.
  - $vec(king) - vec(queen) \approx vec(man) - vec(woman)$

- Popularized the use of word vectors to initialize NLP models like
  - Named entity recognition

# Word Representations(4): Capturing SubWords

- **Approach:** train word embeddings jointly for words and sub word units.
- Example Models:
  - **Character Word Embedding** (Chen et al., 2015)
  - **FastText** (Bojanowski et al, 2017):
    - combining character n-grams embedding
    - Multilingual (157 languages)
    - Trained on Common Crawl & Wikipedia

*fast*Text

Library for efficient text classification and representation learning

GET STARTED

DOWNLOAD MODELS

Download pre-trained models

```
010100100
101001001
010010101
001011011
110101110
101001011
```

English word vectors

Pre-trained on English webcrawl and Wikipedia

```
010100100
101001001
010010101
001011011
110101110
101001011
```

Multi-lingual word vectors

Pre-trained models for 157 different languages

# Word Representations: FastText

- Introduced evaluation on **WordSim-353** on **multiple languages**
  - WordSim-353 (Finkelstein et al., 2001): a collection of **353 pairs** (humanly) annotated with **semantic similarity scores in a scale from 0 to 10**
  - **Spearman correlation** between **human scores** and **word embedding similarity scores** for each pair
- Number of **African languages** represented: 7 (Common Crawl)

*What is the **quality of embeddings** on **low-resourced languages**?*

Evaluation on Skip-Gram Model

Language	Word2Vec	FastText
AR	51	<b>55</b>
EN	<b>72</b>	71
ES	57	<b>59</b>
RO	48	<b>54</b>

Result from Bojanowski et al, 2017

# Word Embeddings for Low-resourced languages (1)

- **Word embeddings** are trained on **large corpus (> 1 billion tokens)**
- However, **low resource languages** have smaller unlabelled corpus
- **Evaluation** is mostly on **high-resourced languages** like English
  - *With a lot of downstream tasks*
- Since there are **few/no evaluation dataset** for **low-resourced languages**.

## Word Embeddings for Low-resourced languages (2)

- Alabi et al (2020) investigates the quality of FastText word embeddings on
  - Two African languages: **Yoruba** and **Twi**
  - **Evaluated** on *translated WordSim-353 corpus*
- Unlabelled corpus used for training has some issues
  - Wikipedia: **too small, absent diacritics, mixed dialects**
  - Common Crawl: **often mixed with other languages**

*How does pre-trained models compare to word embeddings from curated corpus?*

# Word Embeddings for Low-resourced languages (3)

Model	Twi		Yorùbá	
	Vocab Size	Spearman $\rho$	Vocab Size	Spearman $\rho$
F1: Pre-trained Model (Wiki)	935	0.143	21,730	0.136
F2: Pre-trained Model (Common Crawl & Wiki)	NA	NA	151,125	0.073
C1: Curated <i>Small</i> Dataset (Clean text)	9,923	0.354	12,268	0.322
C2: Curated <i>Small</i> Dataset (Clean + some noisy text)	18,494	<b>0.388</b>	17,492	0.302
C3: Curated <i>Large</i> Dataset (All Clean + Noisy texts)	47,134	0.386	44,560	<b>0.391</b>

**C1:**

**Yoruba:** 1.6 M tokens  
**Twi:** 735K tokens

**C2:**

**Yoruba:** 2M (+ **noisy** News)  
**Twi:** 742K (+ **noisy** Wiki)

**C3:**

**Yoruba:** 13M tokens  
**Twi:** 15M tokens

Spearman Correlation between human judgements and similarity scores on the wordSim-353 English Correlation: 71



# Why contextualized word embeddings?

- The same **word** can mean several things depending on the **context**
- *Having a single word embedding in all contexts is misleading.*

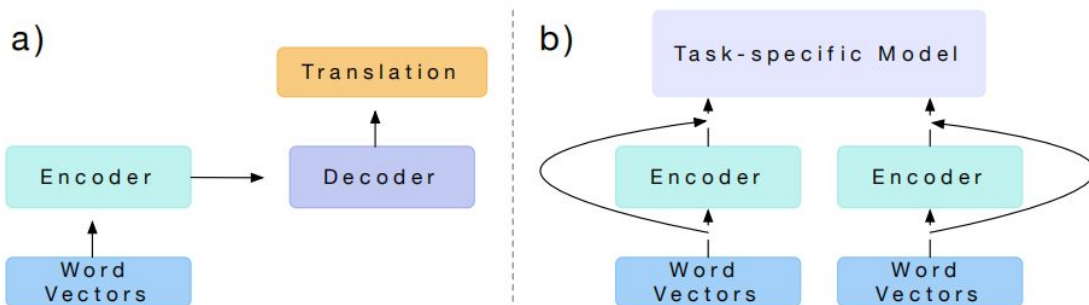
Hey kids, go and **play** in the garden.

Which team do you **play** for?

He learned to **play** the piano at the age of seven.

# Contextualized Word Representations: CoVe

- Inspired by the **transfer learning of CNNs trained on ImageNet** to other tasks in **Computer Vision**
- **CoVe** makes use of **LSTM encoder** trained for **machine translation** task.
  - Words are encoded in context
  - Beginning of contextualized word representation

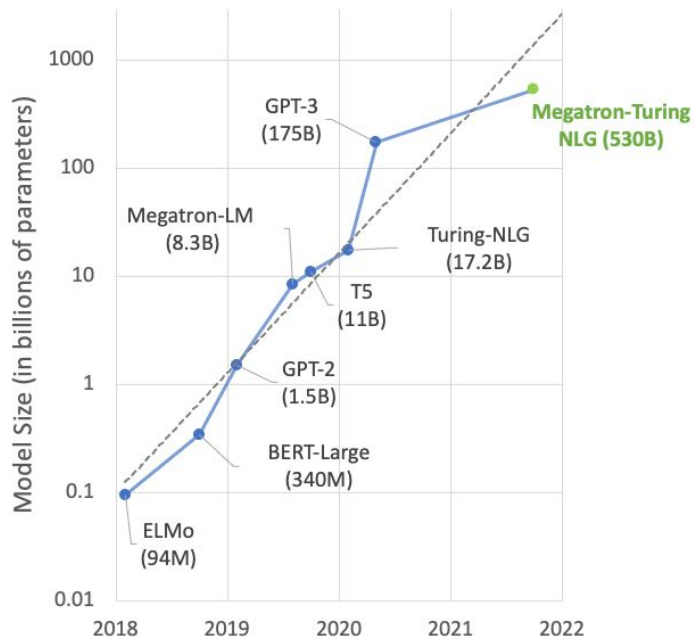


# Contextualized Word Representations: ELMo

- Peters et al (2018) proposed extracting contextualized word representation from **bi-LSTM**
  - Trained on 1 billion word corpus
- **ELMo** uses the **concatenation of independently trained left-to-right and right-to left LSTMs** to generate features

**ELMo** provides **three layers of representations** for each **input token** instead of a fixed-sized representation

# The journey of Pre-trained LM begins



- The **bigger, the better** the LM
- Also, clever **self-supervised techniques** have been developed.
- Some are **multilingual**
  - **How about low-resource languages?**
  - Can they be misused?

# Language Models Needs Better Architecture

---

## Attention Is All You Need

---

**Ashish Vaswani\***  
Google Brain  
avaswani@google.com

**Noam Shazeer\***  
Google Brain  
noam@google.com

**Niki Parmar\***  
Google Research  
nikip@google.com

**Jakob Uszkoreit\***  
Google Research  
usz@google.com

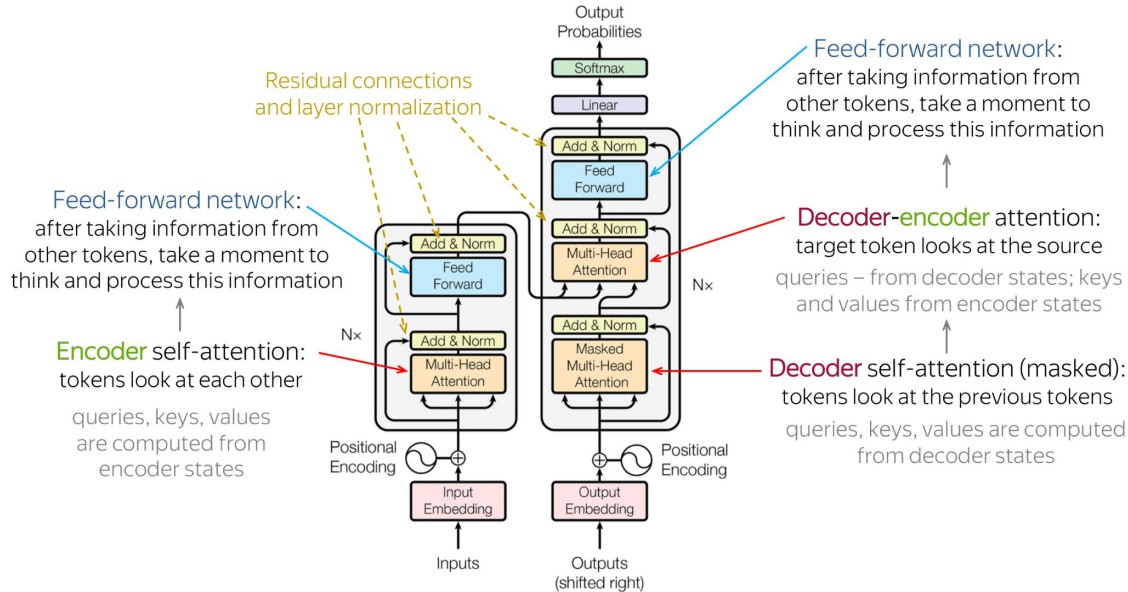
**Llion Jones\***  
Google Research  
llion@google.com

**Aidan N. Gomez\* †**  
University of Toronto  
aidan@cs.toronto.edu

**Lukasz Kaiser\***  
Google Brain  
lukaszkaizer@google.com

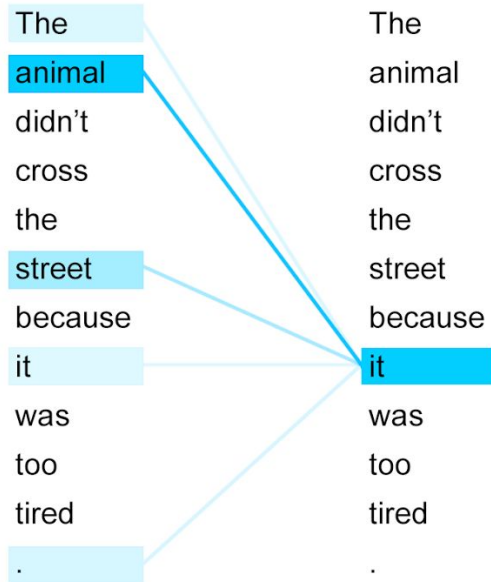
**Illia Polosukhin\* ‡**  
illia.polosukhin@gmail.com

# The Transformer

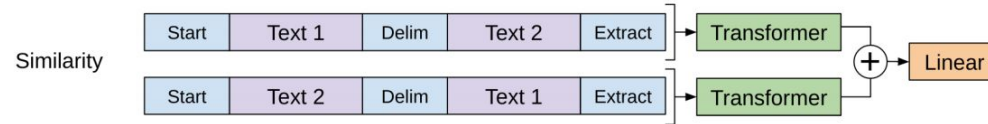
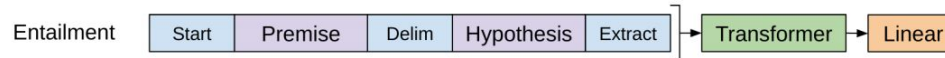
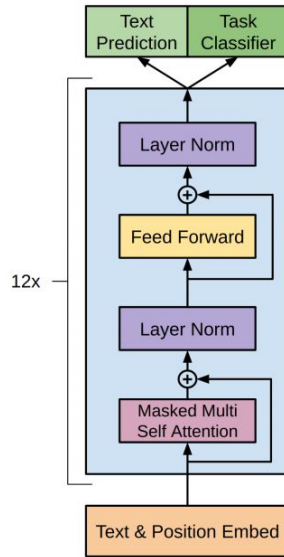


# The Transformer: Attention is all you need

- **Attention** is often combined with **Encoder-Decoder architecture** for **seq2seq** tasks.
- Computing multiple heads of attention *without recurrence or convolution* produces impressive result.
- Highly **parallelizable**
- Suitable for **long-range dependencies**.



# Language Models for Representations: GPT



- Makes use of **Transformer decoder** with 12 layers
- GPT introduced **fine-tuning LM & linear classifier** end-to-end
- Unfortunately, **unidirectional LMs** are sub-optimal for sentence-level NLP tasks.



