The State and Fate of Linguistic Diversity in the NLP world


Data Science in India – IKDD
The Actual State

<table>
<thead>
<tr>
<th></th>
<th>Dutch</th>
<th>Somali</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Speakers</td>
<td>29M</td>
<td>18M</td>
</tr>
<tr>
<td>#Resources (LDC+ELRA)</td>
<td>69</td>
<td>2</td>
</tr>
<tr>
<td>Other details</td>
<td>SOTA</td>
<td>Very few, inferior translation systems</td>
</tr>
</tbody>
</table>

Dutch Translation
De tijger bewoog over het gras

Somali Translation
Digirta ayaa ku dul dhaqaaqday cawska

English Translation
The tiger moved across the grass

English Translation
Beans moved on grass

The tiger moved across the grass
### The Actual State

<table>
<thead>
<tr>
<th></th>
<th>Dutch</th>
<th>Somali</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>#Speakers</strong></td>
<td>29M</td>
<td>18M</td>
</tr>
<tr>
<td><strong>#Resources</strong></td>
<td>69</td>
<td>2</td>
</tr>
<tr>
<td><strong>(LDC+ELRA)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other details</strong></td>
<td>SOTA translation systems</td>
<td>Very few, inferior translation systems</td>
</tr>
</tbody>
</table>

(a) ACL + NAACL + EACL + EMNLP  
(b) LREC + WS
The Questions

How has the fate of different languages changed with current language technologies?

1. How many resources are available across the World's languages, and do they correlate with the number of speakers?

2. Which typological features have NLP systems been exposed to? Which features have been underrepresented?

3. How inclusive has ACL been in conducting and publishing research for different languages?

4. Does resource availability influence the research questions and publication venue?

5. What role does an individual researcher or community have in bridging the resource divide?
The Language Taxonomy

Setup

**Labeled data**
LDC catalog, ELRA Map

**Unlabeled data**
Wikipedia pages (used in pretraining language models)
The Language Taxonomy

Visualization
The Language Taxonomy

Visualization

Class 0 (The left-behinds) - Gondi, Mundari
Class 1 (The Scraping-Bys) - Bhojpuri, Assamese
Class 2 (The Hopefuls) - Konkani, Wolof
Class 3 (The Rising Stars) - Tamil, Marathi
Class 4 (The Underdogs) - Bengali, Hindi
Class 5 (The Winners) - English, French
The Language Taxonomy
Large population left behind

<table>
<thead>
<tr>
<th>Class</th>
<th>5 Example Languages</th>
<th>#Langs</th>
<th>#Speakers</th>
<th>% of Total Langs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Dahalo, Warlpiri, Popoloca, Wallisian, Bora</td>
<td>2191</td>
<td>1.2B</td>
<td>88.38%</td>
</tr>
<tr>
<td>1</td>
<td>Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo</td>
<td>222</td>
<td>30M</td>
<td>5.49%</td>
</tr>
<tr>
<td>2</td>
<td>Zulu, Konkani, Lao, Maltese, Irish</td>
<td>19</td>
<td>5.7M</td>
<td>0.36%</td>
</tr>
<tr>
<td>3</td>
<td>Indonesian, Ukrainian, Cebuano, Afrikaans, Hebrew</td>
<td>28</td>
<td>1.8B</td>
<td>4.42%</td>
</tr>
<tr>
<td>4</td>
<td>Russian, Hungarian, Vietnamese, Dutch, Korean</td>
<td>18</td>
<td>2.2B</td>
<td>1.07%</td>
</tr>
<tr>
<td>5</td>
<td>English, Spanish, German, Japanese, French</td>
<td>7</td>
<td>2.5B</td>
<td>0.28%</td>
</tr>
</tbody>
</table>
Typological features refer to properties/attributes of a language.

- Categories in languages of classes 0, 1, 2 but not 3, 4, 5 are ‘ignored’ categories.

- We then look at typological features with most ‘ignored’ vs. least ‘ignored’ categories.
Typological Representation
Far-reaching repercussions

Table 2: Most and Least ‘ignored’ typological features, the number of categories in each feature which have been ignored, and the number of languages which contain this feature.
## Typological Representation

Transfer Learning for Family Languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Class</th>
<th>#Speakers</th>
<th>‘Ignored’</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amharic</td>
<td>2</td>
<td>22M</td>
<td>9</td>
<td>60.71</td>
</tr>
<tr>
<td>Arabic</td>
<td>4</td>
<td>300M</td>
<td>0</td>
<td>7.8</td>
</tr>
</tbody>
</table>

Semitic Family
Language and Conference

What?

