Machine Translation for Humanitarian Response

Alp Öktem - Translators without Borders

May 26th, 2021

GourMeT Open NLP Workshop 2021

🕛 The next twenty minutes

- Translators without Borders (TWB)
- Gamayun - Language equality initiative
  - Language data collection
  - Machine translation
  - Use-case oriented evaluation
- Questions, feedback, discussion

Information is power!


in the right language
Information is power!

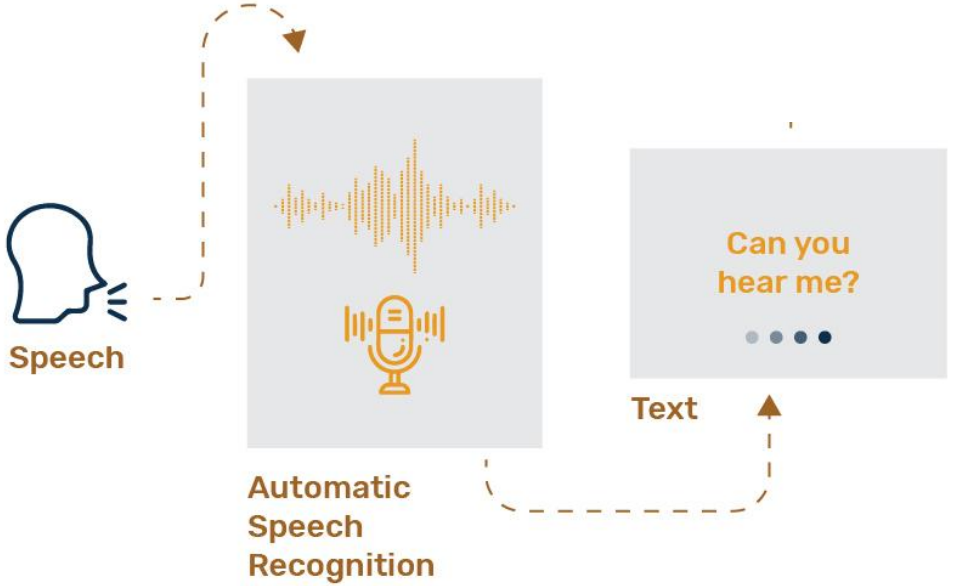
Helping people to get vital information, and be heard, in the languages they speak and understand