# Language Models for Representations: BERT

- **BERT**: a **bi-directional masked LM** jointly trained on **left and right contexts**
- Trained on the **Transformer Encoder** architecture
- Introduced two pre-training tasks
  - **Masked LM prediction**: 15% of tokens are masked.
  - **Next Sentence prediction**

Berlin is the [MASK] of Germany

# Language Models for Representations: BERT

- **BERT**: a **bi-directional masked LM** jointly trained on **left and right contexts**
- Trained on the **Transformer Encoder** architecture:
- Introduced two pre-training tasks
  - **Masked LM prediction**: 15% of tokens are masked.
  - **Next Sentence prediction**

Berlin is the [MASK] of Germany

$\text{IsNext}(\text{"dogs are friendly"}, \text{"malaria is deadly"}) = 0$

# BERT established strong baselines

- Pre-trained on
  - **Book corpus** (800M tokens)
  - **Wikipedia** (2.5B tokens)

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT <sub>BASE</sub>	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT <sub>LARGE</sub>	<b>86.7/85.9</b>	<b>72.1</b>	<b>92.7</b>	<b>94.9</b>	<b>60.5</b>	<b>86.5</b>	<b>89.3</b>	<b>70.1</b>	<b>82.1</b>

Table 1: GLUE Test results, scored by the evaluation server (<https://gluebenchmark.com/leaderboard>).

# Improving over BERT: RoBERTa and Others

- **RoBERTa (Liu et al, 2019)**
  - Pre-trained on more texts (from **16G to 160GB** )
  - Removed next sentence prediction pre-training tasks
    - Less important when there are more texts
- **XLNET (Yang et al, 2019):** An autoregressive LM approach to BERT
- **ERNIE (Sun et al, 2019)** - BERT + Knowledge Graph
- **ALBERT (Lan et al, 2019)** - A Lite BERT (reducing memory)
- **ELECTRA (Clark et al, 2020):** trains a *discriminator for replaced token detection* instead of *predicting masked tokens*
- .....

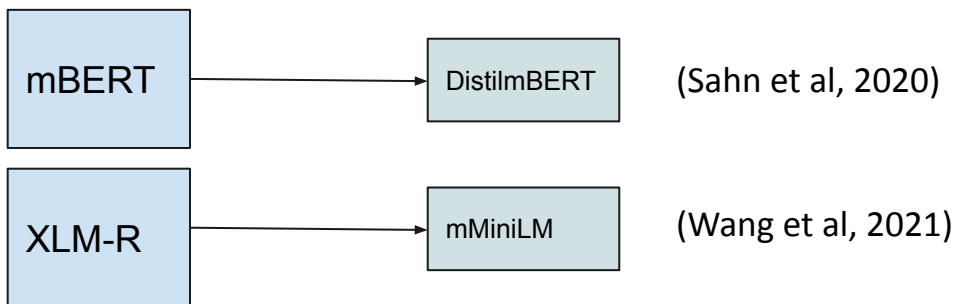
# Multilingual Pre-trained Language Models (1)

Language Model	Largest size (Million)	# languages	# African languages
mBERT (Devlin et al, 2019)	172M	104	3
XLM-RoBERTa (Conneau et al 2020)	559M	100	7
VECO (Luo et al, 2021)	662M	50	2
InfoXLM (Chi et al, 2021)	559M	94	2
ERNIE-M (Ouyang et al, 2021)	559M	96	4
RemBERT (Chung et al, 2021)	559M	110	11

} Most popular

# Multilingual Pre-trained Language Models (2)

- Multilingual pre-trained LMs (PLMs) are **quite big** (>170 million parameters)
  - Most labs in low-resource communities cannot run the model
  - Fine-tuning also takes more time for bigger models.
- **Knowledge distillation** comes to the rescue with minimal drop in performance



# Performance of PLMs on African languages

- What is the performance of multilingual PLMs on African languages:
  - **Named entity recognition**
    - Recognizing entities like personal name, organization, location or date.
    - Evaluation on 10 African languages

The	Emir	of	Kano	turbaned	Zhang	who	has	spent	18	years	in	Nigeria
O	O	O	B-LOC	O	B-PER	O	O		B-DATE	I-DATE	O	B-LOC

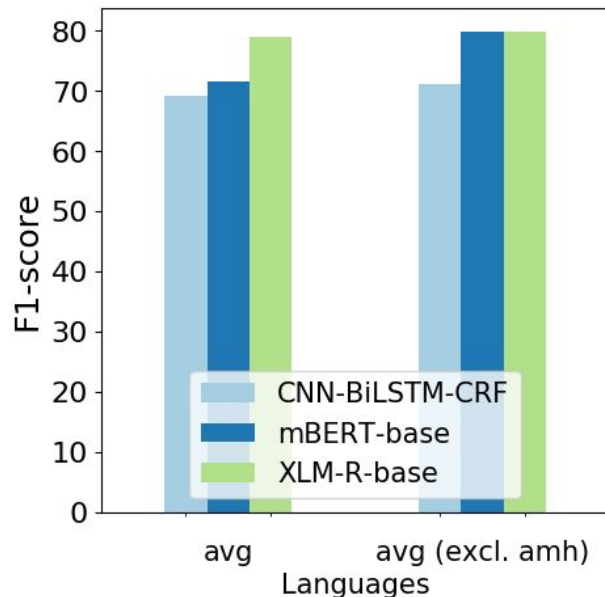
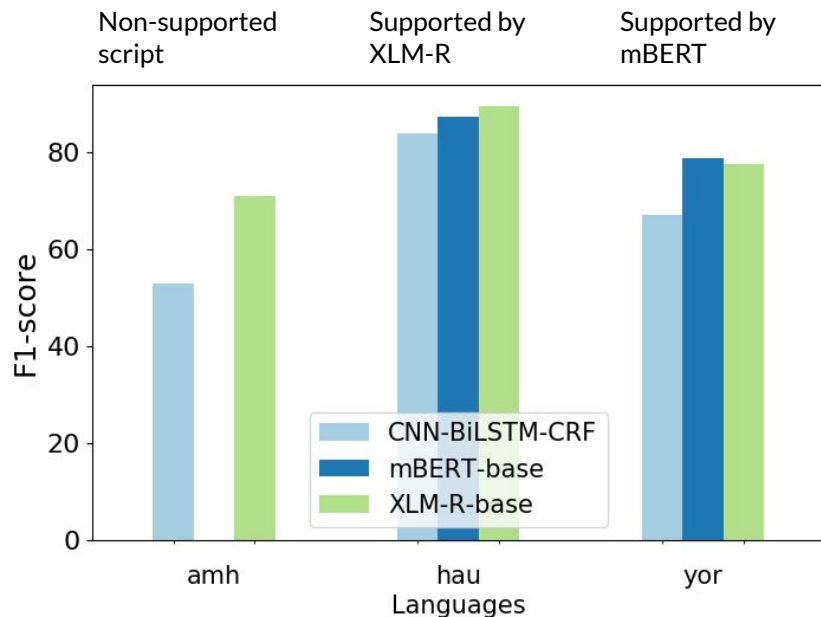
# Performance of PLMs on African languages

- What is the performance of multilingual PLMs on African languages:
  - **Named entity recognition**
    - Recognizing entities like personal name, organization, location or date.
    - MasakhaNER: 10 African languages
  - **News-Topic classification e.g**
    - **Topics like** World, africa, nigeria, politics, sports, health

The	Emir	of	Kano	turbaned	Zhang	who	has	spent	18	years	in	Nigeria
O	O	O	B-LOC	O	B-PER	O	O		B-DATE	I-DATE	O	B-LOC

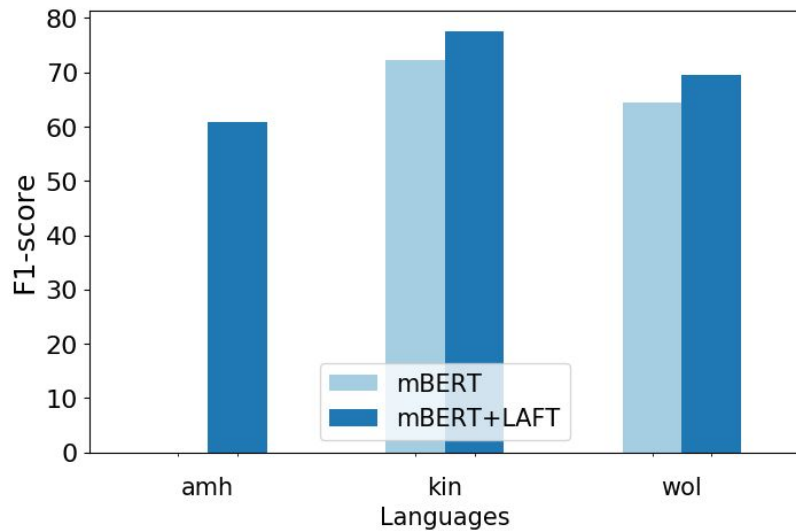


# Performance on Named Entity Recognition



# Language Adaptive Fine-tuning for PLMs

- (optionally swap vocab) fine-tune BERT Masked LM on new language
- Adapt to the downstream task



## LAFT improves over languages

- with **non-supported script**
- **Not seen** by pre-trained LM
- **Morphologically rich** languages
- with **small unlabelled corpora** (40k) for adaptation

# Multilingual PLMs with Small Sized Corpus

- Can we learn multilingual PLMs on small sized corpus?
- Ogueji et al., 2021 pre-train a **multilingual RoBERTa** architecture on **10 African languages**
  - Competitive with state of the art models

LM	Size (GB)	# Lang.	# African Lang.
mBERT	100 GB	104	3
XLM-R	2,395 GB	100	7
AfriBERTa	0.94 GB	10	10

Pre-training size and languages

LM	F1-score
mBERT (172M)	71.61
XLM-R (270M)	78.96
AfriBERTa (126M)	79.10
mBERT + LAFT (172M)	80.69
XLM-R + LAFT (270M)	<b>82.63</b>

NER: Evaluation on 10 African languages

# Multilingual PLMs with Small Sized Corpus

- Can we learn multilingual PLMs on small sized corpus?
- Ogueji et al., 2021 pre-train a **multilingual RoBERTa** architecture on **10 African languages**
  - Competitive with state of the art models

LM	Size (GB)	# Lang.	# African Lang.
mBERT	100 GB	104	3
XLm-R	2,395 GB	100	7
AfriBERTa	0.94 GB	10	10

Pre-training size and languages

LM	hau F1	yor F1
mBERT (172M)	83.03	71.61
XLm-R (270M)	85.62	71.07
AfriBERTa (126M)	90.86	<b>83.22</b>
mBERT + LAFT (172M)	<b>90.98</b>	79.11
XLm-R + LAFT (270M)	90.87	79.19

News Topic Classification: hau & yor

# Pre-trained LMs for Text Generation

- Several PLMs have also been developed for **Text generation tasks** such as
  - **Question & answering, Summarization and machine translation**
  - They are BERT-like, however they use **encoder-decoder** architecture

Objective	Inputs	Targets
Prefix LM e.g GPT-*	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party <b>apple</b> week .	<i>(original text)</i>
MASS/BART-style	Thank you <M> <M> me to your party <M> week .	<i>(original text)</i>
T5-Style	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>

## Pre-training approaches

# Multilingual Pre-trained LM for Text Generation

- **Based on BART**
  - **mBART25** - 25 languages (**0 African language**)
  - **mBART50** - 50 languages (**2 African languages**)
  - **M2M-100** - Machine translation for **100** languages (**15 African languages**)
- **Based on T5:**
  - **mT5** - multilingual T5 for **101** languages (**11 African languages**)
  - **byT5** - **byte level** multilingual T5
  - **CharFormer** - **Token-free** character-level multilingual T5

# mT5 for Low-resource Machine Translation

- Comparing **Supervised MT** with fine-tuning **mT5-base** model (**580M** parameters)
- Fine-tuning **mT5 is competitive or better** with few training epochs.

Domain	Number of Sentences		
	Train. Set	Dev. Set	Test Set
<b>MENYO-20k</b>	10,070	3,397	6,633
<b>Bible</b>	30,760	–	–
<b>JW300</b>	459,871	–	–
<i>TOTAL</i>	500,701	3,397	6,633

Number of Training sentences. Evaluation on MENYO-20k: A multi-domain corpus

Model	en-yo	yo-en
Supervised MT	10.9 ± 0.3	14.0 ± 0.3
mT5-base FT	<b>11.5 ± 0.3</b>	<b>16.3 ± 0.4</b>

BLEU score on English-Yoruba (en-yo)

# Conclusion

- We discussed **static word embeddings** and **contextualized embedding** and their use in NLP
- We also discussed the **quality issues of massively trained pre-trained word embeddings**
- We further introduced **pre-trained language models** and their **multilingual variants**
  - Illustrating their performance through through transfer learning on **NER** and **text classification**
- **Multilingual pre-trained models** only **supports few** low-resourced languages
  - Also, the **huge parameter size makes them difficult to fine-tune**
  - **Knowledge distillation helps to compress** the models



# Low resource Machine Translation



# TLDR

ML/Statistical Models have led to huge gains in MT -> Neural Machine Translation (NMT)

Build tools that allow us to go from one language to another.

