Natural Language Processing

Lecture 22: NLP for Low-Resource Languages

Some *slides & instruction ideas borrowed from:*
Greg Durret, Mohit Iyyer, & Mar'Aurelio Ranzato

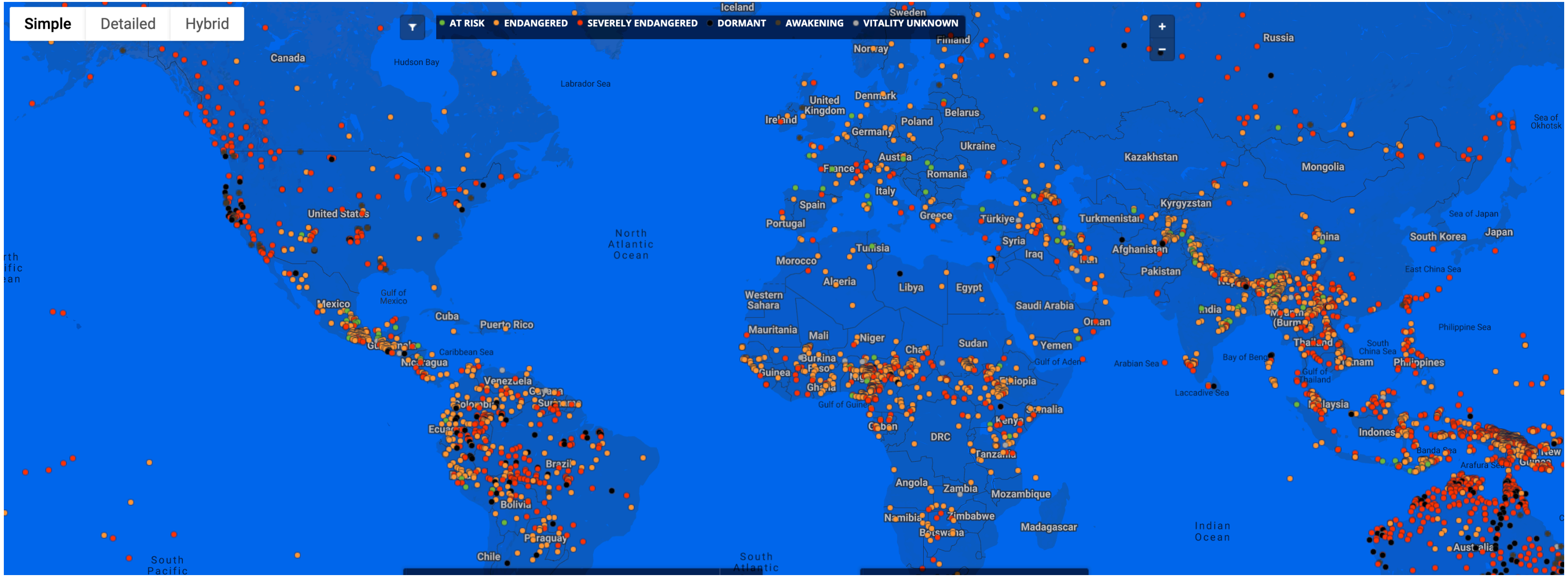
Logistics

- Homework 6 due this Thursday (April 18)
- AP2 and 259 Mid-project reports are being graded.
- AP3 is due April 26
- Tonight: NLP for low resource languages

So far …

- Mostly: NLP for English
- Other languages:
    - Machine Translation
    - Tokenization
    - Parsing & Semantics:
        - Universal Dependency Bank
        - FrameNet