TRANSLATORS WITHOUT BORDERS

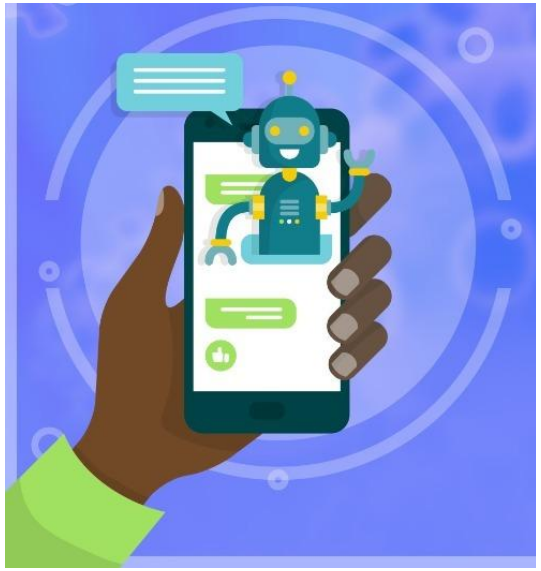
Gamayun

Language equality initiative

Machine Translation



Automatic Speech Recognition



Chatbots

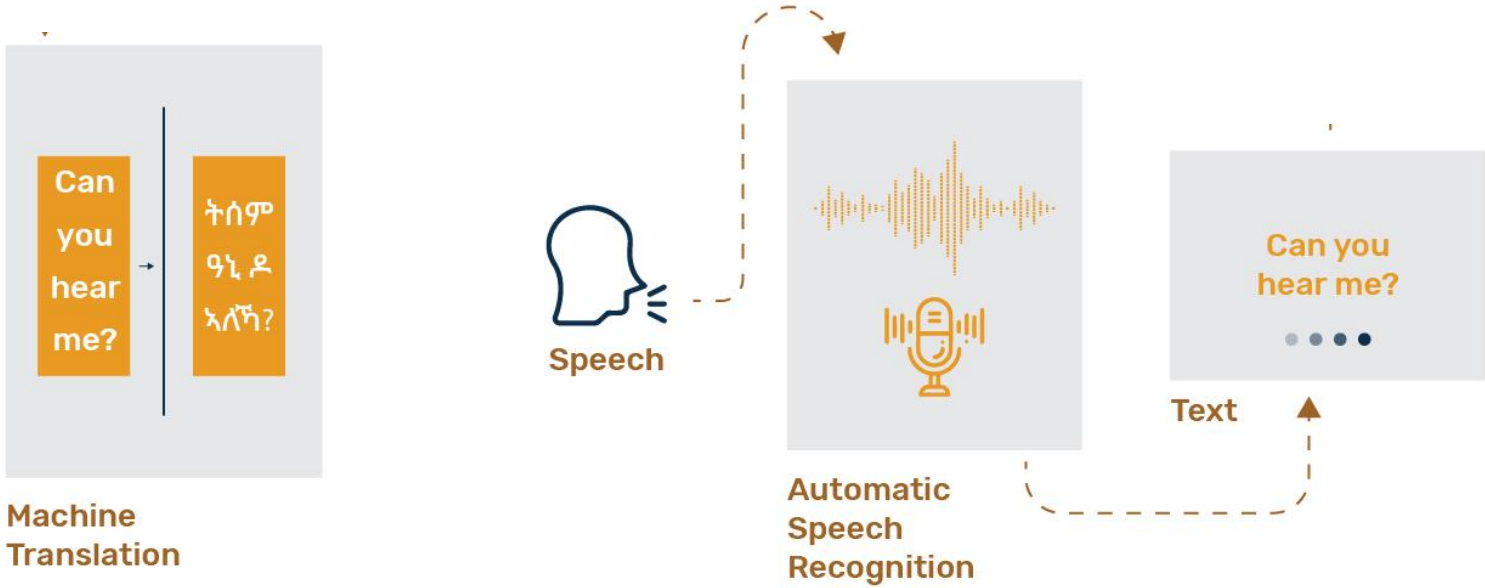
Language data collection

Machine translation

Use-case oriented evaluation

Language data collection

Language data collection



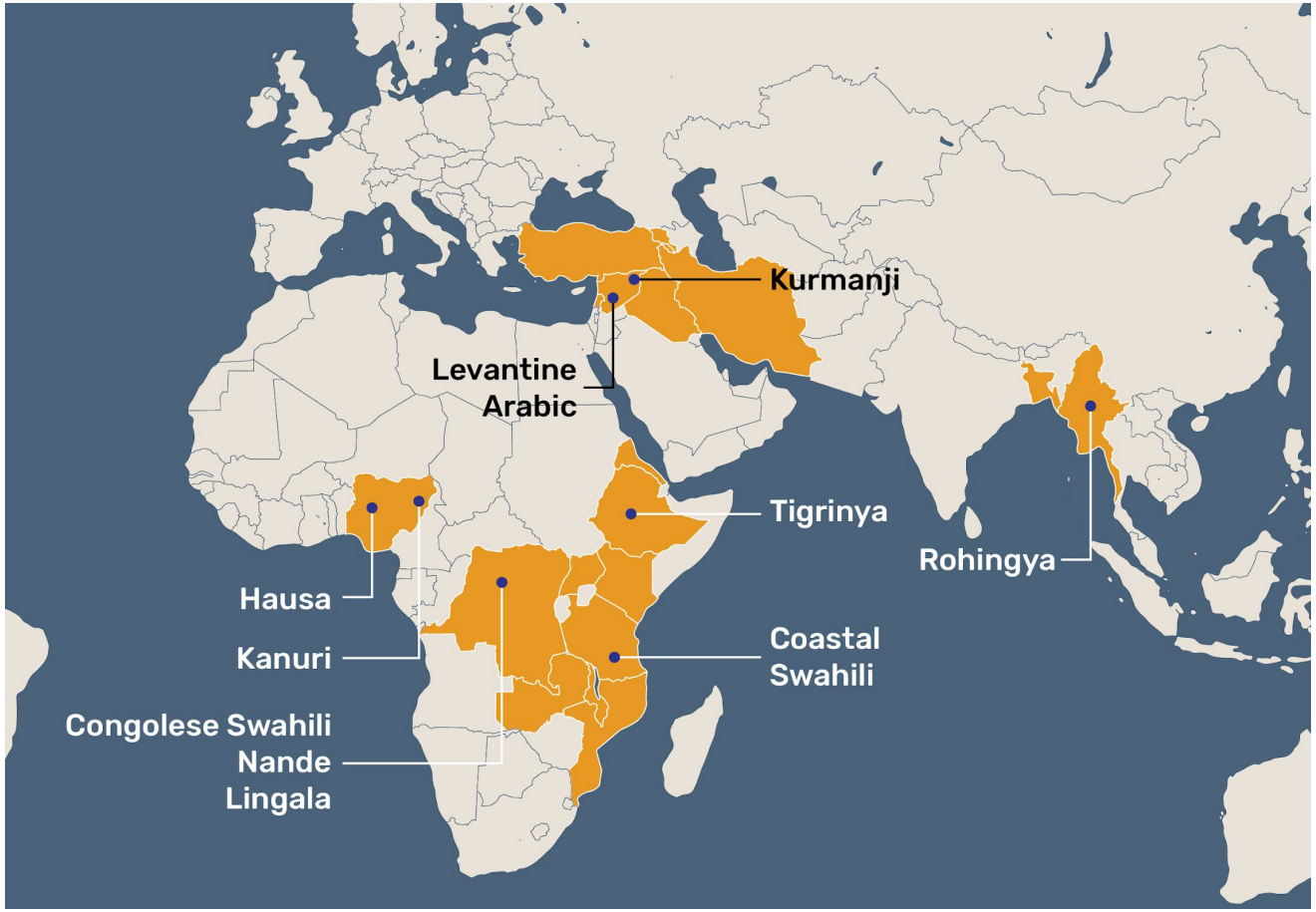
Parallel data

Audio data

Gamayun Kits

- Language data for zero/low-resource languages
- General domain sentences
  - from Tatoeba corpus
  - in English, Spanish, French
- Translated for parallel data → MT
- Recorded for audio data → ASR, TTS

Gamayun Kits



- Currently: Hausa, Kanuri, Rohingya, Coastal/Congolese Swahili, Nande
- http://gamayun.translatorswb.org/data/

Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, Grace Tang. ***Gamayun – Language Technology for Humanitarian Response.*** In: 2020 IEEE Global Humanitarian Technology Conference (GHTC) 2020 November 1: Virtual.

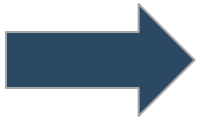
TICO-19

SUPPORT TICO-19

Volunteer	Translate	Share

- **T**ranslation **I**nitiative for **CO**vid-**19**
- Partners: *George Mason University, Carnegie Mellon U., Johns Hopkins U., Google, Microsoft, Facebook, Amazon, Translated* and *Appen*
- Covid-19 domain data for:
  - Translation memories
  - Building specialized MT
  - Terminology

- COVID-19 domain source content from:
  - Medical conversations
  - PubMed articles
  - News
  - Travel restriction announcements
  - Organization announcements
  - Informative Articles



	Amharic	Arabic	Bengali	Kurdish Sorani	Dinka	English	Spanish LA	Farsi	French	Nigerian Fulfulde	Hausa	Hindi	I
Amharic (am)													
Arabic (ar)							zipped .tmx		zipped .tmx			zipped .tmx	z
English (en)		zipped .tmx	zipped .tmx	zipped .tmx	zipped .tmx		zipped .tmx	zipped .tmx	zipped .tmx	zipped .tmx	zipped .tmx	zipped .tmx	z
Spanish LA (es-LA)		zipped .tmx							zipped .tmx			zipped .tmx	z
Farsi (fa)													
French (fr)		zipped .tmx					zipped .tmx					zipped .tmx	z
Hindi (hi)		zipped .tmx	zipped .tmx				zipped .tmx		zipped .tmx				z
Indonesian (id)		zipped .tmx					zipped .tmx		zipped .tmx			zipped .tmx	
Kurdish Kurmanji (ku)				zipped .tmx									
Portuguese Brazilian (pt-BR)		zipped .tmx					zipped .tmx		zipped .tmx			zipped .tmx	z

- 105 Translation Memories,
- 38 languages,
- Based on 3000 translated sentences
- https://tico-19.github.io/

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, Sylwia Tur. ***TICO-19: The Translation Initiative for COvid-19.*** In: NLP COVID-19 Workshop (Part 2) @ EMNLP. 2020 November 19-20; Online.