- Hmm, what about from low resource to high resource (access to more NLP tools)?

Not simply a matter of getting more data (parallel and monolingual)

# The NLP Task: Translation

Today, is a good day [en: English]

- Namunthla, i siku ra kahle [tso: Xitsonga]
- Namunthla, i siku ro saseka [tso: Xitsonga]

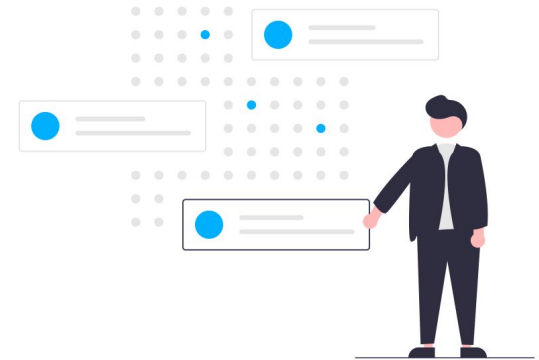
# Rule Based Translation - Classical Approach

A bilingual dictionary [Needs to have been compiled]

Grammatical Rules [Needs to have been studied]

Translation Rules [Needs to have been developed]

**Resources:** Money, People, Time and Archives.



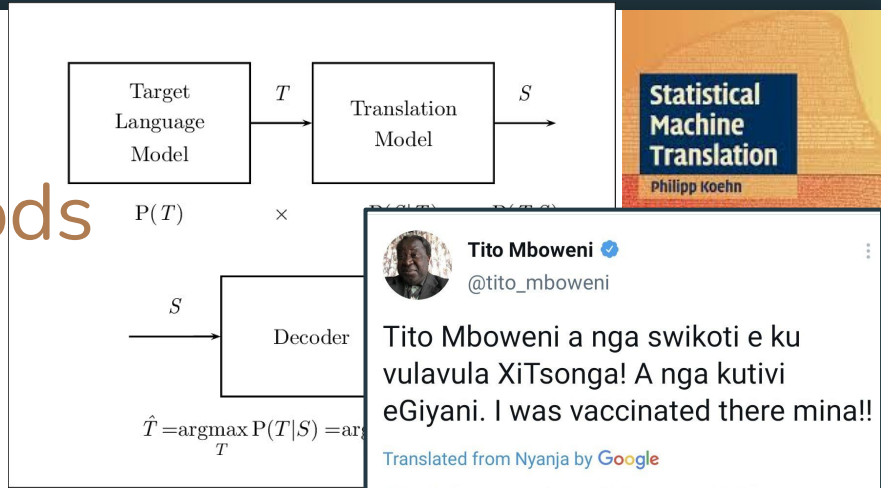
# Enter Statistical Methods

$$P(T|S)$$

Learn the probability distribution from parallel data

- Word Based
- Phrase Based
- Approach can generalise

Costs to corpora creation.  
Errors hard to catch.



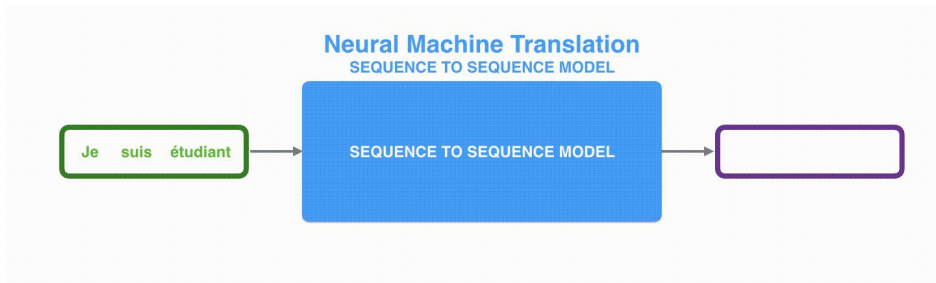
A screenshot of a Twitter post by Tito Mboweni (@tito\_mboweni). The tweet reads: "Tito Mboweni a nga swikoti e ku vulavula XiTsonga! A nga kutivi eGiyani. I was vaccinated there mina!!". Below the text is a blue link: "Translated from Nyanja by Google". The tweet also includes a translated version: "Tito Mboweni doesn't have a skirt to say XiTsonga! He did not hear from Giyani. I was vaccinated there myself !!". The post features four small images showing people in a community setting, some wearing high-visibility vests. The timestamp is "20:51 · 31 Jul 21 · Twitter for iPhone".

# The Modern Age - How Deep Can you Go?

---

## Sequence to Sequence Learning with Neural Networks

---



**Ilya Sutskever**  
Google  
ilyasu@google.com

**Oriol Vinyals**  
Google  
vinyals@google.com

**Quoc V. Le**  
Google  
qvl@google.com

### Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

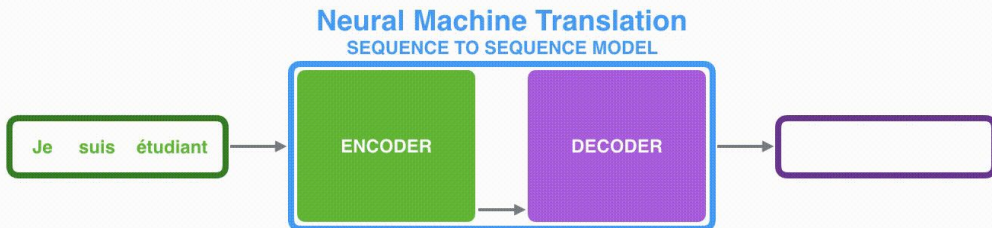
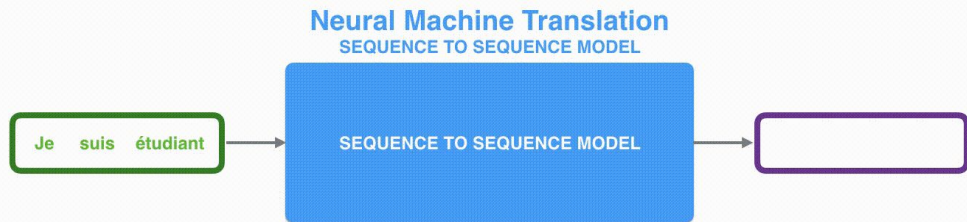
**Kyunghyun Cho**  
**Bart van Merriënboer** **Caglar Gulcehre**  
Université de Montréal  
firstname.lastname@umontreal.ca

**Dzmitry Bahdanau**  
Jacobs University, Germany  
d.bahdanau@jacobs-university.de

**Fethi Bougares** **Holger Schwenk**  
Université du Maine, France  
firstname.lastname@lium.univ-lemans.fr

**Yoshua Bengio**  
Université de Montréal, CIFAR Senior Fellow  
find.me@on.the.web

# What is in a Sequence?



## Sequence to Sequence Learning with Neural Networks

Ilya Sutskever  
Google  
ilyasu@google.com

Oriol Vinyals  
Google  
vinyals@google.com

Quoc V. Le  
Google  
qvl@google.com

## Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

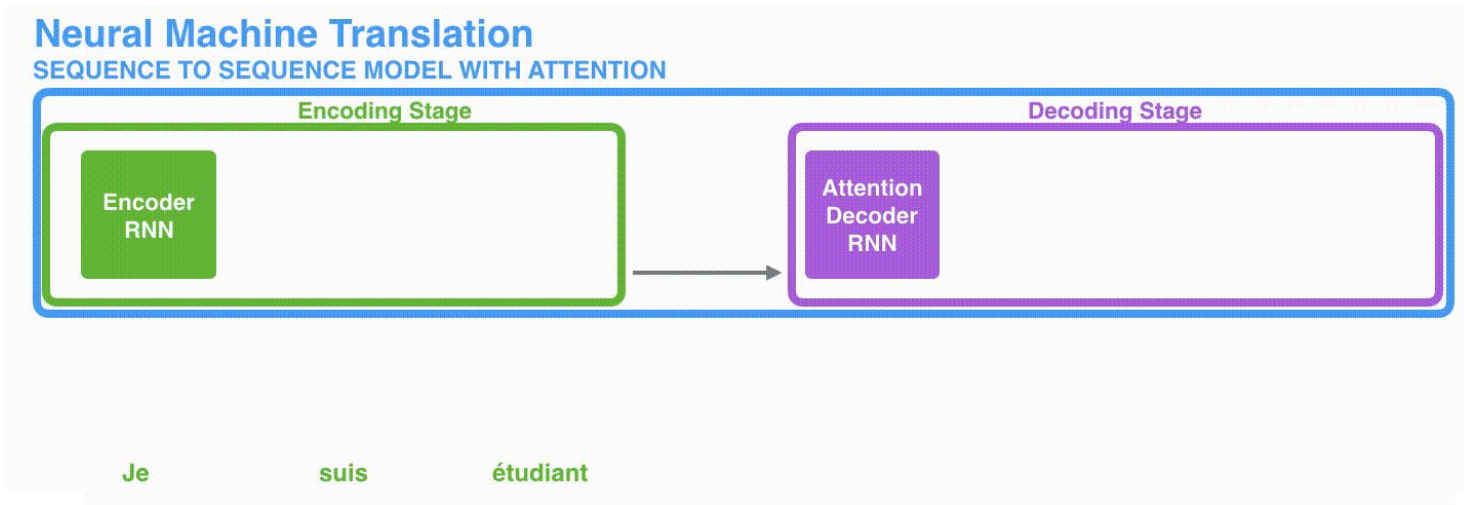
Kyunghyun Cho  
Bart van Merriënboer Caglar Gulcehre  
Université de Montréal  
firstname.lastname@umontreal.ca

Dzmitry Bahdanau  
Jacobs University, Germany  
d.bahdanau@jacobs-university.de

Fethi Bougares Holger Schwenk  
Université du Maine, France  
firstname.lastname@lium.univ-lemans.fr

Yoshua Bengio  
Université de Montréal, CIFAR Senior Fellow  
find.me@on.the.web

# Recurrent Neural Networks with Attention





# Things Fall Apart

- Not enough parallel data [D]
- Expertise to extract data or train models [E]
- Focus and Benchmarks [B]

# Where is the Data?

Lack of Parallel Data

Parallel data is skewed

- Religious or Official Texts

Cost of translation is a real barrier

# The focus on English

Much of translation is en->

HRL to LRL

What about LRL-LRL?

# Human Evaluation and Evaluation Metrics

- Human Evaluation [**Gold Standard**]
- Round Trip Translation [**Heuristic**]
- **BLEU** (bilingual evaluation understudy)
- **METEOR** (Metric for Evaluation of Translation with Explicit ORdering)
- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation)

METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output

Abhaya Agarwal and Alon Lavie  
Language Technologies Institute  
Carnegie Mellon University  
Pittsburgh, PA, 15213, USA  
{abhayaa,alavie}@cs.cmu.edu

# LRL: So what do we do? Our Approach Toolbelt

Data Innovations

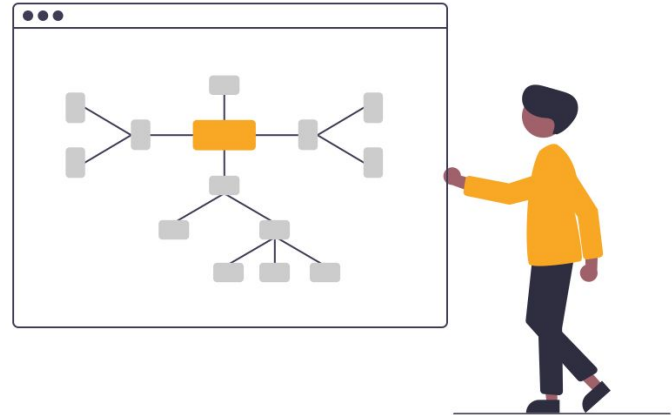
Augmentation

Multilingual/Cross Lingual Models

Transfer Learning/ LLM

Benchmarks

The big Picture



# Approach: Synthetic Data Creation

Corpus
GoURMET v1
SAWA
Tanzil v1
GV v2017q3
GV v2015
Ubuntu v14.10
EUbookshop v2
GNOME v1
total

**Table 1:** Parallel English NMT systems described in the Voices corpus.

1. With the best identified hyper-parameters for each direction we built a system using only parallel data.
2. en and sw monolingual data were back-translated with the systems built in the previous step.
3. Systems in both directions were trained on the combination of the back-translated data and the parallel data.
4. Steps 2–3 were re-executed 3 more times. Back-translation in step 2 was always carried out with the systems built in the most recent execution of step 3, hence the quality of the system used for back-translation improved with each iteration.

