NLP beyond the top 100 languages

Antonis Anastasopoulos
antonis@gmu.edu
Some recent trends

![Diagram showing a classifier with 85% for Spam and 15% for Not Spam.]

USE BERT FOR EVERYTHING!!1!
Some recent trends

- BERT
- RoBERTa
- XLM-R
- ELECTRA
- GPT-2
- PaLM
- Chinchila

Who cares

USE BERT FOR EVERYTHING!!1!
Make it multilingual!

Good recap of the current state of multilingual AI: https://ruder.io/state-of-multilingual-ai/
Lang Tech utility is unequally distributed!
Lang Tech utility is unequally distributed!

Compare:
- American English speaker
- Arabic speaker
  - Tunisian vs Egyptian vs …
- Bemba speaker
Global Utility Metrics
Global Utility Metrics

A language technology should be measured by the utility it provides to every person in the world.
Global Utility Metrics

A language technology should be measured by the utility it provides to every person in the world:

\[ M = \sum_i u_i \]
Global Utility Metrics

A language technology should be measured by the utility it provides to every person in the world

\[ M = \sum_i u_i \]

Two problems:
Global Utility Metrics

A language technology should be measured by the utility it provides to **every person in the world**

\[ M = \sum_{i} u_i \]

Two problems:

- **Problem 1**: how to measure utility of an NLP system?
  - Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)
Global Utility Metrics

A language technology should be measured by the utility it provides to every person in the world

\[ M = \sum_{i} u_i \]

Two problems:

• **Problem 1**: how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

• **Problem 2**: how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient \( \tau \).
Global Utility Metrics

A language technology should be measured by the utility it provides to every person in the world

\[ M = \sum_{i} u_i \]

Two problems:

• **Problem 1**: how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

• **Problem 2**: how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

\[ M_\tau = \sum_{l \in \mathcal{L}} d_l^{(\tau)} \cdot u_l \]
Global Utility Metrics

A language technology should be measured by the utility it provides to every person in the world

\[ M = \sum_i u_i \]

Two problems:

• **Problem 1**: how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

• **Problem 2**: how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient \( \tau \).

\[ M_\tau = \sum_{l \in L} \left[ d_l^{(\tau)} \right] \cdot u_l \]

"normalized demand"
Global Utility Metrics

A language technology should be measured by the utility it provides to every person in the world

\[ M = \sum_i u_i \]

Two problems:

- **Problem 1**: how to measure utility of an NLP system?
  - Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

- **Problem 2**: how to consider different utility provided to every person in the world?
  - Measure over subgroups (here, languages), weighted by demand + coefficient \( \tau \).

\[ M_\tau = \sum_{l \in \mathcal{L}} d_l^{(\tau)} \cdot u_l \]

"normalized demand"  "utility"
Global Utility Metrics

A language technology should be measured by the utility it provides to every person in the world

\[ M = \sum_i u_i \]

Two problems:

• **Problem 1**: how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

• **Problem 2**: how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient \( \tau \).

\[ M_\tau = \sum_{l \in L} \left[ d_l(\tau) \cdot u_l \right] \]

\[ d_l(\tau) = \frac{n_l^\tau}{\sum_{l' \in L} n_{l'}^\tau} \]

"normalized demand"  "utility"
Global Utility Metrics

A language technology should be measured by the utility it provides to every person in the world

\[ M = \sum_{i} u_i \]

Two problems:

- **Problem 1:** how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

- **Problem 2:** how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient \( \tau \).

\[ M_{\tau} = \sum_{l \in \mathcal{L}} d_l^{(\tau)} \cdot u_l \]  
\[ d_l^{(\tau)} = \frac{n_l^{\tau}}{\sum_{l' \in \mathcal{L}} n_l^{\tau}} \]

\( \tau = 1 : \) every person equal  

("demographic-average utility")

"normalized demand" "utility"
Global Utility Metrics

A language technology should be measured by the utility it provides to every person in the world

\[ M = \sum_{i} u_i \]