Language data collection

Language data collection

Machine translation

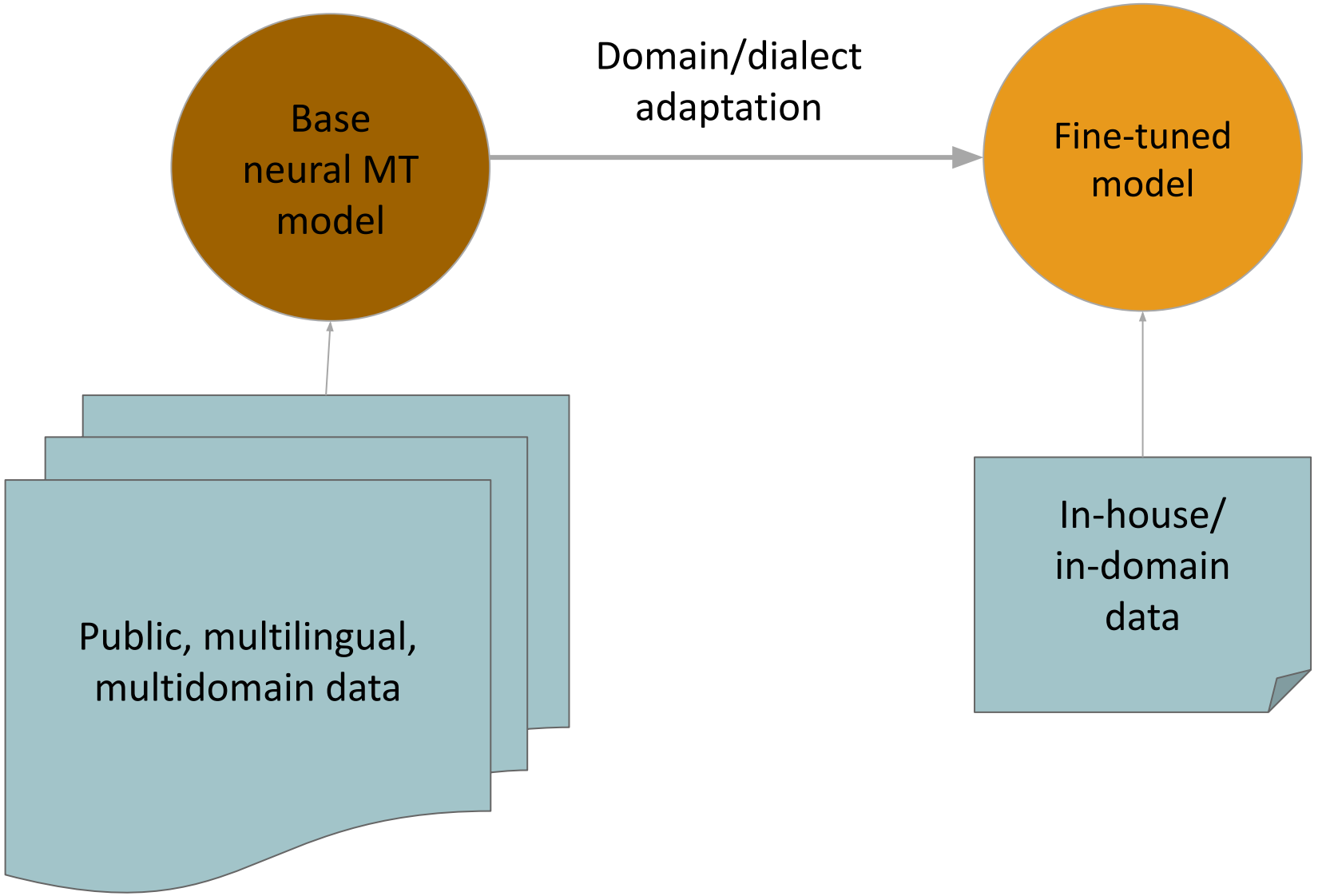
Machine translation

Objectives

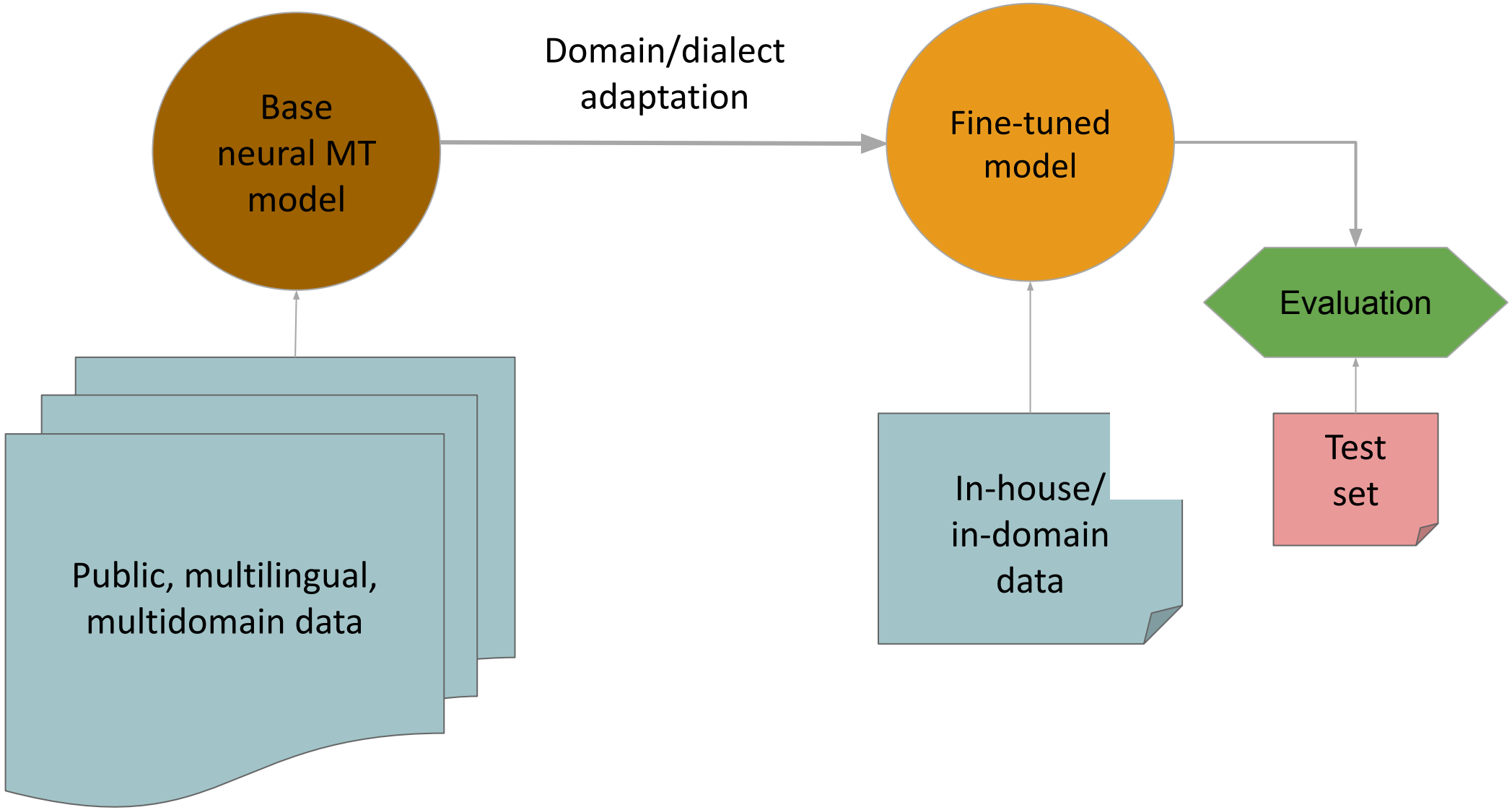
- Closing the technology gap for marginalized languages
- *Empowering* professional/non-professional translators
- Accelerate the flow of vital information in crises
- Automatic monitoring of humanitarian data