## An English–Swahili parallel corpus and its use for neural machine translation in the news domain

Felipe Sánchez-Martínez,<sup>†</sup> Víctor M. Sánchez-Cartagena,<sup>‡</sup> Juan Antonio Pérez-Ortiz,<sup>‡</sup> Mikel L. Forcada,<sup>†</sup> Miquel Esplà-Gomis,<sup>†</sup> Andrew Secker,<sup>†</sup> Susie Coleman,<sup>†</sup> Julie Wall<sup>†</sup>

<sup>†</sup>Dep. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant  
E-03690 Sant Vicent del Raspeig (Spain)  
{fsanchez, vmsanchez, japerez, mlf, mespla}@dlsi.ua.es

<sup>‡</sup>The British Broadcasting Corporation  
BBC Broadcasting House, Portland Place, London, W1A 1AA. (UK)  
{andrew.secker, susie.coleman, julie.wall}@bbc.co.uk

## Synthetic Parallel Data

Corpus	Sent's	Tokens
NewsCrawl (en)	18 113 311	359 823 264
NewsCrawl (sw)	174 425	3 603 035
GoURMET (sw)	5 687 000	174 867 482

**Table 2:** Monolingual Swahili and English corpora used to build synthetic parallel data through back-translation.

# Approach: Identify Linguistic Differences

## Incorporating Information about Linguistic Differences

Feature	Value in English	Value in Swahili	Examples
Coding of plurality in nouns	Plural suffix	Plural prefix	<i>kichwa</i> ('head'), <i>vichwa</i> ('heads'); <i>jicho</i> ('eye'), <i>macho</i> ('eyes')
Number of categories encoded in a single-word verb	Few (number, person, tense)	Many ("STROVE", that is, number and person of subject, tense, aspect and mood, optional relatives, number and person of object, verb root, and optional extensions)	<i>nimekinunua kitabu</i> 'I have bought the book', where: <i>ni</i> 'I', subject; <i>me</i> , present perfect; <i>ki</i> , 'it', object; <i>nunua</i> , 'buy', verb root.
Definite articles	Definite word distinct from demonstrative	Demonstrative (selectivity) used as definite article	<i>kitabu</i> ('book', 'the book', 'a book').
Noun Phrase Conjunction	<i>And</i> different from <i>with</i>	<i>And</i> identical to <i>with</i>	<i>Lete chai na maziwa</i> ('Bring tea and milk'); <i>Yesu alikuja na Baba yake</i> ('Jesus came with his Father').

**Table 3:** A summary of linguistic contrasts between English and Swahili.

Strategy	it.	BLEU	chrF++
en→sw			
only parallel	-	22.23	46.34
iter. backt.	1	25.59	50.08
iter. backt.	2	26.22	50.91
iter. backt.	3	26.36	51.09
iter. backt.	4	26.58	51.39
+ NewsCrawl	4	26.77	51.46
+ NewsCrawl + tags	4	27.42	52.11
<i>Google Translate</i>	-	23.24	48.80
sw→en			
only parallel	-	22.66	44.62
iter. backt.	1	29.29	51.19
iter. backt.	2	29.70	51.82
iter. backt.	3	29.99	51.98
iter. backt.	4	30.19	52.10
+ tags	4	30.55	52.72
<i>Google Translate</i>	-	30.36	53.32

**Table 5:** Automatic evaluation results obtained for the different development steps of the MT systems: *only parallel* stands for the systems trained only on parallel data with the best hyper-parameters; *iter. backt.* represents systems obtained after iteratively back-translating monolingual data (iteration number is shown in column *it.*); *+NewsCrawl* means that the <sub>sw</sub> NewsCrawl corpus was back-translated and added; and *+tags* indicates that TL linguistic tags were interleaved.

Mikel Artetxe, Gorka Labaka & Eneko Agirre  
 IXA NLP Group  
 University of the Basque Country (UPV/EHU)  
 {mikel.artetxe, gorka.labaka, e.agirre}@ehu.eus

Kyunghyun Cho  
 New York University  
 CIFAR Azrieli Global Scholar  
 kyunghyun.cho@nyu.edu

# Approach: Unsupervised Machine Translation

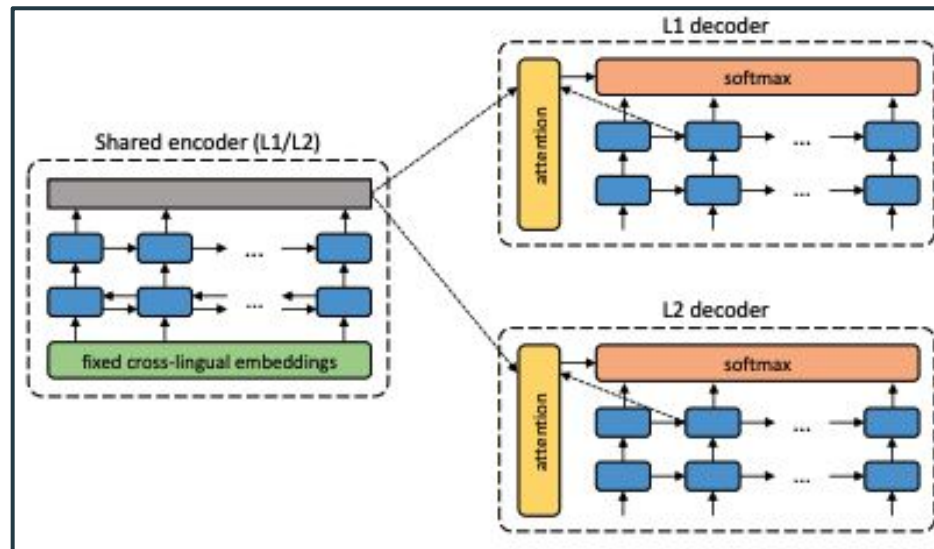
Architecture of the proposed system. For each sentence in language L1,

the system is trained alternating two steps: denoising, which optimizes the probability of encoding a noised version of the sentence with the shared encoder and reconstructing it with the L1 decoder, and on-the-fly backtranslation.

which translates the sentence in inference mode (encoding it with the shared encoder and decoding it with the L2 decoder) and then optimizes the probability of encoding this translated sentence with the shared encoder and recovering the original sentence with the L1 decoder.

Training alternates between sentences in L1 and L2, with analogous steps for the latter.

**Paper:** *Unsupervised Neural Machine Translation* [URL](#)





# PidginUNMT: Unsupervised Neural Machine Translation from West African Pidgin to English

Kelechi Ogueji  
InstaDeep  
k.ogueji@instadeep.com

Orevaoghene Ahia  
InstaDeep  
o.ahia@instadeep.com

## Approach: Unsupervised Machine Translation

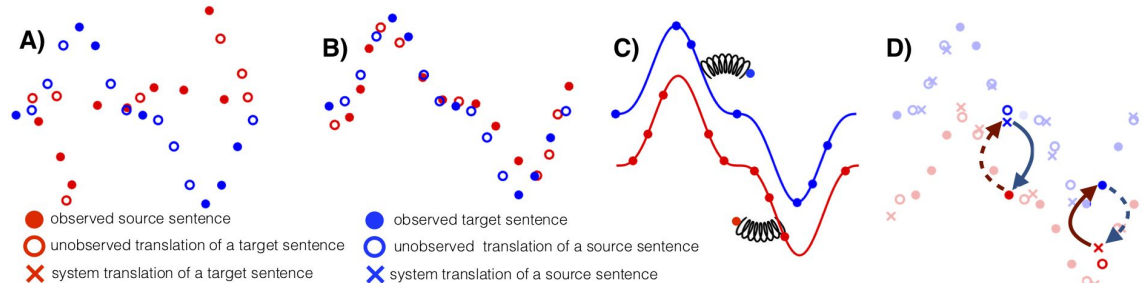
Trained Cross Lingual Embeddings

Utilised the UnsupervisedMT Library

Pidgin to English:  
English to Pidgin:

Source	what are most people today not aware of ?
Reference	wetin many people today no know ?
Model Translation	wetin most people are today no dey aware of
Source	one student began coming to the kingdom hall .
Reference	one of my student come start to come kingdom hall .
Model Translation	one student wey begin dey come di kingdom hall .

Table 3: Model Translation Results from English to Pidgin



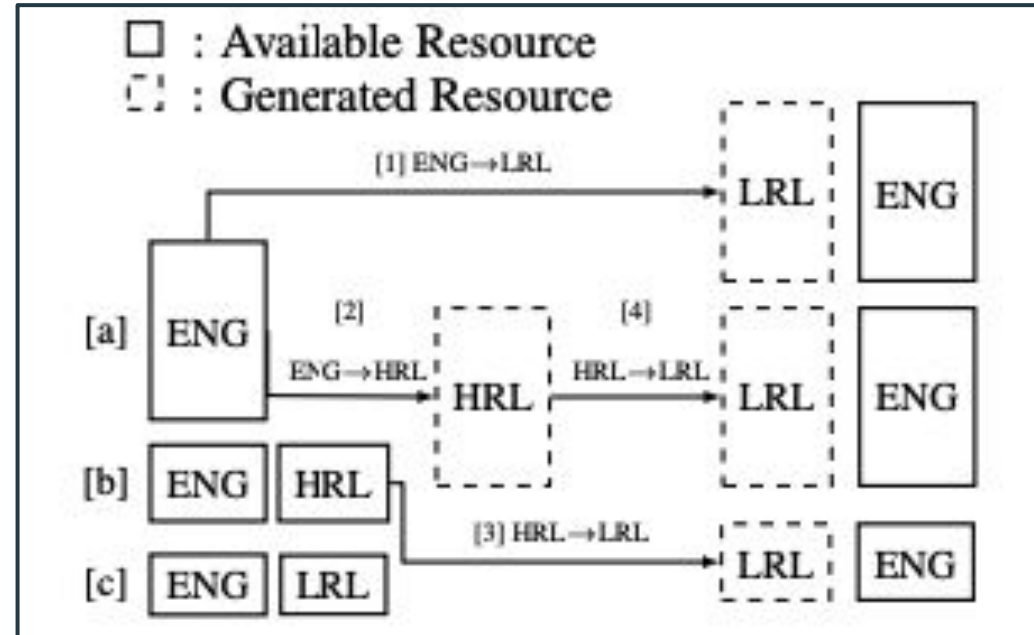
**Paper:** *PidginUNMT: Unsupervised Neural Machine Translation from West African Pidgin to English* [URL](#)  
**UMT System Paper:** *Phrase-Based & Neural Unsupervised Machine Translation* [URL](#)

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, Graham Neubig  
 Language Technologies Institute, Carnegie Mellon University  
 {mengzhox, xiangk, aanastas, gneubig}@andrew.cmu.edu

# Approach: Augmentation

Exploiting in between HRL to generate new parallel data.

Figure 1: With a low-resource language (LRL) and a related high-resource language (HRL), typical data augmentation scenarios use any available parallel data [b] and [c] to back-translate English monolingual data [a] and generate parallel resources ([1] and [2]). We additionally propose scenarios [3] and [4], where we pivot through HRL in order to generate a LRL-ENG resource.



# Approach: Augmentation

Marzieh Fadaee    Arianna Bisazza    Christof Monz  
 Informatics Institute, University of Amsterdam  
 Science Park 904, 1098 XH Amsterdam, The Netherlands  
 {m.fadaee, a.bisazza, c.monz}@uva.nl

**Targeted words selection:** Following common practice, our NMT system limits its vocabulary  $V$  to the  $v$  most common words observed in the training corpus. We select the words in  $V$  that have fewer than  $R$  occurrences and use this as our targeted rare word list  $V_R$ .