Two problems:

- **Problem 1:** how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

- **Problem 2:** how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient \( \tau \).

\[ M_{\tau} = \sum_{l \in \mathcal{L}} d_l^{(\tau)} \cdot u_l \quad d_l^{(\tau)} = \frac{n_l^{\tau}}{\sum_{l' \in \mathcal{L}} n_l'^{\tau}} \]

\( \tau = 1 \) : every person equal

("demographic-average utility")

\( \tau = 0 \) : every subgroup equal

("linguistic-average utility")
Zooming In (Analysis Tasks)

Dependency Parsing

Inflection
Zooming In (Analysis Tasks)

Dependency Parsing

Inflection
Zooming In (User-facing Tasks)

Question Answering

Natural Language Inference

Speech Synthesis
Zooming In (User-facing Tasks)

Question Answering

Natural Language Inference

Speech Synthesis
Zooming In (Machine Translation)

MT To English

MT To Spanish

MT To Bengali
Going Deeper: Dialects
Going Deeper: Dialects

Very few languages are monoliths!
Very few languages are monoliths!

Need to model dialectal/regional/user variations.
Going Deeper: Dialects

Very few languages are monoliths!

Need to model dialectal/regional/user variations.

Problem: most are spoken (like 45% of all languages)
Very few languages are monoliths!

Need to model dialectal/regional/user variations.

Problem: most are spoken (like 45% of all languages)

SD-QA: Spoken Dialectal Question Answering for the Real World

Fahim Faisal, Shartina Keshava, Md Mahfuz ibn Alam, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{ffaisal, skeshav, malam21, antonis}@gmu.edu

(EMNLP Findings 2021)
SD-QA: Spoken, Dialectal, Multilingual QA
SD-QA: Spoken, Dialectal, Multilingual QA
SD-QA: Spoken, Dialectal, Multilingual QA

Transcription: Teolejia ya dogma ni nini?

Minimal Answer: neno linalotumika hapa kuna anisha fundisho la imani lisilweza kukanushwa na wafuasi wa dini fulani.
SD-QA: Spoken, Dialectal, Multilingual QA

Transcription: Teolejia ya dogma ni nini?

Minimal Answer: neno linalotumika hapa kuna mishina fundisho la imani lisiloweza kukanusha na wafuasi wa dini fulani.
SD-QA: Spoken, Dialectal, Multilingual QA
Let's make a plan
Going beyond the top-100 languages

Make MLMs highly multilingual

Train them on 100 languages

Apply them on the other 6400 languages

Apply them on the *other* 6400 languages
Going beyond the top-100 languages

Make MLMs highly multilingual

Train them on 100 languages

Apply them on the other 6400 languages

Apply them on the *other* 6400 languages

Dominant Written (Latin) Standardized high(ish)-resource
Going beyond the top-100 languages

Make MLMs highly multilingual
Train them on 100 languages

Apply them on the other 6400 languages
Apply them on the *other* 6400 languages

Dominant
Written (Latin)
Standardized
high(ish)-resource

Local
Oral
non-Standardized
Very low-resource
Are all unseen languages equally hard?

When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models

Benjamin Muller†  Antonis Anastasopoulos‡  Benoît Sagot†  Djamé Seddah†
†Inria, Paris, France
‡Department of Computer Science, George Mason University, USA
firstname.lastname@inria.fr  antonis@gmu.edu

(NAACL 2021)

https://github.com/benjamin-mlr/mbert-unseen-languages.git
Are all unseen languages hard?
Are all unseen languages hard?

Some are “easy”
Are all unseen languages hard?

Some are “easy”

Similar languages in pre-training + same script

e.g. Faroese, Swiss German
Are all unseen languages hard?

Some are “easy”

- Similar languages in pre-training + same script
  - e.g. Faroese, Swiss German

Some are “intermediate”
Are all unseen languages hard?