Pilot languages

- Levantine Arabic, Tigrinya, Congolese Swahili



General methodology



General methodology

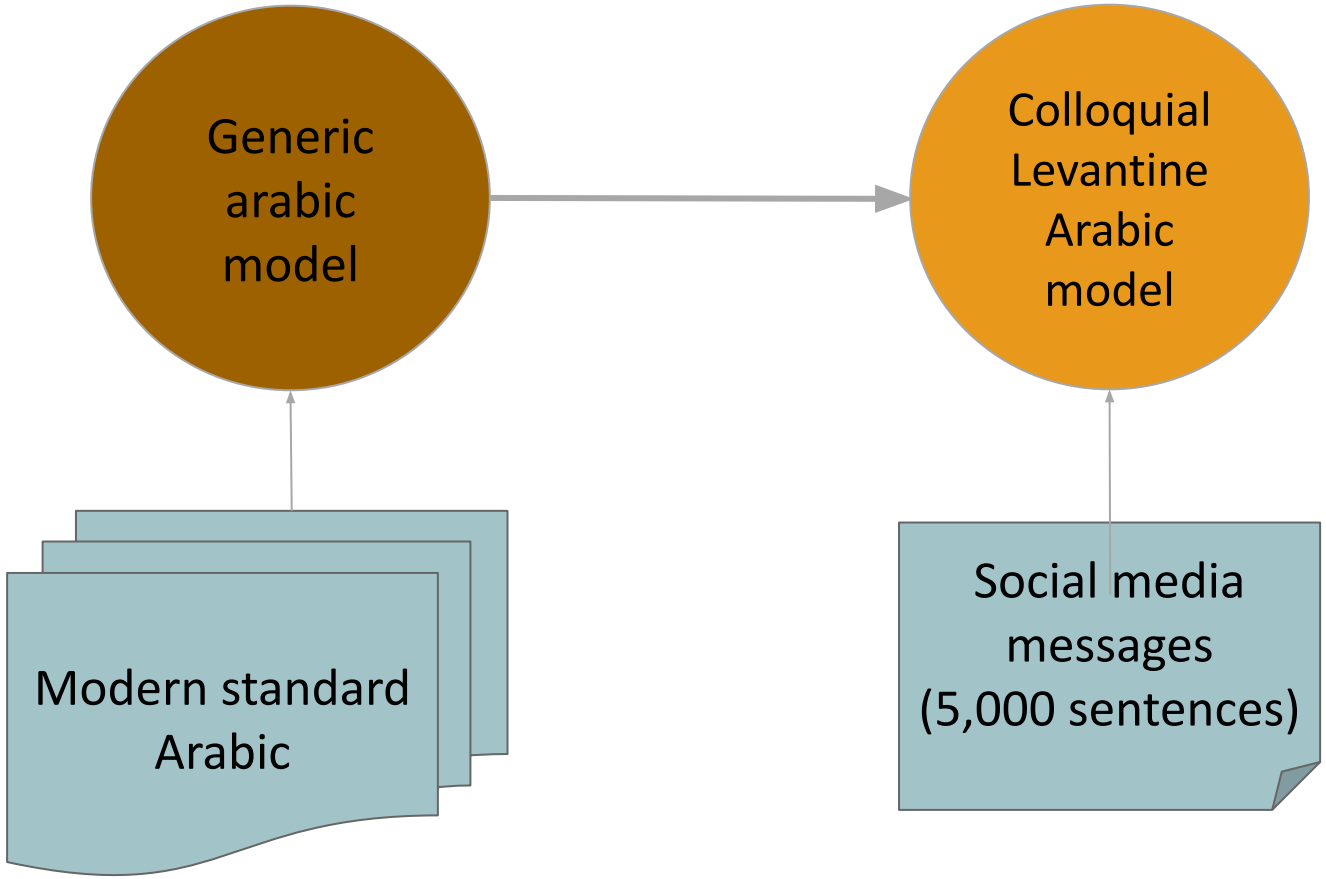
Levantine Arabic
Social media monitoring

Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, Grace Tang. ***Gamayun – Language Technology for Humanitarian Response.*** In: 2020 IEEE Global Humanitarian Technology Conference (GHTC.) 2020 November 1: Virtual.

WFP
World Food Programme

KHABRONA .INFO خبرونا

Levantine Arabic
Social media monitoring



KHABRONA .INFO خبرونا

WFP World Food Programme

Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, Grace Tang. ***Gamayun – Language Technology for Humanitarian Response.*** In: 2020 IEEE Global Humanitarian Technology Conference (GHTC.) 2020 November 1: Virtual.

Levantine Arabic
Social media monitoring

Evaluation

- BLEU scoring with respect to reference translation
  - Test set: 200 messages
  - Domain adaptation: 19.5% -> 24.8%
  - Google MT: 21.2%

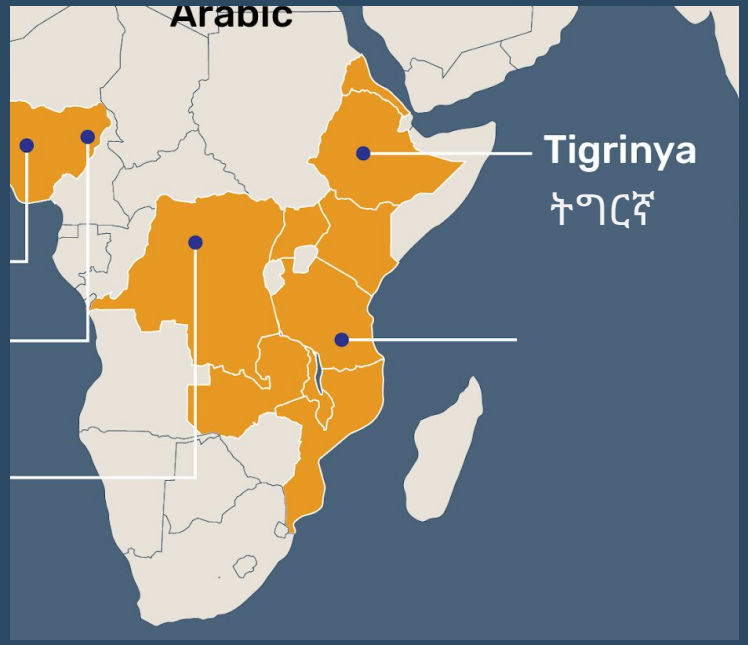
Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, Grace Tang. ***Gamayun – Language Technology for Humanitarian Response.*** In: 2020 IEEE Global Humanitarian Technology Conference (GHTC.) 2020 November 1: Virtual.

Levantine Arabic
Social media monitoring

- "الغلاء انقطاع في المياه" - "Excessive water cuts and high prices"
  - TWB-MT: "high prices , water cuts"
  - Google MT: "Expensive interruption in the water"
- "الظروف المادية" - "Financial conditions"
  - TWB-MT: "Financial conditions"
  - Google MT: "Physical conditions"

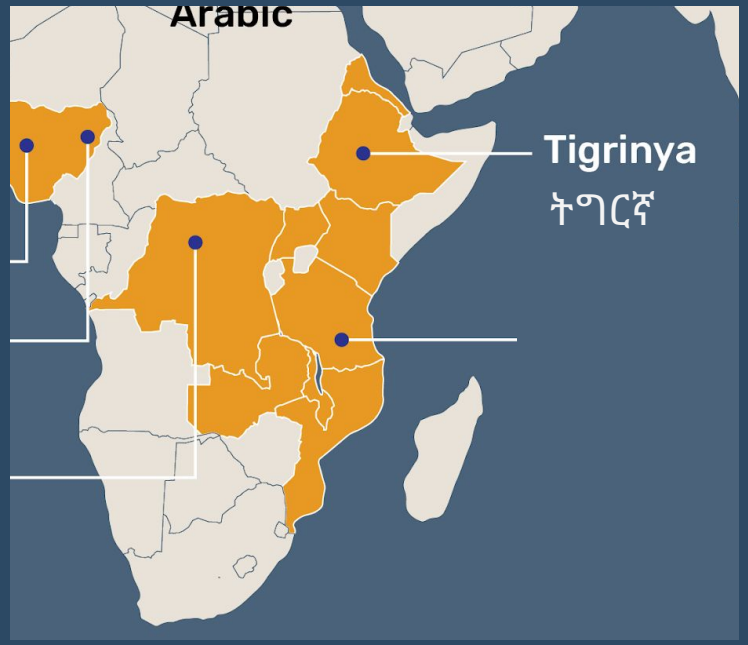
Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, Grace Tang. ***Gamayun – Language Technology for Humanitarian Response.*** In: 2020 IEEE Global Humanitarian Technology Conference (GHTC.) 2020 November 1: Virtual.