**Rare word substitution:** If the LM suggests a rare substitution in a particular context, we replace that word and add the new sentence to the training data. Formally, given a sentence pair  $(S, T)$  and a position  $i$  in  $S$  we compute the probability distribution over  $V$  by the forward and backward LMs and select rare word substitutions  $\mathcal{C}$  as follows:

$$\vec{\mathcal{C}} = \{s'_i \in V_R : \text{topK } P_{\text{ForwardLM}_S}(s'_i | s_1^{i-1})\}$$

$$\overleftarrow{\mathcal{C}} = \{s'_i \in V_R : \text{topK } P_{\text{BackwardLM}_S}(s'_i | s_n^{i+1})\}$$

$$\mathcal{C} = \{s'_i | s'_i \in \vec{\mathcal{C}} \wedge s'_i \in \overleftarrow{\mathcal{C}}\}$$

Source	der tunnel hat einen querschnitt von 1,20 meter höhe und 90 zentimeter breite .
Baseline translation	the wine consists of about 1,20 m and 90 of the canal .
TDA <sub>r=1</sub> translation	the tunnel has a UNK measuring meters 1.20 metres high and 90 <b>centimetres</b> wide .
Reference	the tunnel has a cross - section measuring 1.20 metres high and 90 centimetres across .
Examples of augmented data for the word <i>centimetres</i>	<ul style="list-style-type: none"> <li>• the average speed of cars and buses is therefore around 20 [kilometres / <b>centimetres</b>] per hour .</li> <li>• grab crane in special terminals for handling capacities of up to 1,800 [tonnes / <b>centimetres</b>] per hour .</li> <li>• all suites and rooms are very spacious and measure between 50 and 70 [m / <b>centimetres</b>]</li> <li>• all we have to do is lower the speed limit everywhere to five [kilometers / <b>centimetres</b>] per hour .</li> </ul>

Table 3: An example from newstest2014 illustrating the effect of augmenting rare words on generation during test time. The translation of the baseline does not include the rare word *centimetres*, however, the translation of our TDA model generates the rare word and produces a more fluent sentence. Instances of the augmentation of the word *centimetres* in training data are also provided.

Barret Zoph<sup>1</sup>, Deniz Yuret<sup>2</sup>, Jonathan May<sup>1</sup>, Kevin Knight<sup>3</sup>

<sup>1</sup>Information Sciences Institute, University of Southern California  
 {zoph, jonmay}@isi.edu

<sup>2</sup>Computer Engineering, Koç University  
 dyuret@ku.edu.tr

<sup>3</sup>Information Sciences Institute &  
 Computer Science Department, University of Southern California  
 knight@isi.edu

## Approach: Transfer Learning

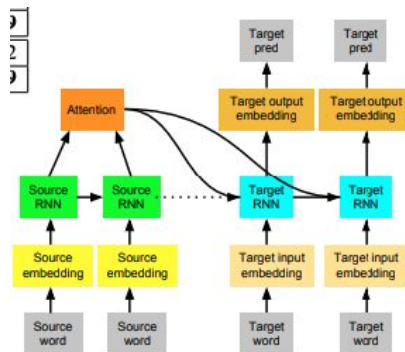


Figure 2: Our NMT model architecture, showing six blocks of parameters, in addition to source/target words and predictions. During transfer learning, we expect the source-language related blocks to change more than the target-language related blocks.

Language	Train size	Test size	SBMT BLEU	NMT BLEU
Hausa	1.0m	11.3K	23.7	16.8
Turkish	1.4m	11.6K	20.4	11.4
Uzbek	1.8m	11.5K	17.9	10.7
Urdu	0.2m	11.4K	17.9	5.2

Table 1: NMT models with attention are outperformed by standard string-to-tree statistical MT (SBMT) when translating low-resource languages into English. Train/test bitext corpus sizes are given in word tokens on the English side. Single-reference, case-insensitive BLEU scores are given for held-out test corpora.

# Approach: Benchmarks

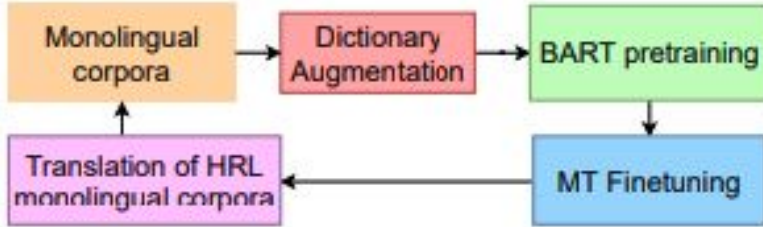


Figure 2: Iterative approach to pretraining using pseudo monolingual data and dictionaries

Direction	En-Run		En-Zu		En-Af		En-Xh	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
Random	22.92	51.89	34.84	65.54	48.33	68.11	24.36	52.91
mNMT	21.53	50.62	31.53	62.95	43.39	64.73	22.28	54.81
AfroBART Baseline	24.33	52.87	<b>35.59</b>	66.14	49.09	68.54	25.65	58.09
AfroBART-Dictionary	24.42	53.22	35.48	66.16	49.25	68.75	25.77	58.15
AfroBART	<b>24.62</b>	<b>53.24</b>	35.58	<b>66.30</b>	<b>49.80</b>	<b>69.03</b>	<b>25.80</b>	<b>58.22</b>

Direction	En-Ln		En-Bem		En-St		En-Sw	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
Random	28.23	52.62	18.96	45.85	43.04	62.68	33.61	58.56
mNMT	27.29	53.16	18.54	46.20	40.26	60.65	30.55	56.44
AfroBART Baseline	29.12	54.31	20.07	47.50	43.79	63.22	34.19	59.08
AfroBART-Dictionary	29.13	54.40	20.48	47.69	43.74	63.33	34.30	59.08
AfroBART	<b>29.46</b>	<b>54.68</b>	<b>20.60</b>	<b>48.00</b>	<b>43.87</b>	<b>63.42</b>	<b>34.36</b>	<b>59.11</b>

Table 2: Results on AFROMT’s En-XX Machine Translation

Laura Martinus

Explore Data Science Academy, South Africa  
laura@explore-ai.net

Jade Z. Abbott

Retro Rabbit, South Africa  
jabbott@rettorabbit.co.za

## Approach: Benchmarks

Table 3: **English to isiZulu Translations:** We show the reference translation, translation by the Transformer model, and translation back to English performed by an isiZulu speaker.

Source	Note that the funds will be held against the Vote of the Provincial Treasury pending disbursement to the SMME Fund .
Target	Lemali izohlala emnyangweni wezimali .
Transformer	Qaphela ukuthi izimali zizobanjwa kweVME esifundazweni saseTreasury zezifo ezithunyelwa ku-MSE .
Back Translation	Be aware that the money will be held by VME with facilities of Treasury with diseases sent to MSE .

Table 4: **English to Xitsonga Translations:** We show the reference translation, translation by the Transformer model, and translation back to English performed by a Xitsonga speaker.

Source	we are concerned that unemployment and poverty persist despite the economic growth experienced in the past 10 years .
Target	hi na swivilelo leswaku mpfumaleko wa mitirho na vusweti swi ya emahlweni hambileswi ku nga va na ku kula ka ikhonomi eka malembe ya 10 lawa ya hundzeke .
Transformer	hi na swivilelo leswaku mpfumaleko wa mitirho na vusweti swi ya emahlweni hambileswi ku nga va na ku kula ka ikhonomi eka malembe ya 10 lawa ya hundzeke .
Back Translation	We have concerns that there is still lack of jobs and poverty even though there has been economic growth in the past 10 years.

Surafel M. Lakew<sup>†+</sup>, Matteo Negri<sup>+</sup> & Marco Turchi<sup>+</sup>

<sup>†</sup>University of Trento, Trento, Italy

<sup>+</sup>Fondazione Bruno Kessler, Trento, Italy

<sup>†</sup>{name.surname}@unitn.it, <sup>+</sup>{surname}@fbk.eu

## Approach: Benchmarks

Table 1: BLEU scores for the SATOS ↔ En directions, domain-specific best performing results are highlighted for each direction, whereas bold shows the overall best in terms of the AVG score.

Model	Domain	Sw-En		Am-En		Ti-En		Om-En		So-En	
		En	Sw	En	Am	En	Ti	En	Om	En	So
S-NMT	Jw300	48.71	47.58	32.86	25.72	29.89	25.54	26.92	23.38		
	Bible			30.35	23.36					29.87	24.64
	Tanzil	18.83	31.67	11.71	5.71					8.51	2.46
	Ted	16.63	11.92	4.26	1.32					1.35	0.39
	AVG	<b>28.06</b>	<b>30.39</b>	19.80	14.03	29.89	25.54	26.92	23.38	13.38	9.16
SS-NMT	Jw300	48.90	47.45	32.76	26.54	29.84	25.99	26.45	23.47		
	Bible			30.53	24.21					27.68	22.89
	Tanzil	19.44	32.17	12.55	7.29					6.75	2.25
	Ted	18.62	14.72	6.92	1.41					1.21	0.52
	AVG	28.99	<b>31.45</b>	20.69	<b>14.86</b>	29.84	25.99	26.45	23.47	11.88	8.55
TL	Jw300	48.74	47.39	32.95	26.49	29.81	26.47	27.77	24.54		
	Bible			30.36	24.26					32.07	27.67
	Tanzil	19.9	31.78	12.28	7.34					10.14	3.34
	Ted	19.74	14.81	7.42	1.31					1.57	0.56
	AVG	<b>29.46</b>	31.33	20.75	14.85	29.81	<b>26.47</b>	27.77	24.54	14.73	10.52
M-NMT	Jw300	46.62	44.47	33.21	24.39	32.21	26.4	32.24	24.96		
	Bible			29.78	20.01					34.99	28.76
	Tanzil	18.75	24.22	13.68	10.95					12.68	3.73
	Ted	17.54	14.65	6.78	1.32					3.09	1.01
	AVG	27.64	27.78	<b>20.86</b>	14.17	<b>32.21</b>	26.40	<b>32.24</b>	<b>24.96</b>	<b>16.92</b>	<b>11.17</b>

# The Full Picture

It takes a village.

## Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages

∇\*, Wilhelmina Nekoto<sup>1</sup>, Vukosi Marivate<sup>2</sup>, Tshinondiwa Matsila<sup>1</sup>, Timi Fasubaa<sup>3</sup>,  
Tajudeen Kolawole<sup>4</sup>, Taiwo Fagbohunge<sup>5</sup>, Solomon Oluwole Akinola<sup>6</sup>,  
Shamsuddee Hassan Muhammad<sup>7,39</sup>, Salomon Kabongo<sup>4</sup>, Salomey Osei<sup>4</sup>,  
Sackey Freshia<sup>8</sup>, Rubungo Andre Niyongabo<sup>9</sup>, Ricky Macharm<sup>10</sup>, Perez Ogayo<sup>11</sup>,  
Orevaoghene Ahia<sup>12</sup>, Musie Meressa<sup>13</sup>, Mofe Adeyemi<sup>14</sup>, Masabata Mokgesi-Selinga<sup>15</sup>,  
Lawrence Okegbemi<sup>5</sup>, Laura Jane Martinus<sup>16</sup>, Kolawole Tajudeen<sup>4</sup>, Kevin Degila<sup>17</sup>,  
Kelechi Ogueji<sup>12</sup>, Kathleen Siminyu<sup>18</sup>, Julia Kreutzer<sup>19</sup>, Jason Webster<sup>20</sup>,  
Jamiil Toure Ali<sup>1</sup>, Jade Abbott<sup>21</sup>, Iroro Orife<sup>3</sup>, Ignatius Ezeani<sup>38</sup>,  
Idris Abdulkabir Dangana<sup>23,7</sup>, Herman Kamper<sup>24</sup>, Hady Elsahar<sup>25</sup>, Goodness Duru<sup>26</sup>,  
Ghollah Kioko<sup>27</sup>, Espoir Murhabazi<sup>1</sup>, Elan van Biljon<sup>12,24</sup>, Daniel Whitenack<sup>28</sup>,  
Christopher Onyefuluchi<sup>29</sup>, Chris Emezue<sup>40</sup>, Bonaventure Dossou<sup>31</sup>, Blessing Sibanda<sup>32</sup>,  
Blessing Ito Bassey<sup>4</sup>, Ayodele Olabiyi<sup>33</sup>, Arshath Ramkilowan<sup>34</sup>, Alp Öktem<sup>35</sup>,  
Adewale Akinfaderin<sup>36</sup>, Abdallah Bashir<sup>37</sup>

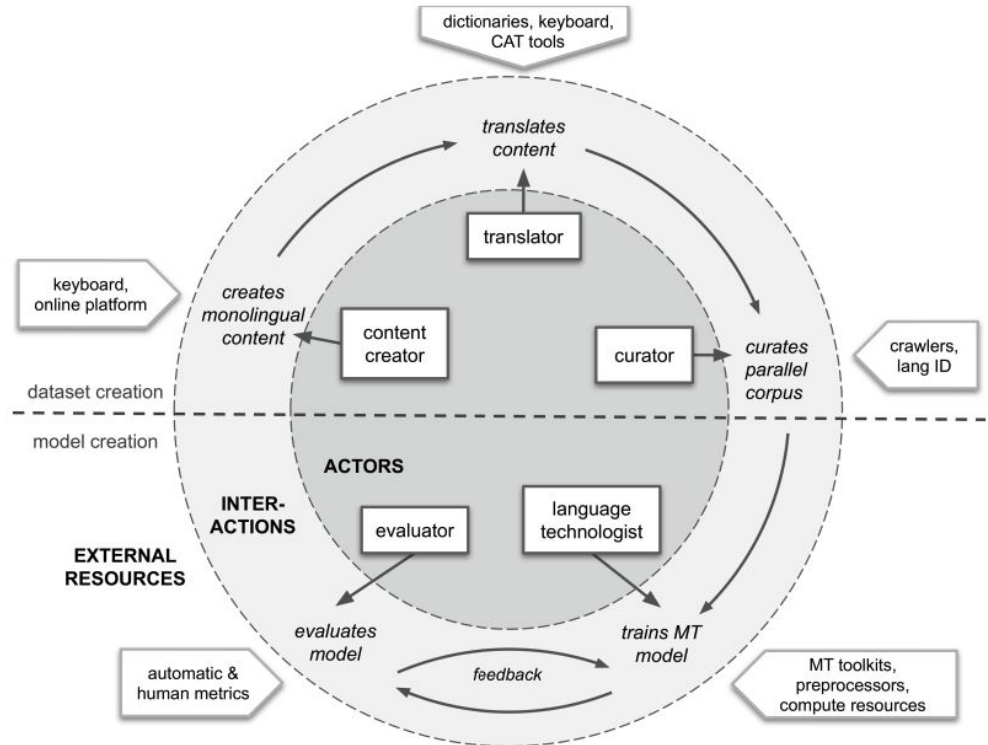


Figure 1: The MT Process, in terms of the necessary agents, interactions and external constraints and demand (excluding stakeholders).