Some are “easy”

- Similar languages in pre-training + same script
  - e.g. Faroese, Swiss German

Some are “intermediate”

- Simple approach (continued fine-tuning) leads to good results
  - e.g. Maltese, Bambara, Wolof
Are all unseen languages hard?

Some are “easy”
  Similar languages in pre-training + same script
  e.g. Faroese, Swiss German

Some seem “hard”

Some are “intermediate”
  Simple approach (continued fine-tuning) leads to good results
  e.g. Maltese, Bambara, Wolof
Are all unseen languages hard?

Some are “easy”

Similar languages in pre-training + same script

e.g. Faroese, Swiss German

Some are “intermediate”

Simple approach (continued fine-tuning) leads to good results

e.g. Maltese, Bambara, Wolof

Some seem “hard”

Hard == Static monolingual embeddings >> mBERT adaptation
Are all unseen languages hard?

Some are “easy”

Similar languages in pre-training + same script

- e.g. Faroese, Swiss German

Some are “intermediate”

Simple approach (continued fine-tuning) leads to good results

- e.g. Maltese, Bambara, Wolof

Some seem “hard”

- Hard == Static monolingual embeddings >> mBERT adaptation

- Similar languages in pre-training BUT different script
Are all unseen languages hard?

Some are “easy”
- Similar languages in pre-training + same script
  - e.g. Faroese, Swiss German

Some are “intermediate”
- Simple approach (continued fine-tuning) leads to good results
  - e.g. Maltese, Bambara, Wolof

Some seem “hard”
- Hard == Static monolingual embeddings >> mBERT adaptation
- Similar languages in pre-training BUT different script
  - e.g. Uyghur, Sorani, Sindhi
Are all unseen languages hard?

Some are “easy”

- Similar languages in pre-training + same script
  - e.g. Faroese, Swiss German

Some are “intermediate”

- Simple approach (continued fine-tuning) leads to good results
  - e.g. Maltese, Bambara, Wolof

Some seem “hard”

Hard == Static monolingual embeddings >> mBERT adaptation

- Similar languages in pre-training BUT different script
  - e.g. Uyghur, Sorani, Sindhi

- Transliteration helps
Doing better by hard-coding linguistic information

Phylogeny-Inspired Adaptation of Multilingual Models to New Languages

Fahim Faisal, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{ffaisal, antonis}@gmu.edu

(AACL 2022)

https://github.com/ffaisal93/adapt_lang_phylogeny
Revisiting Adapters
Revisiting Adapters
Revisiting Adapters
Revisiting Adapters
Revisiting Adapters

...
Revisiting Adapters
Revisiting Adapters
Revisiting Adapters
Revisiting Adapters
Revisiting Adapters
Revisiting Adapters
Revisiting Adapters

![Diagram of layers and adapters](image)
Revisiting Adapters
Revisiting Adapters
Revisiting Adapters
Revisiting Adapters
Revisiting Adapters

Easy zero-shot adaptation to new languages at a low cost (additional parameters)
Revisiting Adapters

Easy zero-shot adaptation to new languages at a low cost (additional parameters)

Avoids catastrophic forgetting
Revisiting Adapters

Easy zero-shot adaptation to new languages at a low cost (additional parameters)

Avoids catastrophic forgetting

Performance comparable to full-model fine-tuning
Revisiting Adapters

Easy zero-shot adaptation to new languages at a low cost (additional parameters)

Avoids catastrophic forgetting

Performance comparable to full-model fine-tuning

Can we do better?
Follow Phylogeny for Parameter Sharing
Follow Phylogeny for Parameter Sharing

For Dutch input
Follow Phylogeny for Parameter Sharing

For Bengali input
Results
Results

DEPENDENCY PARSING

UAS
80
60
40
20
0

[T] [LT] [FGLT] [T] [LT] [FGLT]