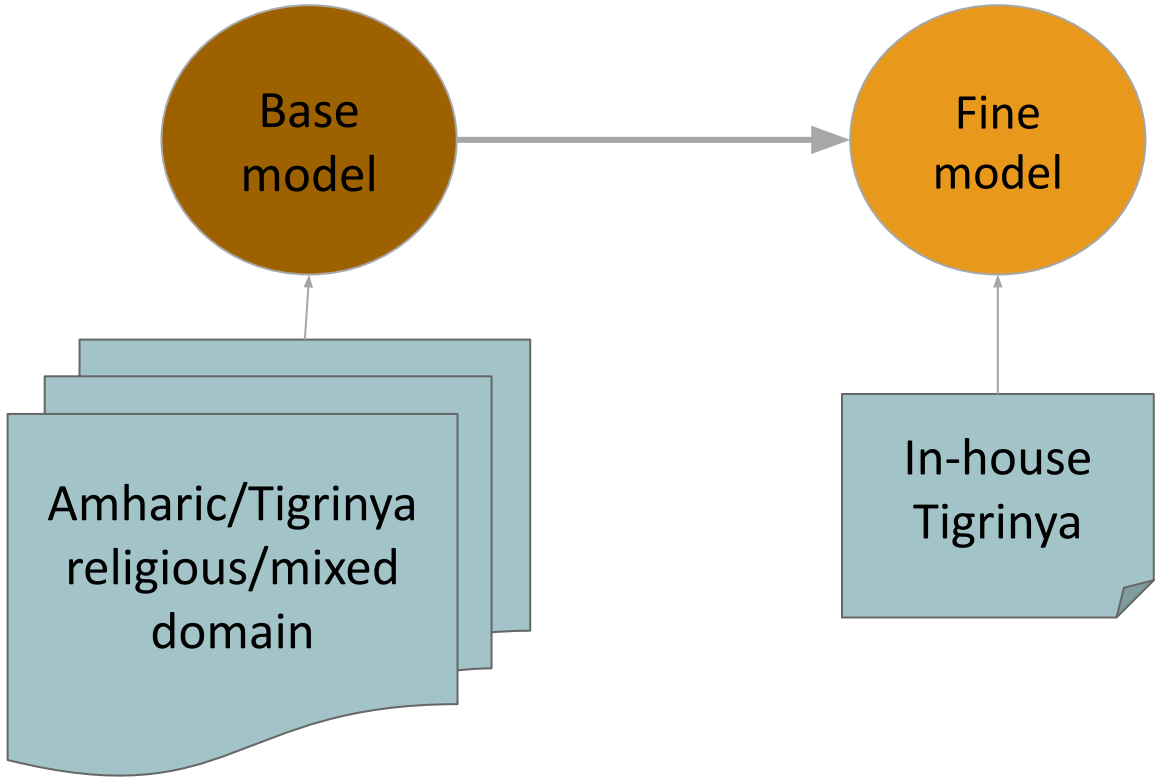
Tigrinya



Alp Öktem, Mirko Plitt, Grace Tang. ***Tigrinya Neural Machine Translation with Transfer Learning for Humanitarian Response.*** In: AfricaNLP workshop @ Eighth International Conference on Learning Representations (ICLR 2020) 2020 April 26; Addis Ababa, Ethiopia (Online).

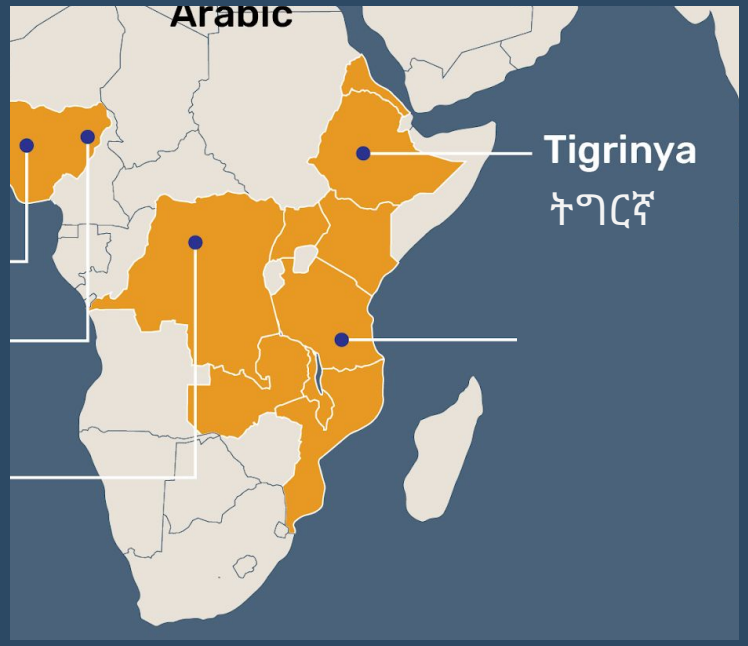
Tigrinya

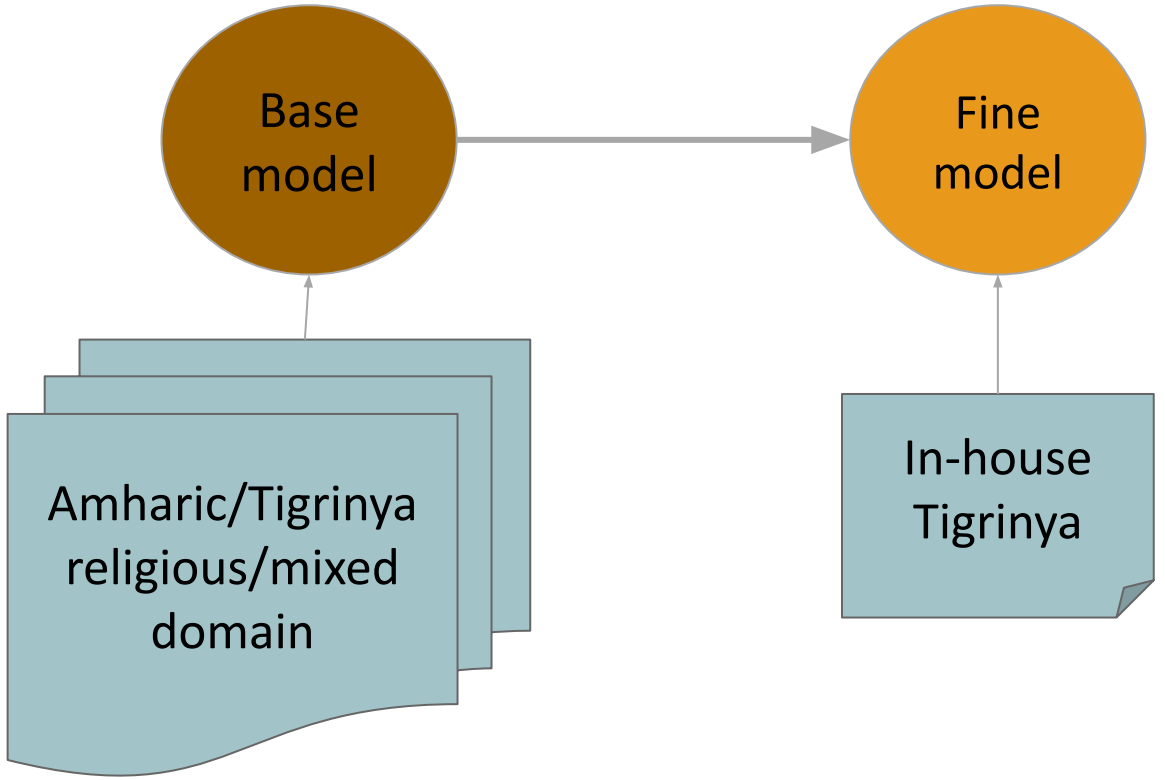




Alp Öktem, Mirko Plitt, Grace Tang. ***Tigrinya Neural Machine Translation with Transfer Learning for Humanitarian Response.*** In: AfricaNLP workshop @ Eighth International Conference on Learning Representations (ICLR 2020) 2020 April 26; Addis Ababa, Ethiopia (Online).