# Promising Directions: Large Language Models

Extending Language Models to also be used for NMT

Same challenges exist with LRL or LR NLP within these scenarios.

# A great resource

## **Neural Machine Translation for Low-Resource Languages: A Survey**

SURANGIKA RANATHUNGA, University of Moratuwa, Sri Lanka

EN-SHIUN ANNIE LEE, University of Toronto, Canada

MARJANA PRIFTI SKENDULI, University of New York Tirana, Albania

RAVI SHEKHAR, Queen Mary University, London

MEHREEN ALAM, National University of Computer and Emerging Sciences, Pakistan

RISHEMJIT KAUR, CSIR-Central Scientific Instruments Organisation, India



# Few-shot learning for Low-resource languages



# Low-resource settings

- **Few-shot learning** targets *no/small available labelled data* in different *low-resource scenarios*
  - **Task**
  - Domain
  - Language



# Low-resource settings

- **Few-shot learning** targets *no/small available labelled data* in different *low-resource scenarios*
  - **Task**
  - **Domain**
  - Language



# Low-resource settings

- **Few-shot learning** targets *no/small available labelled data* in different *low-resource scenarios*
  - Task
  - Domain
  - Language
    - A combination e.g new language, new domain



# Zero-shot settings

- In **zero-shot settings - no labelled data** in the target
  - **Task**
  - **Domain**
  - **Language**

- However, **few examples in the target language/domain/task** can help a lot (Lauscher et al. 2020)
  - Even as little as **10 labelled sentences**
  - Making **Few-shot learning** an attractive research direction



# Evaluation Tasks, Language and Datasets

Tasks	Source language(s)	Target African languages	Dataset
Cross-lingual Natural Language Inference (XNLI)	English	Swahili (sw)	XNLI
Parts of Speech (POS)	English / Multilingual	Yoruba (yo)	UD
Named Entity Recognition (NER)	English / Multilingual	Swahili (sw), Yoruba (yo), Hausa (ha)	WikiANN, MasakhaNER

# Sequence Classification: XNLI

- **Natural language Inference (NLI)**
  - a model is tasked with reading two sentences and determining whether **one entails the other**, **contradicts it**, or **neither (neutral)**.

Sentence 1	Sentence 2	Label
Your words have rankled with him.	He did not like what you said.	entailment
Your words have rankled with him.	He is going to punish you for those words.	neutral
Your words have rankled with him.	He really enjoyed what you said.	contradiction

# Token Classification: POS & NER

- Parts of Speech (POS)
  - Grammatical tagging of words in text e.g.
    - Noun
    - Verb
    - Adjective etc.
- Named entity recognition (NER)
  - Recognizes entities like personal names, location and organization

Only	ADV	O
France	PROPN	B-LOC
and	CCONJ	O
Britain	PROPN	B-LOC
backed	VERB	O
Fischler	PROPN	B-PER
's	PART	O
proposal	NOUN	O
.	PUNCT	O

# Few Shot learning for Low-resource languages

- **Transfer learning**
  - Zero-shot
  - Small labelled data
- **Distant Supervision**
  - Leverages *automatic annotation* of large unlabelled data in the target language/domain/task
- **Data Augmentation**
  - Create additional **synthetic labelled data**
- **Meta-learning**
  - **Model-agnostic approach** to quickly adapt to any new task/domain/language

# Transfer Learning

- Inspired by **Computer Vision** where
  - **Pre-trained Imagenet model** are *fine-tuned on several CV tasks* like object detection and classification
- **Transfer learning** has shown very impressive results in NLP (Howard and Ruder. 2018)
  - E.g **BERT-based models** are *fine-tuned* on multiple NLP tasks
  - Including **cross-lingual** settings.
  - Works *effectively even for unseen languages* provided the script is supported by the pre-trained model

# Transfer Learning: XTREME Dataset

- XTREME - aggregation of several cross-lingual datasets in multiple tasks
  - **Diverse languages** including **two African languages: Swahili and Yoruba**

Task	Corpus	Train	Dev	Test	Test sets	Lang.	Task	Metric	Domain
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI	Acc.	Misc.
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase	Acc.	Wiki / Quora
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS	F1	Misc.
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER	F1	Wikipedia
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction	F1 / EM	Wikipedia
	MLQA			4,517-11,590	translations	7	Span extraction	F1 / EM	Wikipedia
	TyDiQA-GoldP	3,696	634	323-2,719	ind. annot.	9	Span extraction	F1 / EM	Wikipedia
Retrieval	BUCC	-	-	1,896-14,330	-	5	Sent. retrieval	F1	Wiki / news
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval	Acc.	misc.

# Transfer Learning: Zero-shot

- Cross-lingual XNLI
- Named Entity Recognition (NER)

<b>Model</b>	<b>fr</b>	<b>zh</b>	<b>sw</b>
mBERT	73.4	67.8	49.7
<b>XLNLI-R</b>	<b>82.2</b>	<b>78.2</b>	<b>71.2</b>

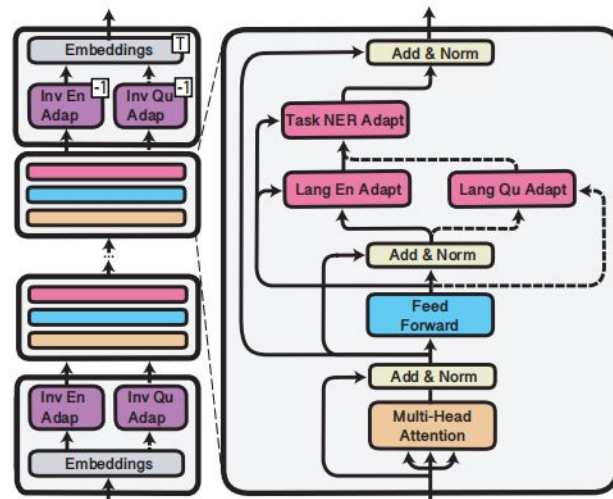
XLNLI: Trained on English

<b>Model</b>	<b>fr</b>	<b>zh</b>	<b>sw</b>	<b>yo</b>
mBERT	79.6	<b>42.7</b>	67.5	<b>44.9</b>
<b>NER-R</b>	<b>80.5</b>	33.1	<b>70.5</b>	33.6

NER: Trained on English

# Parameter Efficient Transfer Learning (1)

- **Cross-lingual Transfer** involves fine-tuning end-to-end
  - Modifying all the parameters of the original model
  - Not reusable for a **new language** once fine-tuned.
  - Also, **not parameter efficient**, huge models for several tasks.
- **Pfeiffer et al. 2020** introduced MAD-X
  - A parameter-efficient approach that adds adapter in between transformer models.
  - This achieves competitive result with less than 1% additional parameters





# Parameter Efficient Transfer Learning (2)

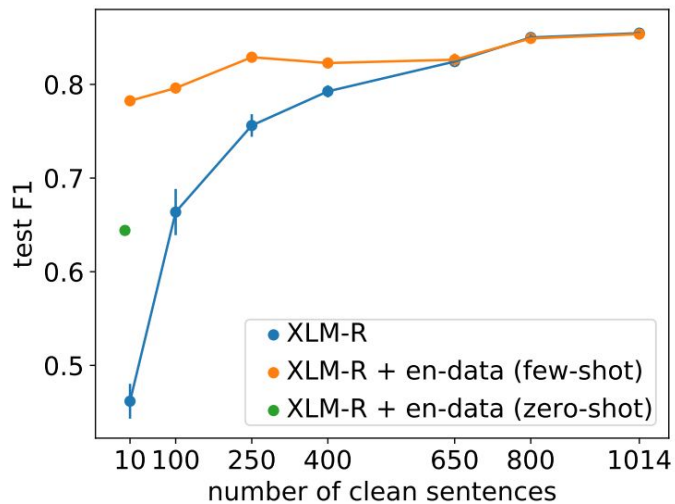
- Baselines:
  - XLM-R - fine-tuned on English NER dataset
  - XLM-R MLM-SRC
    - MLM on English language corpus
  - XLM-R MLM-TRG
    - MLM on target language corpus
  - MAD-X
    - Train **language adapter** of English
    - Train **language adapter** on target language
    - Train **task adapter** on English
    - **Swap language adapter** before zero-shot transfer

<b>Model</b>	<b>en</b>	<b>zh</b>	<b>sw</b>	<b>gn</b>
XLM-R	83.0	19.6	63.5	41.0
XLM-R MLM-SRC	84.2	11.0	57.9	41.7
XLM-R MLM-TRG	84.2	15.5	<b>77.7</b>	50.6
MAD-X	82.3	<b>20.5</b>	73.8	<b>55.1</b>

Zero-shot NER: English as source-language

# Transfer Learning: Small Additional data

- Having **small additional labelled data (e.g 10 or 100 sentences)** can give impressive results.
  - Less time consuming e.g 30 minutes of annotation.
  - Cost less/no money.
- **Example**
  - First **train on English as source language (zero-shot)**
  - Additional **fine-tune on small in-language labelled dataset (few-shot)**



(c) Transfer Learn NER Hausa

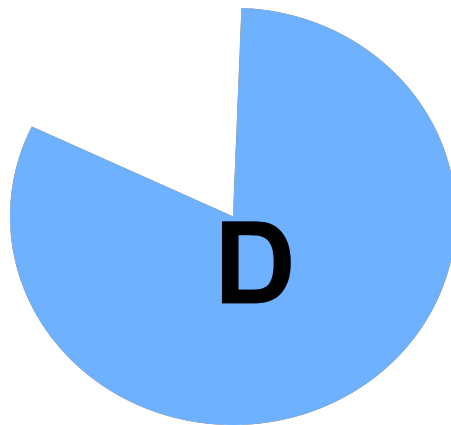
# Distant Supervision

Clean, expensive,  
manually-annotated text



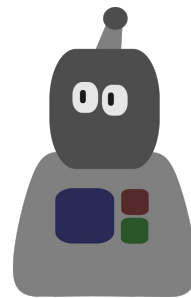
Unlabeled text

+ automatic annotation  
(quick + cheap)



Leverage

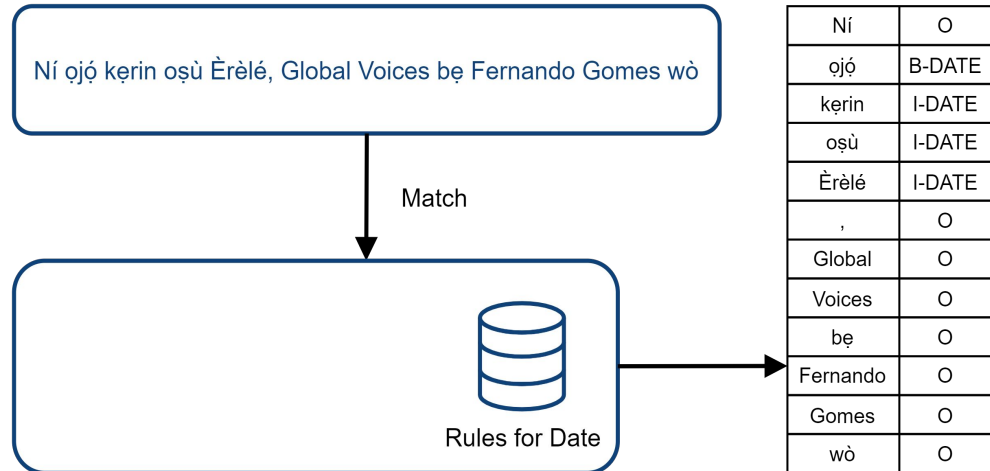
- context
- expert insights
- external knowledge and resources
- self-training