GERMANIC (12) URALIC (11)
Results

**DEPENDENCY PARSING**

<table>
<thead>
<tr>
<th>Language</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GERMANIC (12)</td>
<td>70.6</td>
</tr>
<tr>
<td>URALIC (11)</td>
<td>48.3</td>
</tr>
</tbody>
</table>

The chart above shows the UAS (Uparse Accuracy Score) for dependency parsing across different languages. The highest UAS score is 70.6 for GERMANIC (12), while the lowest is 48.3 for URALIC (11).
Results

DEPENDENCY PARSING

GERMANIC (12)

<table>
<thead>
<tr>
<th>Language</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>[T]</td>
<td>70.6</td>
</tr>
<tr>
<td>[LT]</td>
<td>69.2</td>
</tr>
<tr>
<td>[FGLT]</td>
<td></td>
</tr>
</tbody>
</table>

URALIC (11)

<table>
<thead>
<tr>
<th>Language</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>[T]</td>
<td>48.3</td>
</tr>
<tr>
<td>[LT]</td>
<td>51.4</td>
</tr>
<tr>
<td>[FGLT]</td>
<td></td>
</tr>
</tbody>
</table>
Results

DEPENDENCY PARSING

<table>
<thead>
<tr>
<th>Language</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GERMANIC (12)</td>
<td>70.6</td>
</tr>
<tr>
<td></td>
<td>69.2</td>
</tr>
<tr>
<td></td>
<td>72.3</td>
</tr>
<tr>
<td>T</td>
<td>48.3</td>
</tr>
<tr>
<td>LT</td>
<td>51.4</td>
</tr>
<tr>
<td>FGLT</td>
<td>58.3</td>
</tr>
<tr>
<td>URALIC (11)</td>
<td>51.4</td>
</tr>
<tr>
<td>T</td>
<td>69.2</td>
</tr>
<tr>
<td>LT</td>
<td>48.3</td>
</tr>
<tr>
<td>FGLT</td>
<td>58.3</td>
</tr>
</tbody>
</table>

Germanic and Uralic languages show different performance in dependency parsing, with Germanic languages having higher UAS scores than Uralic languages.
Results on unseen languages
Results on unseen languages

DEPENDENCY PARSING

[Diagram showing UAS scores for different language families: Germanic (3), Uralic (8), Tupian (8)]
Results on unseen languages

DEPEDENCY PARSING

<table>
<thead>
<tr>
<th>Language</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GERMANIC</td>
<td>52.5</td>
</tr>
<tr>
<td>URALIC</td>
<td>36.9</td>
</tr>
<tr>
<td>TUPIAN</td>
<td>24.1</td>
</tr>
</tbody>
</table>

Legend:
- [T]
- [LT]
- [FGLT]
Results on unseen languages

Dependency Parsing

UAS

<table>
<thead>
<tr>
<th>Language</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GERMANIC (3)</td>
<td>52.5</td>
</tr>
<tr>
<td>[T]</td>
<td></td>
</tr>
<tr>
<td>[LT]</td>
<td>50.8</td>
</tr>
<tr>
<td>[FGLT]</td>
<td></td>
</tr>
<tr>
<td>GERMANIC (3)</td>
<td></td>
</tr>
<tr>
<td>[T]</td>
<td></td>
</tr>
<tr>
<td>[LT]</td>
<td>36.9</td>
</tr>
<tr>
<td>[FGLT]</td>
<td></td>
</tr>
<tr>
<td>URALIC (8)</td>
<td>41.1</td>
</tr>
<tr>
<td>[T]</td>
<td></td>
</tr>
<tr>
<td>[LT]</td>
<td></td>
</tr>
<tr>
<td>[FGLT]</td>
<td>24.1</td>
</tr>
<tr>
<td>TUPIAN (8)</td>
<td>19.0</td>
</tr>
<tr>
<td>[T]</td>
<td></td>
</tr>
<tr>
<td>[LT]</td>
<td></td>
</tr>
<tr>
<td>[FGLT]</td>
<td></td>
</tr>
</tbody>
</table>
# Results on unseen languages