Language technologies - driven by NLP Research

<table>
<thead>
<tr>
<th>Conference</th>
<th>h-index</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACL/NAACL/EMNLP/EACL</td>
<td>106/61/88/36</td>
<td>Data-Driven lately</td>
</tr>
<tr>
<td>CL (Journal)</td>
<td>25</td>
<td>Computational Linguistics focused</td>
</tr>
<tr>
<td>COLING</td>
<td>41</td>
<td>Oldest conference</td>
</tr>
<tr>
<td>LREC</td>
<td>45</td>
<td>Multilingual Research</td>
</tr>
<tr>
<td>WS (Workshop Proceedings)</td>
<td>n/a</td>
<td>Factoring papers accepted in workshops of above conferences</td>
</tr>
</tbody>
</table>
Language and Conference

Year-wise Language Occurrence

- Understand how multilinguality is changing over conference iterations
- Language mentions in papers are a measure for language inclusion
- Use Entropy as a unified measure to calculate skew in language distribution for a conference iteration.
- No. of languages = \(2^{\text{entropy}}\)
Language and Conference

Year-wise Language Occurrence

- Understand how multilinguality is changing over conference iterations
- Language mentions in papers are a measure for language inclusion
- Use Entropy as a unified measure to calculate skew in language distribution for a conference iteration.
- No. of languages $= (2)^{\text{entropy}}$

Conference = ACL 2019

| Langs  | af | am | ar | de | en | es | hi | it | pt | tr | vi | ....... | zh |
|--------|----|----|----|----|----|----|----|----|----|----|----|         |    |
| All Papers | 3  | 4  | 10 | 9  | 31 | 11 | 7  | 9  | 8  | 4  | 1  | ....... |
|          | 10 |    |    |    |    |    |    |    |    |    |    |         |    |


Language and Conference

Year-wise Language Occurrence

- Understand how multilinguality is changing over conference iterations
- Language mentions in papers are a measure for language inclusion
- Use Entropy as a unified measure to calculate skew in language distribution for a conference iteration.
- No. of languages = $2^{\text{entropy}}$

Conference = ACL 2019

Entropy

4.1
Language and Conference
Year-wise Language Occurrence

- Understand how multilinguality is changing over conference iterations
- Language mentions in papers are a measure for language inclusion
- Use Entropy as a unified measure to calculate skew in language distribution for a conference iteration.
- No. of languages = \(2^{\text{entropy}}\)
Language and Conference

Year-wise Language Occurrence

(a) $c = \text{ACL}$  
(b) $c = \text{NAACL}$  
(c) $c = \text{EMNLP}$  
(d) $c = \text{EACL}$  
(e) $c = \text{COLING}$  

(f) $c = \text{CL}$  
(g) $c = \text{WS}$  
(h) $c = \text{CONLL}$  
(i) $c = \text{SEMEVAL}$  
(j) $c = \text{LREC}$
Determine the standing of each language class in a conference.

\[
\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}
\]

\(\text{rank}_i \rightarrow\) language rank in a particular conference ordered by mention frequency.

\(Q \rightarrow\) number of languages in each class.
Heterogenous Entity Embeddings

Motivation

Previous analysis indicates variance in acceptance of different languages across different NLP conferences.

Vanilla statistics fail to capture the subtle nuances in the data that might be affecting these outcomes.

Embeddings have been shown to capture complex relationships directly from the data without supervision.

Proposal: Jointly learn the representations of Conferences, Authors and Languages, collectively termed as Entities.
Heterogeneous Entity Embeddings

Spatial Representation

Long distance relationships?

Time waits for no conference

Friends?

Languages
- Class 1
- Class 2
- Class 3
- Class 4
- Class 5
Heterogenous Entity Embeddings

Role of Community

- Mean Reciprocal Rank (MRR) of a language signifies how many authors in the research community are exclusively close to this language.

- Higher MRR indicates more focused research community

<table>
<thead>
<tr>
<th>Class</th>
<th>MRR(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.69146</td>
</tr>
<tr>
<td>1</td>
<td>0.52585</td>
</tr>
<tr>
<td>2</td>
<td>0.45265</td>
</tr>
<tr>
<td>3</td>
<td>0.52670</td>
</tr>
<tr>
<td>4</td>
<td>0.47795</td>
</tr>
<tr>
<td>5</td>
<td>0.51471</td>
</tr>
</tbody>
</table>

Not all superheroes wear capes
Takeaways
Recommendations

Evident Taxonomy
Typology Consideration

Inclusive Conferences
Focused Communities

A call to look at the language disparity at the conferences
Linguistic Diversity & Inclusion clauses

aka.ms/statefate