# Distant Supervision

## Rules

- Native speaker (domain expert)
- Date detection using keywords like "oḡo" (day) "oṣu" (month)



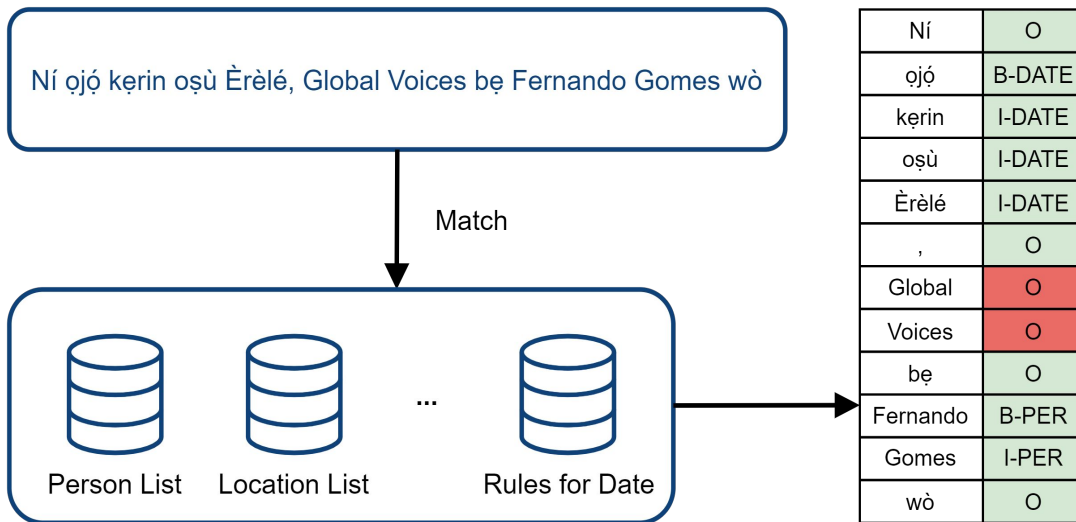
# Distant Supervision

## Rules

- Native speaker (domain expert)
- Date detection using keywords like "oḵó" (day) "oṣù" (month)

## Entity lists

- From sources like gazetteers, dictionaries, phone books, and Wikipedia



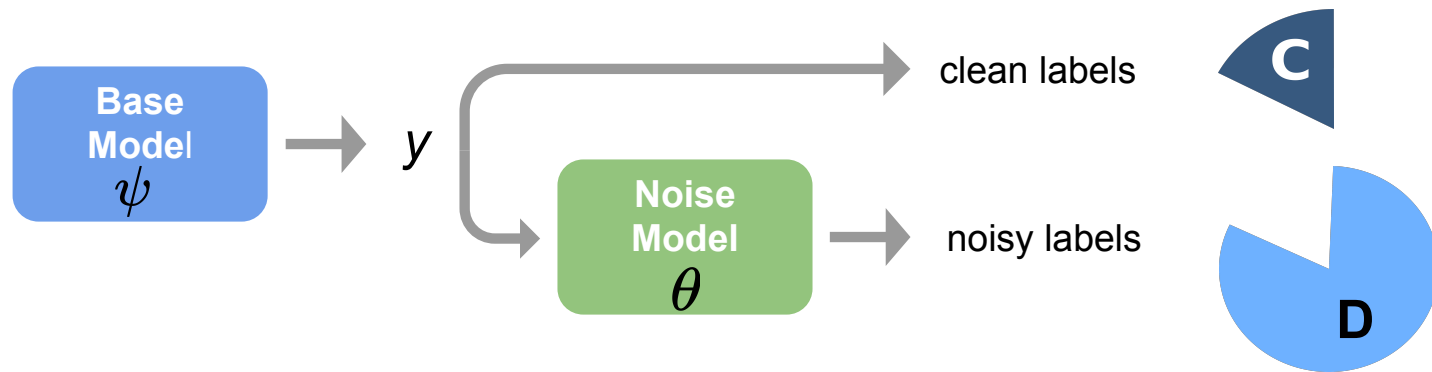
# Label-noise handling

- Distant supervision usually more errors → noisy labels
- Can deteriorate performance
- Explicit noise handling
  - Noise modeling
  - Label cleaning

Named Entity	hau F1	yor F1
LOC	65	76
ORG	18	44
PER	51	19
DATE	54	62
<b>Overall</b>	<b>51.6</b>	<b>54.9</b>

Quality of Distant Supervision

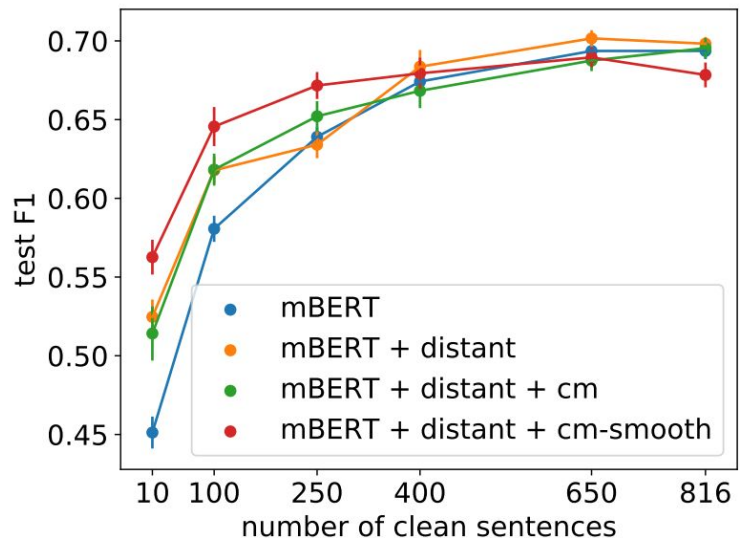
# Noise Modeling



- Confusion Matrix [Hedderich & Klakow, DeepLo 2018]
  - Pairs of clean and noisy labels
  - Performance affected by how good is the initial noise matrix.

# Distant Supervision + Noisy Handling

- mBERT
  - Baseline trained on **only clean sentences**
- mBERT + distant
  - trained on **only clean & distant sentences**
  - **Distant**: automatically annotated
- mBERT + distant + cm
  - **Noise handling** with **Confusion Matrix**
- mBERT + distant + cm-smooth
  - **cm-smooth** - provides better initialization



(a) Distant Supervision NER Yorùbá



# Data Augmentation techniques

- Annotating large amount of labelled data is **time-consuming and costly**
- **Data augmentation**: synthetic data generation from small labelled data e.g.
  - **Back-translation** approach in machine translation
  - **Synonym replacement** in text classification tasks
- For **NER**, it is difficult since
  - The **tags of each tokens** should be maintained
  - **Word order** should be maintained

# Data Augmentation for NER

- Liu et al. 2021 proposed MulDA - a **three step approach**

**Labeled sentence in the source language:**  
[PER Jamie Valentine] was born in [LOC London].

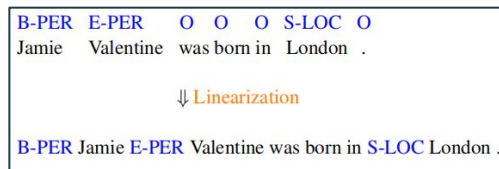
**1. Translate sentence with placeholders:**  
src: PER0 was born in LOC1.  
tgt: PER0 nació en LOC1.

**2. Translate entities with context:**  
PER0  
src: [Jamie Valentine] was born in London.  
tgt: [Jamie Valentine] nació en Londres.

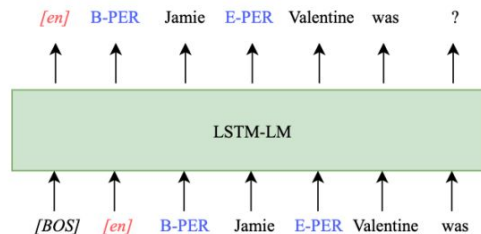
LOC1  
src: Jamie Valentine was born in [London].  
tgt: Jamie Valentine nació en [Londres].

**3. Replace placeholders with translated entities:**  
[PER Jamie Valentine] nació en [LOC Londres].

Labeled sentence translation



Labeled Sequence linearization



Multilingual LSTM-LM on linearized sequences

Ensures diversity i.e new entity generation

# Data Augmentation for NER

- Liu et al. 2021 proposed MulDA - a **three step approach**

**Labeled sentence in the source language:**  
[PER Jamie Valentine] was born in [LOC London].

**1. Translate sentence with placeholders:**  
src: PER0 was born in LOC1.  
tgt: PER0 nació en LOC1.

**2. Translate entities with context:**  
PER0  
src: [Jamie Valentine] was born in London.  
tgt: [Jamie Valentine] nació en Londres.

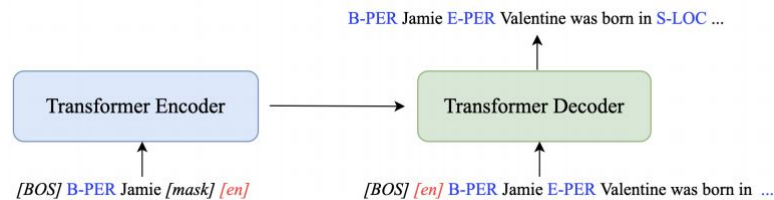
LOC1  
src: Jamie Valentine was born in [London].  
tgt: Jamie Valentine nació en [Londres].

**3. Replace placeholders with translated entities:**  
[PER Jamie Valentine] nació en [LOC Londres].

Labeled sentence translation

B-PER E-PER O O O S-LOC O  
Jamie Valentine was born in London .  
↓ Linearization  
B-PER Jamie E-PER Valentine was born in S-LOC London .

Labeled Sequence linearization



mBART on linearized sequences

[BOS] [de] B-PER Jamie E-PER Valentine wurde in S-LOC London geboren.  
[BOS] [es] B-PER Jamie E-PER Valentine nació en S-LOC Londres.  
[BOS] [nl] B-PER Jamie E-PER Valentine werd geboren in S-LOC Londen.  
...

# Data Augmentation for NER

- **En**
  - train on 1k english data
- **En + Multi-Train**
  - train on **multiple source languages translated from English**
- **MulDA-LSTM**
  - LSTM to generate new sentences
- **MulDA-mBART**
  - mBART to generate new sentences

Method	en	zh	sw	yo
En	74.81	13.46	64.59	44.62
En + Multi-Train	77.27	39.40	67.81	49.75
MulDA-LSTM	78.23	<b>41.77</b>	<b>68.77</b>	48.09
MulDA-mBART	<b>78.79</b>	41.30	67.40	<b>52.97</b>

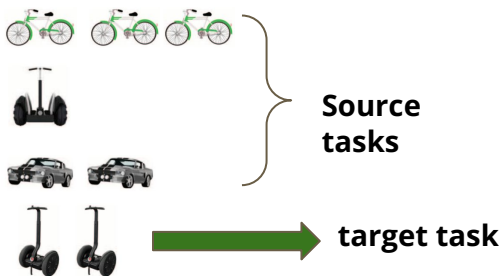
MulDA Results: Source languages are *ar, en, fr, it, ja, tr, zh*

# Meta Learning

- Meta-learning also known as learning to learn
- Models the way humans, especially *kids learn from small examples*
- A method to train a model on a **variety of learning tasks**, such that it can solve *new learning tasks* using only a *small number of training samples*.