## DEPENDENCY PARSING

<table>
<thead>
<tr>
<th>Language</th>
<th>[T]</th>
<th>[LT]</th>
<th>[FGLT]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germanic (3)</td>
<td>52.5</td>
<td>50.8</td>
<td>60.1</td>
</tr>
<tr>
<td>Uralic (8)</td>
<td>36.9</td>
<td>41.1</td>
<td>50.5</td>
</tr>
<tr>
<td>Tupian (8)</td>
<td>24.1</td>
<td>19</td>
<td>23.8</td>
</tr>
</tbody>
</table>

The diagram shows the Universal Average Score (UAS) for dependency parsing across different language families.

- **Germanic (3)**: UAS values of 52.5, 50.8, and 60.1
- **Uralic (8)**: UAS values of 36.9, 41.1, and 50.5
- **Tupian (8)**: UAS values of 24.1, 19, and 23.8
Results on unseen languages

**DEPENDEDENCY PARSING**

<table>
<thead>
<tr>
<th>Language</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GERMANIC (3)</td>
<td>52.5</td>
</tr>
<tr>
<td>TUPIAN (8)</td>
<td>19</td>
</tr>
<tr>
<td>URALIC (8)</td>
<td>36.9</td>
</tr>
<tr>
<td>TUPIAN (8)</td>
<td>24.1</td>
</tr>
<tr>
<td>FSLT</td>
<td>50.5</td>
</tr>
<tr>
<td>TUPIAN (8)</td>
<td>23.8</td>
</tr>
</tbody>
</table>

Much larger improvements for *new, unseen* languages.
Results on unseen languages

DEPEDEGENCY PARSING

Much larger improvements for new, unseen languages

You’re just using more parameters!
Ablations
Ablations

DEPENDENCY PARSING ON URALIC LANGS
Ablations

DEPENDENCY PARSING ON URALIC LANGS

<table>
<thead>
<tr>
<th>Language</th>
<th>UAS (DEP PARSING ACC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[LT]</td>
<td>48</td>
</tr>
<tr>
<td>[FGLT]</td>
<td>53.3</td>
</tr>
<tr>
<td>[FGLT] CONSTRAINED</td>
<td></td>
</tr>
<tr>
<td>[LLLT] CONSTRAINED</td>
<td></td>
</tr>
</tbody>
</table>
Ablations

DEPENDENCY PARSING ON URALIC LANGS

<table>
<thead>
<tr>
<th>UAS (DEP PARSING ACC)</th>
<th>[LT]</th>
<th>[FGLT]</th>
<th>[FGLT] CONSTRAINED</th>
<th>[LLLT] CONSTRAINED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>48</td>
<td>53.3</td>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>
Ablations

DEPENDENCY PARSING ON URALIC LANGS

Even constraining to the same number of parameters, still improvements!
Even constraining to the same number of parameters, still improvements!
Is it language sharing or network depth?
Ablations

DEPENDENCY PARSING ON URALIC LANGS

Even constraining to the same number of parameters, still improvements!
Is it language sharing or network depth?
Same idea applied to Translation:

- 2nd best constrained system at WMT Shared Task on Large-Scale Multilingual Systems for African Languages!
No matter what, we need data in these languages.

What data do we need, though?
Few-Shot is the way

Let’s leave script issues aside for a minute —since we can find solutions, e.g.

It seems that we can do great-well with just a few in-domain in-language task data + data augmentation!
Going forward and beyond
Going forward and beyond

Towards More Equitable Question Answering Systems: How Much More Data Do you Need?

Arnab Debnath, Navid Rajabi, Fardina Fathmiul Alam, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{debnath, nrajsbi, falam5, antonis}@gmu.edu

(ACL 2021)

How much data?
Going forward and beyond

Towards More Equitable Question Answering Systems: How Much More Data Do you Need?