Tigrinya





- Tigrinya-to-English: 23.60% BLEU
  - *+1.32% with cross-lingual transfer*
- English-to-Tigrinya: 9.92% BLEU

Alp Öktem, Mirko Plitt, Grace Tang. ***Tigrinya Neural Machine Translation with Transfer Learning for Humanitarian Response.*** In: AfricaNLP workshop @ Eighth International Conference on Learning Representations (ICLR 2020) 2020 April 26; Addis Ababa, Ethiopia (Online).

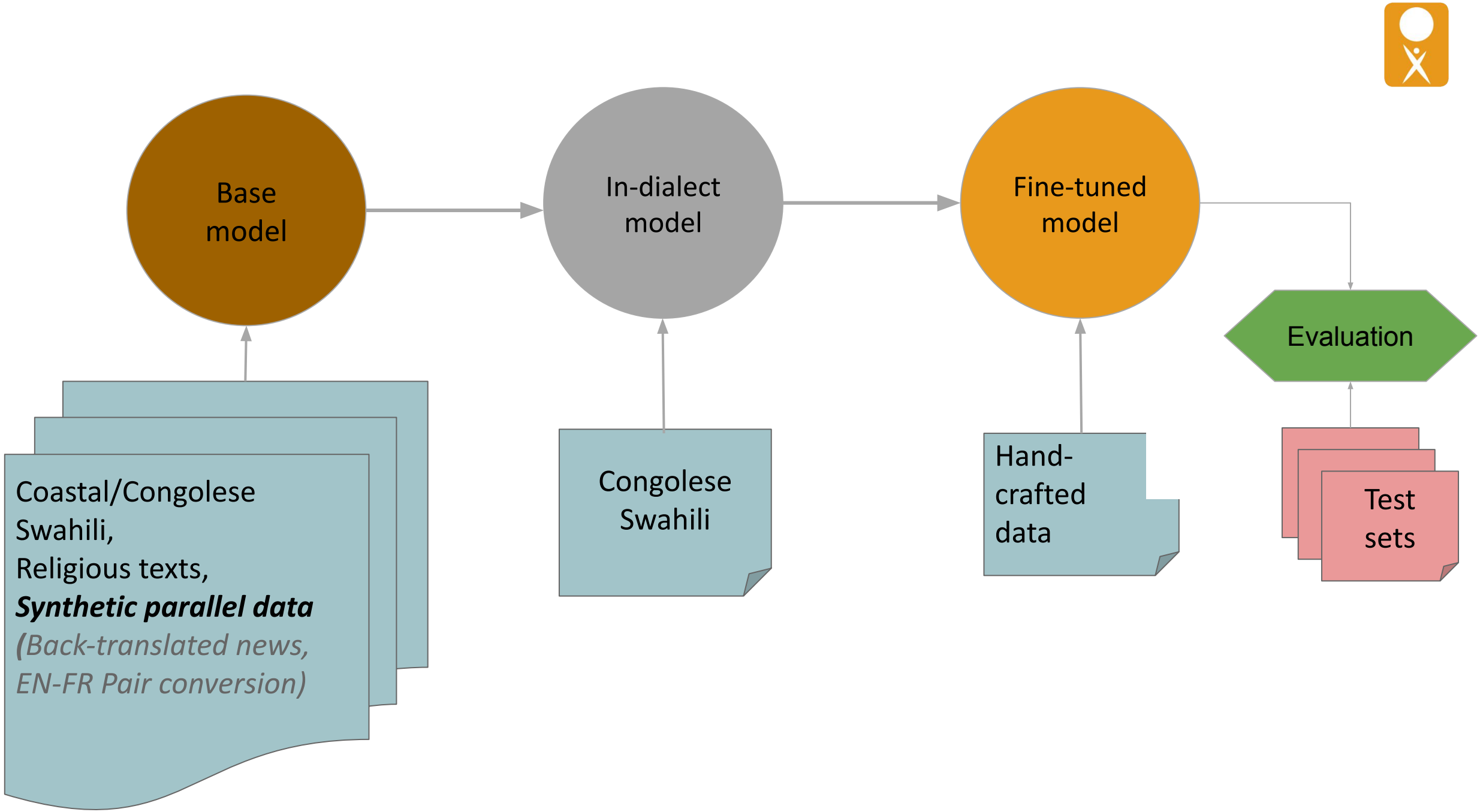
Congolese Swahili

Alp Öktem, Eric DeLuca, Rodrigue Bashizi, Eric Paquin, Grace Tang. ***Congolese Swahili Machine Translation for Humanitarian Response.*** Accepted: Africa-NLP workshop at the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL2021.) 2021 April 19-23; Virtual.

Congolese Swahili

- *“Once the test is negative, the family can take care of the funeral themselves”*
  - **(Coastal)** Mara tu upimaji ukiwa hasi, familia inaweza kushughulikia mazishi wao wenyewe.
  - **(Congo)** Ikiwa tu vipimovinaonesha kama ni mtu ambaye hakuhakikishwa ku kuwa na ugonjwa, familia inaweza kufanya mazishi yenyewe.

Alp Öktem, Eric DeLuca, Rodrigue Bashizi, Eric Paquin, Grace Tang. ***Congolese Swahili Machine Translation for Humanitarian Response.*** In: Africa-NLP workshop at the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL2021.) 2021 April 19-23; Virtual.


Base model
In-dialect model
Fine-tuned model
Evaluation
Coastal/Congolese Swahili,
Religious texts,
Synthetic parallel data
(Back-translated news, EN-FR Pair conversion)
Congolese Swahili
Hand-crafted data
Test sets

	SWC to French	French to SWC
Only Congolese Swahili	17.7	13.0
Coastal Swahili mix	19.4 (+1.7)	15.0 (+2)
Synthetic data augmentation	20.1 (+0.5)	16.5 (+1.5)

Congolese Swahili (SWC) machine translation evaluation (BLEU score percentages) on TICO-19 test set