# Meta-learning Approaches

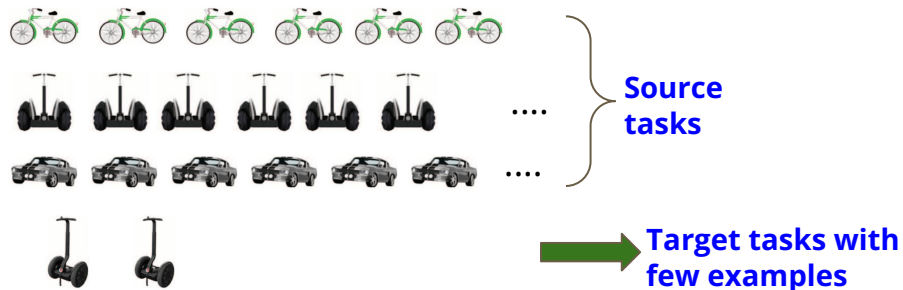
Humans learn from few examples



Supervised learning with a lot of examples



Meta learning



# Meta-learning: MAML

- **Model-Agnostic** - it is compatible with any model (architecture independent) trained with gradient descent.
- including classification, regression and reinforcement learning **tasks**
- **Task** can be a *new domain or language* depending on the setting.
- with applications in several tasks like **image classification, text classification, machine translation** and **named entity recognition**

# Meta-learning: MAML & MetaSGD

---

**Algorithm 1** Model-Agnostic Meta-Learning

---

**Require:**  $p(\mathcal{T})$ : distribution over tasks

**Require:**  $\alpha, \beta$ : step size hyperparameters

- 1: randomly initialize  $\theta$
- 2: **while** not done **do**
- 3:   Sample batch of tasks  $\mathcal{T}_i \sim p(\mathcal{T})$
- 4:   **for all**  $\mathcal{T}_i$  **do**
- 5:     Evaluate  $\nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta})$  with respect to  $K$  examples
- 6:     Compute adapted parameters with gradient descent:  $\theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta})$
- 7:   **end for**
- 8:   Update  $\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_i})$
- 9: **end while**

**MAML (Finn C. et al, 2017)**

---

**Algorithm 1:** Meta-SGD for Supervised Learning

---

**Input:** task distribution  $p(\mathcal{T})$ , learning rate  $\beta$

**Output:**  $\theta, \alpha$

- 1: Initialize  $\theta, \alpha$ ;
- 2: **while** not done **do**
- 3:   Sample batch of tasks  $\mathcal{T}_i \sim p(\mathcal{T})$ ;
- 4:   **for all**  $\mathcal{T}_i$  **do**
- 5:      $\mathcal{L}_{\text{train}(\mathcal{T}_i)}(\theta) \leftarrow \frac{1}{|\text{train}(\mathcal{T}_i)|} \sum_{(\mathbf{x}, \mathbf{y}) \in \text{train}(\mathcal{T}_i)} \ell(f_{\theta}(\mathbf{x}), \mathbf{y})$ ;
- 6:      $\theta'_i \leftarrow \theta - \alpha \circ \nabla \mathcal{L}_{\text{train}(\mathcal{T}_i)}(\theta)$ ;
- 7:      $\mathcal{L}_{\text{test}(\mathcal{T}_i)}(\theta'_i) \leftarrow \frac{1}{|\text{test}(\mathcal{T}_i)|} \sum_{(\mathbf{x}, \mathbf{y}) \in \text{test}(\mathcal{T}_i)} \ell(f_{\theta'_i}(\mathbf{x}), \mathbf{y})$ ;
- 8:   **end**
- 9:    $(\theta, \alpha) \leftarrow (\theta, \alpha) - \beta \nabla_{(\theta, \alpha)} \sum_{\mathcal{T}_i} \mathcal{L}_{\text{test}(\mathcal{T}_i)}(\theta'_i)$ ;
- 10: **end**



# MAML for Cross-Lingual NLI

- Meta-learning **task** is a **new language**
- Nooralahzadeh et al. 2020 proposed **X-MAML** for XNLI:
- **Step 1: Pre-train** on a high-resource language (i.e., English):
- **Step 2: Meta-learn** on one or more auxiliary languages from the low-resource set.
- **Step 3: Zero-shot** or **few-shot** learn on the target languages.

Model	en	fr	sw
mBERT	81.36	73.45	47.58
X-MAML (one aux. lang.) $hi \rightarrow X$	81.88	74.17	47.12
X-MAML (two aux. lang.) $(l_1, l_2) \rightarrow X$	<b>(hi, de)</b> <b>82.59</b>	<b>(hi, ar)</b> <b>75.69</b>	<b>(el, tr)</b> <b>50.42</b>

X-MAML Results

# MetaSGD for Cross-Lingual Part of Speech

- Compare **multi-task learning** on many languages with **meta-learning**.
- **Ponti et. al. 2021** proposed meta-training on **99 POS treebanks** and
  - **Evaluated on 16 treebanks (14 languages)** from low-resource languages.
  - We show an example for **Yoruba**.

Model	Number of shots			
	0	5	10	20
Multi-task	41.46	59.59	63.34	67.23
Meta-SGD	<b>47.34</b>	<b>62.93</b>	<b>66.71</b>	<b>69.14</b>

Meta-SGD performance on Yoruba POS task

# Conclusion

- We discuss various **approaches for few-shot learning** for low-resourced languages including
  - Transfer learning approaches
  - Distant supervision and noise handling techniques
  - Data augmentation approach for named entity recognition, and
  - Meta-learning approaches (MAML & Meta-SGD)
- We show the application of the few-shot learning approaches on **various tasks**
  - Natural language inference
  - Parts of speech, and
  - Named entity recognition

# Data and Resources



## AI4D - African Language Dataset Challenge

**Kathleen Siminyu**

Artificial Intelligence for Development  
Africa

kathleensiminyu@gmail.com

**Sackey Freshia**

Jomo Kenyatta University  
of Agriculture and Technology

freshiasackey@gmail.com

**Jade Abbott**

Retro Rabbit

jabbott@retorabbit.co.za

**Vukosi Marivate**

University of Pretoria

vukosi.marivate@cs.up.ac.za

Language	Tasks	Submissions
Yoruba	MT, Diacritic Verification, Text Classification, NER, misc	7
Kiswahili	Document Classification, misc	6
Igbo	NER, misc	4
Hausa	Sentiment Analysis, Document classification, misc	4
Fongbe	MT, Speech to Text, misc	3
Amharic	Hate speech detection, stop words list, misc	3
Asante Twi	Sentiment Analysis, MT, misc	3
Chichewa	NER, MT	2
Ewe	MT, misc	2
Wolof	ASR	1
Tunizian Arabizi	Sentiment Analysis	1
Kikuyu	misc	1
Kabiwe	MT	1
Oromo	misc	1
Zulu	misc	1

Table 1: Language and Task distribution of submissions.



ARTIFICIAL  
INTELLIGENCE  
FOR  
DEVELOPMENT  
AFRICA

**giz**

**Z:IND!**

# African Natural Language Processing (AfricaNLP)

## Recent uploads



December 1, 2020 (0.2) Dataset Open Access

### Swahili : News Classification Dataset

Davis David;

Swahili is spoken by 100-150 million people across East Africa. In Tanzania, it is one of two national languages (the other is English) and it is the official language of instruction in all schools. News in Swahili is an important part of the media sphere in Tanzania. News contributes to education,

Uploaded on September 17, 2021

1 more version(s) exist for this record

View

July 31, 2021 (1.0) Dataset Open Access

### South African Disinformation [Fake News] Website Data - 2020

Harm, de Wet; Marivate, Vukosi;

See publication: Is it Fake? News Disinformation Detection on South African News Websites We used, as sources, investigations by the news websites MyBroadband (<https://mybroadband.co.za/forum/threads/list-of-known-fake-news-sites-in-south-africa-and-beyond.879854/>) and News24 (<https://exposed>).

View

New upload

### African Natural Language Processing (AfricaNLP)

This is a project that aims to curate African Natural Language Processing projects and data that is uploaded on Zenodo. Please fill free to tag this community when uploading.

#### Curated by:

vukosi

#### Curation policy:

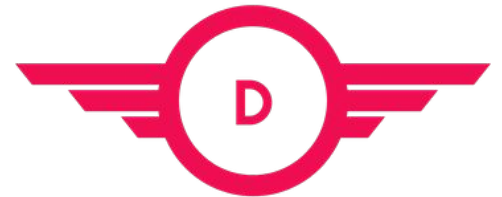
We are very much open to any submissions but may ask for added documentation to projects when needed.

#### Created:

February 15, 2020

#### Harvesting API:

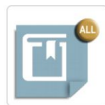
[OAI-PMH Interface](#)



Data Science for Social Impact

A collection of language resource metadata mostly collected during the NHN funded technology audit of 2009, as well as the SADiLaR technology audit of 2018. Not all resources in this collection are available for download.

## Recent Submissions



### **Linguistically enriched corpora for conjunctively written South African languages**

Puttkammer, Martin, et al. (North-West University, Centre for Language Technology (CTeX1), 2021-09)

This resource contains linguistically annotated data for four official South African languages with a conjunctive orthography from the Nguni family ...



### **Description of N|uu**

Sands, Bonny, et al. (Bonny Sands, 2015-10-06)

Recordings of dictionary entries for a pan-dialectal dictionary of the N|uu language (Eastern and Western dialects) made by Bonny Sands, Johanna Brugman, ...



### **Mburisano Covid-19 multilingual corpus**

Marais, Laurette (CSIR Voice Computing, 2020-12-04)

This corpus was created to aid development of the AwezaMed Covid-19 speech-to-speech mobile application. The project within which it was created, ...



### **Denominal adjectives in Afrikaans dataset**

Trollip, Benito (South African Centre for Digital Language Resources, 2020-05-15) ~ Resource Catalogue

This dataset contain a collection of Afrikaans denominal adjectives that were extracted from the Virtual Institute for Afrikaans' corpus portal. The ...



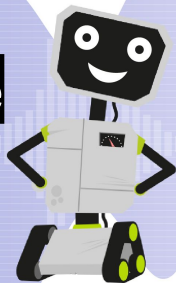
# TRANSLATORS WITHOUT BORDERS



## Global WordNet Association

Common Voice

moz://a



<https://huggingface.co/docs/datasets/>

# CORPUS



# Some Tips

- Reach out local communities.
- Talk to linguists.
- More resources to get funding for data collection/tool creating with local communities.
- Be open to learning.

# THE BEAT OF THE DRUM





If you want to go fast, go alone; if  
you want to go far, go together.

African Proverb



North Africans in NLP



GhanaNLP



DEEP LEARNING  
INDABA



# A beat from West Africa

## Opportunities

Let's hear from **Stephen Moore**

*What makes you excited about working on your languages?*

Department of Mathematics,  
University of Cape Coast,



GhanaNLP





# A beat from East Africa

## Opportunities

- Tapping into available funding opportunities to build datasets for several NLP tasks like Machine Translation, Speech recognition, Topic Modelling, Sentiment analysis, Misinformation detection and Named Entity Recognition.
- Leveraging on the use of transfer learning techniques to build models for example Machine Translation models for low-resourced language based on another language that is in the same family.
- Working closely with language experts in the Institute of African languages throughout the collection and curation of the open datasets.

## Joyce Nakatumba-Nabende

Makerere Artificial Intelligence Lab  
Uganda



# A beat from South East Asia

## Opportunities

- Working to preserve data, publicising and sharing.
- Making more tools available on Github.
- Great opportunities with pre-trained models for low resource languages.
- Dealing with code mixing.
- Closing the gaps to allow tackling of online abuse.

Let's hear from **Surangika Ranathunga**

***What makes you excited about working on your languages?***

Department of Computer Science and Engineering, University of  
Moratuwa  
Sri Lanka

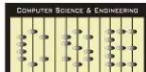






# ආයුෂෝචන් (May you have a long life)

- Sinhala (සිංහල)
  - Primarily used in Sri Lanka
  - Morphologically rich
  - Own alphabet and script
- My work
  - Linguistic tools for Sinhala
  - End-user NLP applications for Sinhala
- Challenges
  - Lack of linguistic expertise
  - Lack of data
  - Noise in data
- Opportunities
  - Supported in Google translate, input tools, Facebook, etc
  - Included in pre-trained multilingual embedding models s.a. XLM-R, mBART, mT5
- Way forward
  - Open up our data and models
  - Participatory and collaborative research



# A beat from the Americas

## Opportunities

- Extending accessibility of local languages.
- Capturing local knowledge and supporting equality.
- Curating and developing multi-parallel evaluation sets for typologically diverse languages.
- Working mostly in a non-English-centric approach.

Let's hear from **Arturo Oncevay**

*What makes you excited about working on your languages?*



PhD student, University of Edinburgh  
Member of the organising committee of AmericasNLP  
From Peru

## Language Diversity



Follow us on Twitter:  
**@AmericasNLP**  
You can join our Slack!  
(link provided by DM)





# A CALL TO ACTION



# More to do, too little time

For some languages, there is not much time left to get good coverage.

You can look for researchers in these spaces and see what might be good approaches to

- Data Collection
- Data Creation
- Data Curation
- Model Development
- Tool Deployment

# More to do, too little time

There are many grassroots organizations/initiatives

Connect

Listen

Engage

# In Closing

We hope through this tutorial

- You have gained a good technical understanding of LRL LRNLP
- That the work is not just technical.
- Document your work
  - Model Card
  - Data Statements
- Engage with the communities

# A Journey Through the Opportunity of Low Resourced Natural Language Processing — An African Lens



UNIVERSITEIT VAN PRETORIA  
UNIVERSITY OF PRETORIA  
YUNIBESITHI YA PRETORIA



Data Science for Social Impact

Vukosi Marivate and David Adelani  
[vukosi.marivate@cs.up.ac.za](mailto:vukosi.marivate@cs.up.ac.za) - @vukosi  
[didelani@lsv.uni-saarland.de](mailto:didelani@lsv.uni-saarland.de) @davlanade



**SIC** Saarland Informatics  
Campus