Arnob Debnath, Navid Rajabi, Fardina Fathmiul Alam, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{adebnath, nrajabi, falam5, antonis}@gmu.edu

Dataset Geography: Mapping Language Data to Language Users

Fahim Faisal, Yinkai Wang, Antonios Anastasopoulos
Department of Computer Science, George Mason University, USA
{ffaisal, ywang88, antonis}@gmu.edu

(ACL 2021)

How much data?

(ACL 2022)

What data?
Example: (Extractive) Question Answering

We have a lot of English-only datasets for QA (e.g. SQuAD)

Can we leverage them, and investigate few-shot approaches in new languages?

Study on TyDi-QA dataset (7 languages)

Towards More Equitable Question Answering Systems: How Much More Data Do you Need?

Arnab Debnath, Navid Rajabi, Fardina Fathmiul Alam, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{debnath, nrajski, faalam5, antonis}@gmu.edu

(ACL 2021)
Few-Shot adaptation
Few-Shot adaptation

MACRO-AVG F1-SCORE

0 22.5 45 67.5 90

ZERO-SHOT  XL+10/LANG  XL+50/LANG  XL+100/LANG  XL+500/LANG  XL+500/LANG+DATAAUG  SKYLINE
Few-Shot adaptation
Few-Shot adaptation

MACRO-AVG F1-SCORE

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZERO-SHOT (SQUAD)</td>
<td>58</td>
</tr>
<tr>
<td>XL +10/LANG</td>
<td>64.2</td>
</tr>
<tr>
<td>XL +50/LANG</td>
<td></td>
</tr>
<tr>
<td>XL +100/LANG</td>
<td></td>
</tr>
<tr>
<td>XL +500/LANG</td>
<td></td>
</tr>
<tr>
<td>XL +500/LANG +DATAAUG</td>
<td></td>
</tr>
<tr>
<td>SKYLINE</td>
<td></td>
</tr>
</tbody>
</table>
Few-Shot adaptation

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Macro-Avg F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot (SQUAD)</td>
<td>58</td>
</tr>
<tr>
<td>XL +10/LANG</td>
<td>64.2</td>
</tr>
<tr>
<td>XL +50/LANG</td>
<td>68.5</td>
</tr>
<tr>
<td>XL +100/LANG</td>
<td>64.2</td>
</tr>
<tr>
<td>XL +500/LANG</td>
<td>68.5</td>
</tr>
<tr>
<td>XL +500/LANG +DATAAUG</td>
<td>68.5</td>
</tr>
<tr>
<td>SKYLINE</td>
<td>58</td>
</tr>
</tbody>
</table>
Few-Shot adaptation

- MACRO-AVG F1-SCORE:
  - ZERO-SHOT (SQUAD): 58
  - XL +10/LANG: 64.2
  - XL +50/LANG: 68.5
  - XL +100/LANG: 71.7
  - XL +500/LANG +DATAAUG: SKYLINE
Few-Shot adaptation

<table>
<thead>
<tr>
<th>Method</th>
<th>Macro-Avg F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZERO-SHOT (SQUAD)</td>
<td>58</td>
</tr>
<tr>
<td>XL +10/LANG</td>
<td>64.2</td>
</tr>
<tr>
<td>XL +50/LANG</td>
<td>68.5</td>
</tr>
<tr>
<td>XL +100/LANG</td>
<td>71.7</td>
</tr>
<tr>
<td>XL +500/LANG</td>
<td>76.7</td>
</tr>
<tr>
<td>XL +500/LANG +DATAAUG</td>
<td></td>
</tr>
<tr>
<td>SKYLINE</td>
<td></td>
</tr>
</tbody>
</table>
Few-Shot adaptation