Language data collection

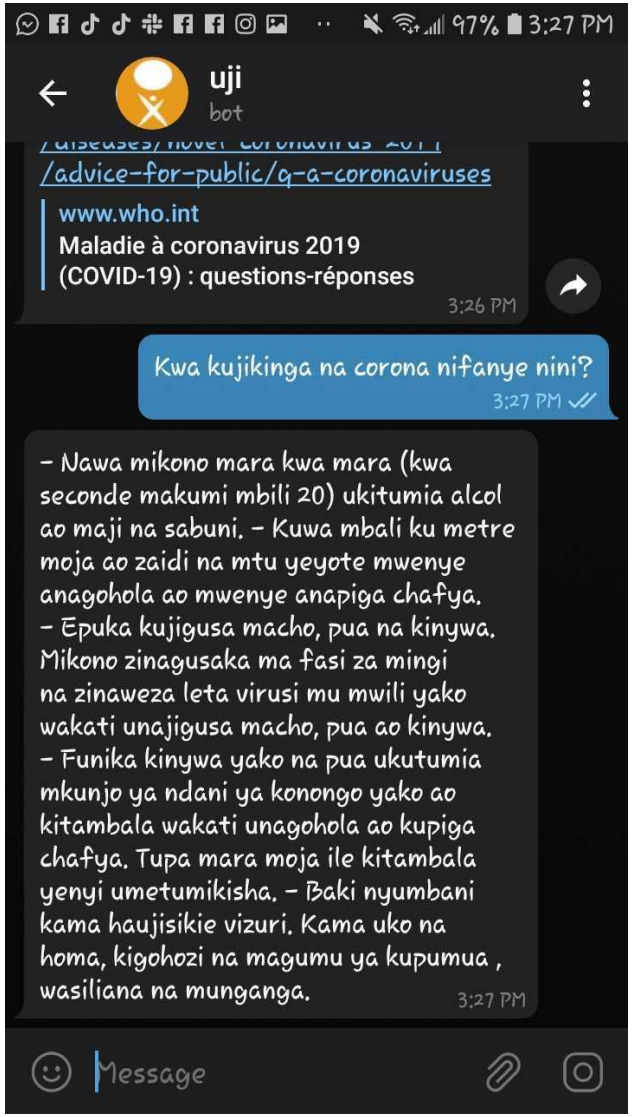
Machine translation

Language data collection

Machine translation

Use-case oriented evaluation

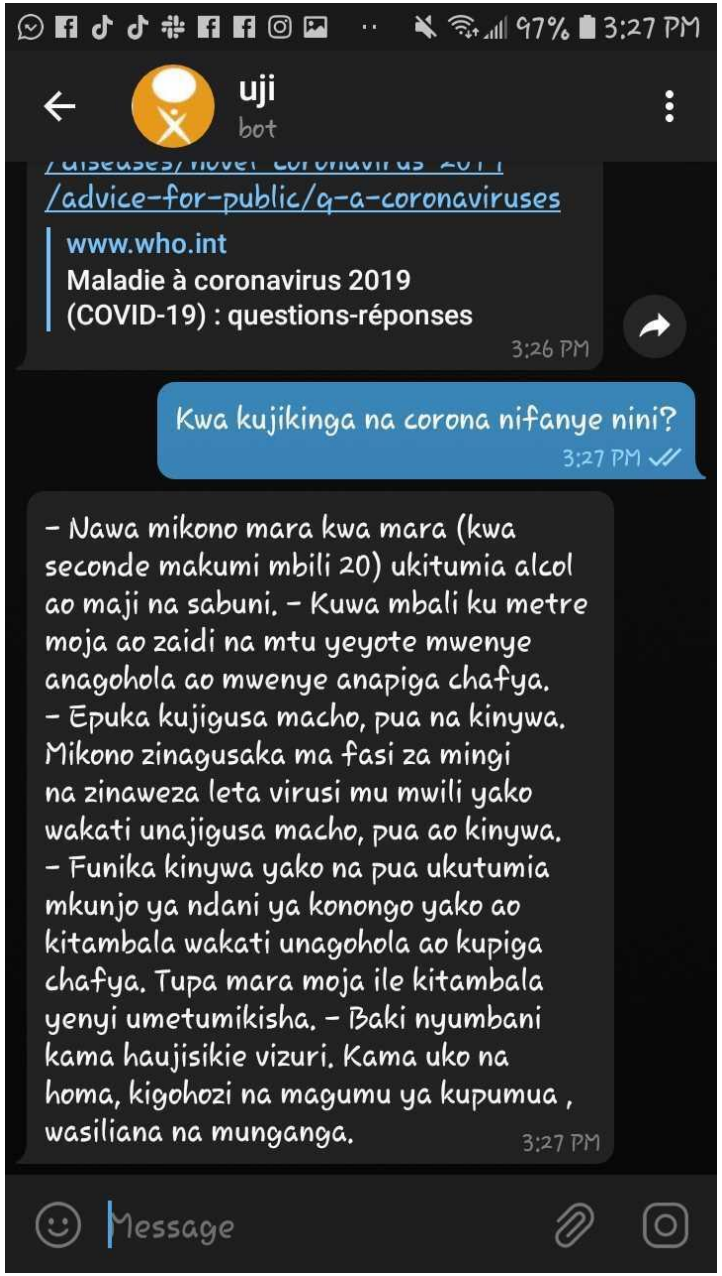
Use-case oriented human evaluation



Talk to Uji - https://t.me/twb_uji_drc_bot

Use-case oriented human evaluation

- Post-editing (FRA->SWC)
- Direct assessment (SWC->FRA)



Use-case oriented human evaluation

- **Post-editing (FRA->SWC)**
- Direct assessment (SWC->FRA)

FR *Des cas de réinfection du COVID-19 ont été signalés mais sont rares. En général, la réinfection signifie qu'une personne a été infectée (est tombée malade) une fois, s'est rétablie, puis est redevenue infectée plus tard. Sur la base de ce que nous savons de virus similaires, certaines réinfections sont attendues.*

Raw-MT *Kesi za uambukizi wa COVID-19 zimeripotiwa lakini hazipatikani kwa urahisi. Kwa ujumla, maambukizi inamaanisha kama mtu aliambukizwa (aliugua) mara moja, s' alirudishwa, na kisha akaambukizwa tena baadaye. Kutokana na kile tunachojua kuhusu virusi kama hivyo, maambukizo mengine yangali mbele.*

PE-MT *Kesi za maambukizi ya COVID-19 zimeripotiwa lakini hazipatikani mara nyingi. Kwa ujumla, maambukizi inamaanisha kama mtu aliambukizwa (aliugua) mara moja, akapona, na kisha akaambukizwa tena baadaye. Kutokana na kile tunachojua kuhusu virusi kama hivyo, maambukizo mengine yangali mbele.*

BLEU	TER	ChrF
55.92	0.37	74.81

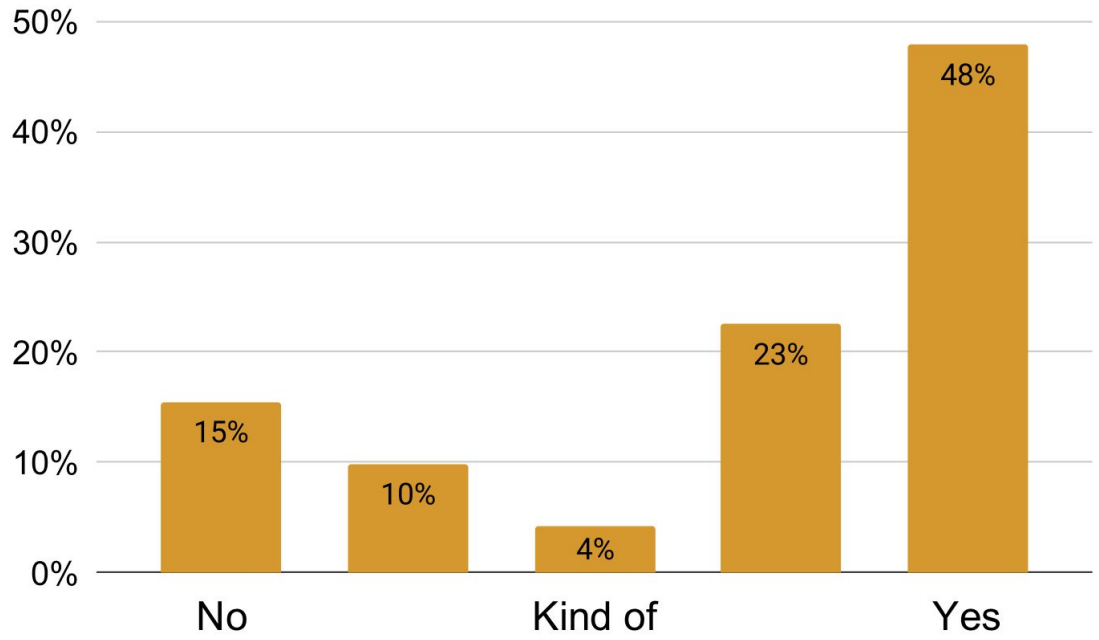
Use-case oriented human evaluation

- Post-editing (FRA->SWC)
- **Direct assessment (SWC->FRA)**

How good is the translation on a scale of 1 to 10?

➢ 6.3±3 /10

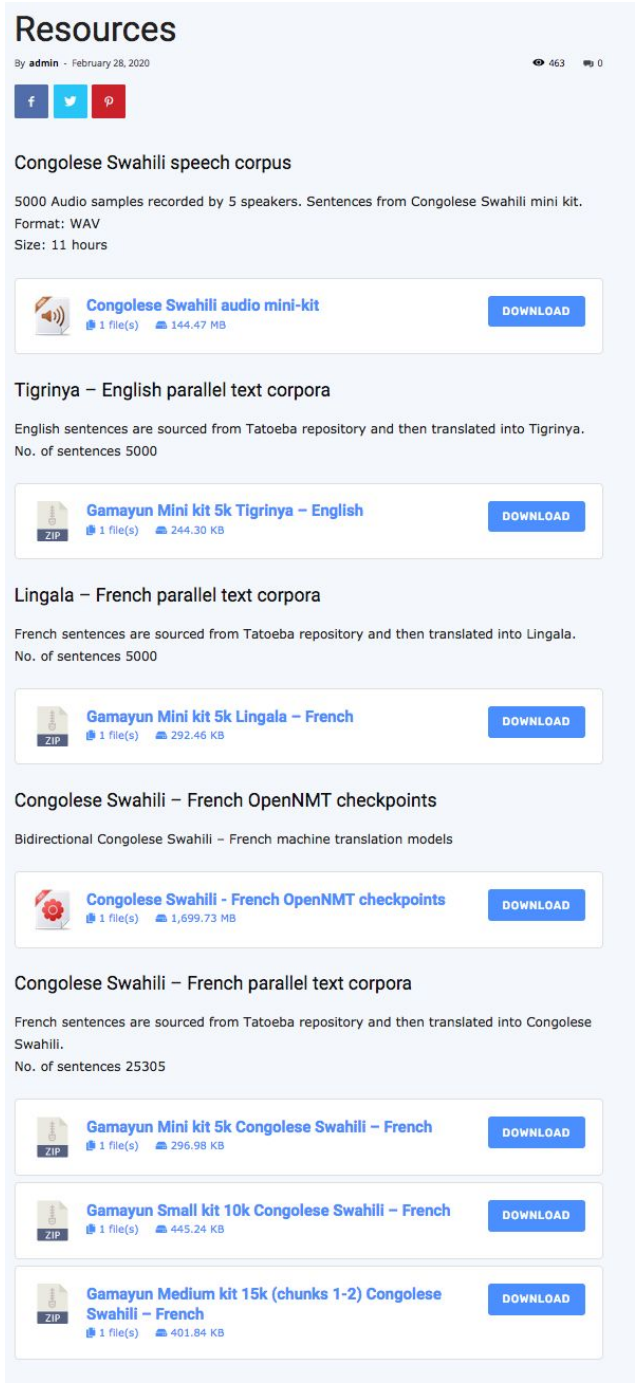
Does this translation convey the main meaning of the source text?



Language data collection

Machine translation

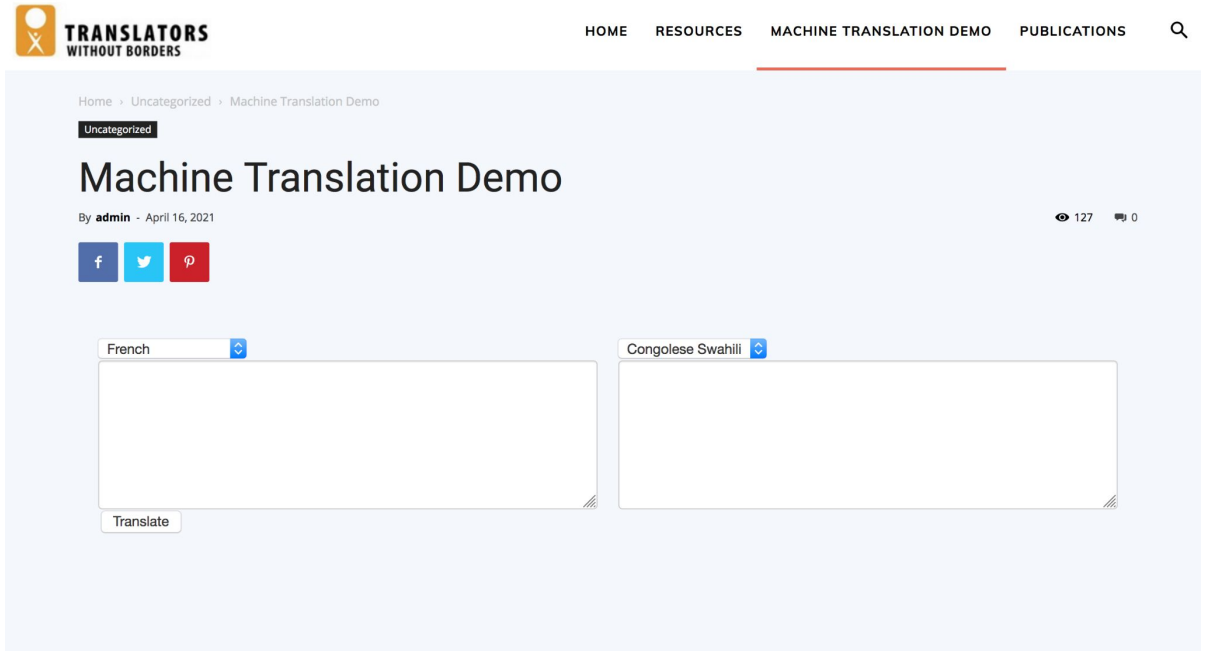
Use-case oriented evaluation



❖ Open data

- ➢ Hand-crafted training/testing data
- ➢ Model checkpoints

❖ Demo translator app



http://gamayun.translatorswb.org


Questions? Feedback?
https://translatorswithoutborders.org
alp@translatorswb.org
@OktemAlp
alpoktem.github.io
TRANSLATORS
WITHOUT BORDERS