<table>
<thead>
<tr>
<th>Description</th>
<th>MACRO-AVG F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZERO-SHOT (SQUAD)</td>
<td>58</td>
</tr>
<tr>
<td>XL +10/LANG</td>
<td>64.2</td>
</tr>
<tr>
<td>XL +50/LANG</td>
<td>68.5</td>
</tr>
<tr>
<td>XL +100/LANG</td>
<td>71.7</td>
</tr>
<tr>
<td>XL +500/LANG</td>
<td>76.7</td>
</tr>
<tr>
<td>XL +500/LANG +DATAAUG</td>
<td>78.7</td>
</tr>
</tbody>
</table>

SKYLINE
Few-Shot adaptation

<table>
<thead>
<tr>
<th>Model Type</th>
<th>MACRO-AVG F1-SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZERO-SHOT (SQUAD)</td>
<td>58</td>
</tr>
<tr>
<td>XL +10/LANG</td>
<td>64.2</td>
</tr>
<tr>
<td>XL +50/LANG</td>
<td>68.5</td>
</tr>
<tr>
<td>XL +100/LANG</td>
<td>71.7</td>
</tr>
<tr>
<td>XL +500/LANG</td>
<td>76.7</td>
</tr>
<tr>
<td>XL +500/LANG +DAIAUG</td>
<td>78.7</td>
</tr>
<tr>
<td>SKYLINE</td>
<td>80.8</td>
</tr>
</tbody>
</table>
Few-Shot adaptation

Within 98% of skyline with less than 10% in-language training data!
So how should we spend our annotation budget?
So how should we spend our annotation budget?

My view:
- Focus on building high-quality evaluation sets
- Spend only a fraction of your budget on training data — combine with stronger baselines
So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

**Terms and Conditions apply**
So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets
Spend only a fraction of your budget on training data — combine with stronger baselines

**Terms and Conditions apply**

4500 training examples in 1 language
So how should we spend our annotation budget?

My view:
Focus on building high-quality evaluation sets
Spend only a fraction of your budget on training data — combine with stronger baselines

**Terms and Conditions apply**

- 4500 training examples in 1 language
- Avg F-score on 6 other languages: 72.3
So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

Avg F-score on 6 other languages:

4500 training examples in 1 language < 1500 training examples in 3 languages

72.3

**Terms and Conditions apply**
So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

**Terms and Conditions apply**

Avg F-score on 6 other languages:

- 4500 training examples in 1 language: 72.3
- 1500 training examples in 3 languages: 74.5
So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

**Terms and Conditions apply**

Avg F-score on 6 other languages:

- 4500 training examples in 1 language < 72.3
- 1500 training examples in 3 languages < 74.5
- 500 training examples in 6 languages
So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

**Terms and Conditions apply

Avg F-score on 6 other languages:

- 4500 training examples in 1 language: 72.3
- 1500 training examples in 3 languages: 74.5
- 500 training examples in 6 languages: 78.7
So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

**Terms and Conditions apply

4500 training examples in 1 language <
1500 training examples in 3 languages <
500 training examples in 6 languages <
250 training examples in 12 languages <

Avg F-score on 6 other languages:

- 72.3
- 74.5
- 78.7
- ...

32
How Representative Are your Data?
How Representative Are your Data?

Where does your data come from?
How Representative Are your Data?

Where does your data come from?

Which speakers are modeled?
How Representative Are your Data?

Where does your data come from?

Which speakers are modeled?

Study for country-level representation

Dataset Geography: Mapping Language Data to Language Users

Fahim Faisal, Yinkai Wang, Antonios Anastasopoulos
Department of Computer Science, George Mason University, USA
{ffaisal, ywang88, antonis}@gmu.edu

(ACL 2022)
Idea
Idea

Named Entities can reveal the information we need!
Idea

Named Entities can reveal the information we need!
Idea

Named Entities can reveal the information we need!
Idea

Named Entities can reveal the information we need!

For a given dataset

RUI COSTA FROM AMADORA PLAYED FOR FIORENTINA
Idea

Named Entities can reveal the information we need!

For a given dataset

- Identify named entities

RUI COSTA FROM AMADOR PLAYED FOR FIORENTINA
Idea

Named Entities can reveal the information we need!

For a given dataset
- Identify named entities
- Link entities to countries through wikidata
Idea

Named Entities can reveal the information we need!

For a given dataset

- ✓ Identify named entities
- ✓ Link entities to countries through wikidata
- ✓ Aggregate through dataset
  - ✓ Representativeness measures
  - ✓ Fairness measures
  - ✓ Visualizations
Dataset Geography

Code & Dataset
https://github.com/ffaisal93/dataset Geography

Project Webpage & Additional Visualizations
https://nlp.cs.gmu.edu/project/datasetmaps
Dataset Geography
Dataset Geography
Dataset Geography

Dataset Map: Masakhaner wolof

Top-10 Represented Countries
Countries Missing: 177 of 243 (72.84%)

Main Countries where language is spoken.
Percentage in-country: 24.20% (Green indicates allocation proportional to the population)
Dataset Geography

Top-10 Represented Countries
Countries Missing: 80 of 243 (32.92%)
What do communities need/want?
What do communities need/want?

Work *with* the communities *for* the communities
What do communities need/want?

Work *with* the communities *for* the communities
What do communities need/want?

Work *with* the communities *for* the communities
What do communities need/want?

Work *with* the communities *for* the communities

---

**Educational Tools for Mapuzugun**

Cristian Ahumada¹  Claudio Gutierrez¹  Antonios Anastasopoulos²

¹Department of Computer Science, Universidad de Chile
²Computer Science Department, George Mason University

ahumada.86@gmail.com  cgtierr@ecc.uchile.cl  antonis@gnu.edu
What do communities need/want?

Work *with* the communities *for* the communities
What do communities need/want?

Work with the communities for the communities

BembaSpeech: A Speech Recognition Corpus for the Bemba Language

Claytone Sikasote¹
Department of Computer Science
University of Zambia
Zambia
claytone.sikasote@cs.unza.zm

Antonios Anastasopoulos
Department of Computer Science
George Mason University
USA
antonis@gmu.edu

BIG-C: Multimodal Dataset for the Bemba Language

Claytone Sikasote², Eunice Mukonde³, and Antonios Anastasopoulos²
¹Department of Computer Science, University of Zambia, Zambia
²Department of Computer Science, George Mason University, USA
claytone.sikasote@cs.unza.zm, antonis@gmu.edu
What do communities need/want?

Work *with* the communities *for* the communities

---

**BembaSpeech: A Speech Recognition Corpus for the Bemba Language**

Claytone Sikasote¹
Department of Computer Science
University of Zambia
Zambia
claytone.sikasote@cs.unza.zm

Antonios Anastasopoulos
Department of Computer Science
George Mason University
USA
antonis@gmu.edu

---

**BIG-C: Multimodal Dataset for the Bemba Language**

Claytone Sikasote², Eunice Mukonde³, and Antonios Anastasopoulos²

¹Department of Computer Science, University of Zambia, Zambia
²Department of Computer Science, George Mason University, USA
claytone.sikasote@cs.unza.zm, antonis@gmu.edu
Thank you!

Shoutout to collaborators:
  Graham Neubig, Damian Blasi,
  Benjamin Muller, Benoît Sagot,
  Djamé Seddah

And students:
  Fahim Faisal, Sharlina Keshava,
  Mahfuz ibn Alam, Yinkai Wang

Other things I’m working on:
- NLP for endangered languages (e.g. OCR for scanned documents from Latin America, building basic tools for Griko, Mapudungun, Pomak)
- NLP for linguists (Machine-aided annotation)
- Machine Translation from/into dialects
- Cross-Lingual and Cross-Cultural Fairness
- Geospatial Language Understanding and Navigation
- SLT for Crisis Response
- ...

GMU and GMNLP is hiring!
Faculty/postdocs/PhD students
The amount of data, labeled or unlabeled, varies wildly across languages!

Image from Joshi et al 2020