Languages of the World

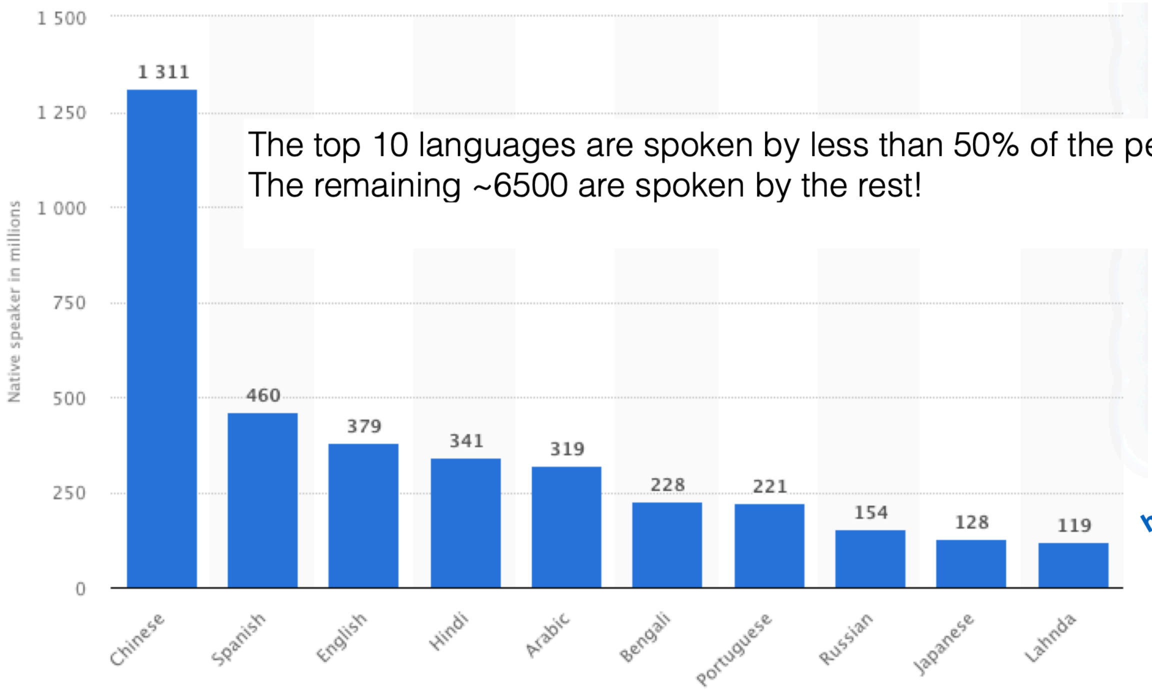


https://endangeredlanguages.com/

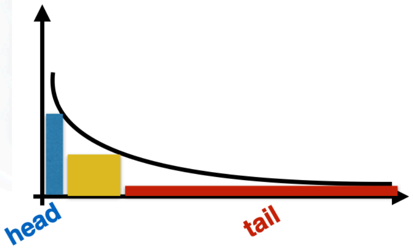
Languages of the World

- 6500+ languages around the world
- ~70% of the world don’t speak English.
- Only 10%- of the world are native English speakers.

NLP Ethics: Exclusion of the underprivileged

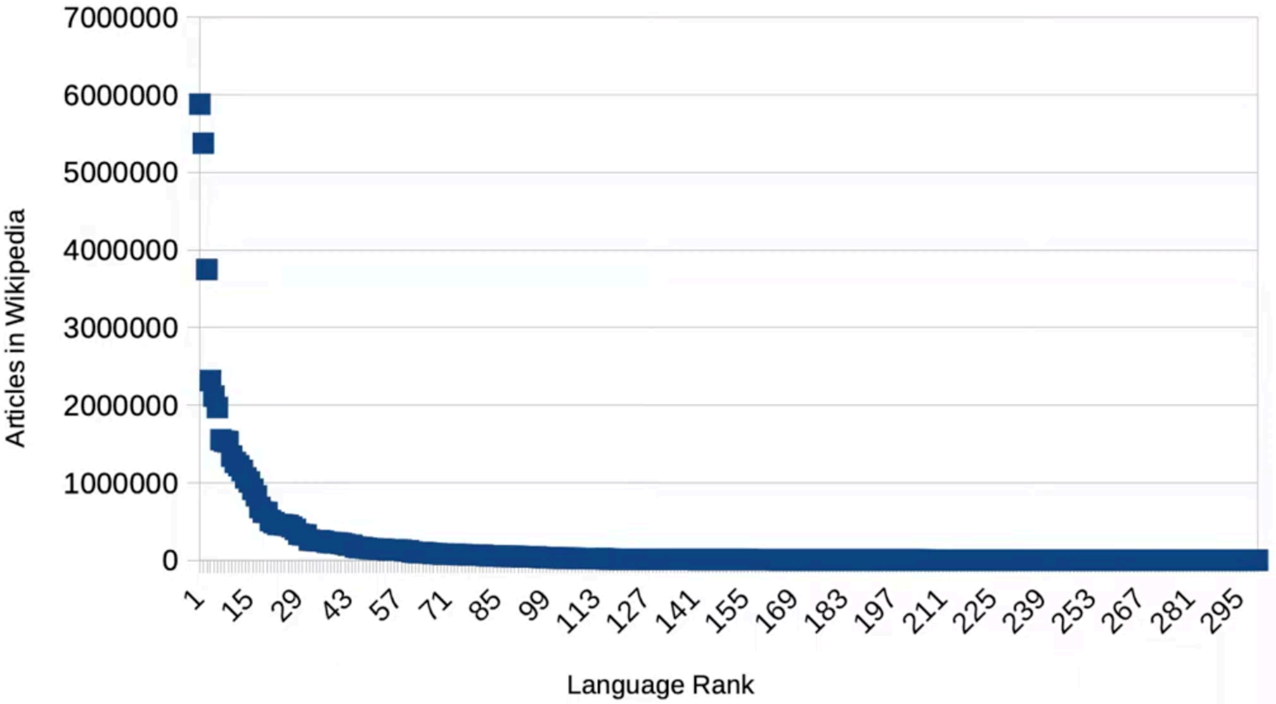


The top 10 languages are spoken by less than 50% of the people.
The remaining ~6500 are spoken by the rest!



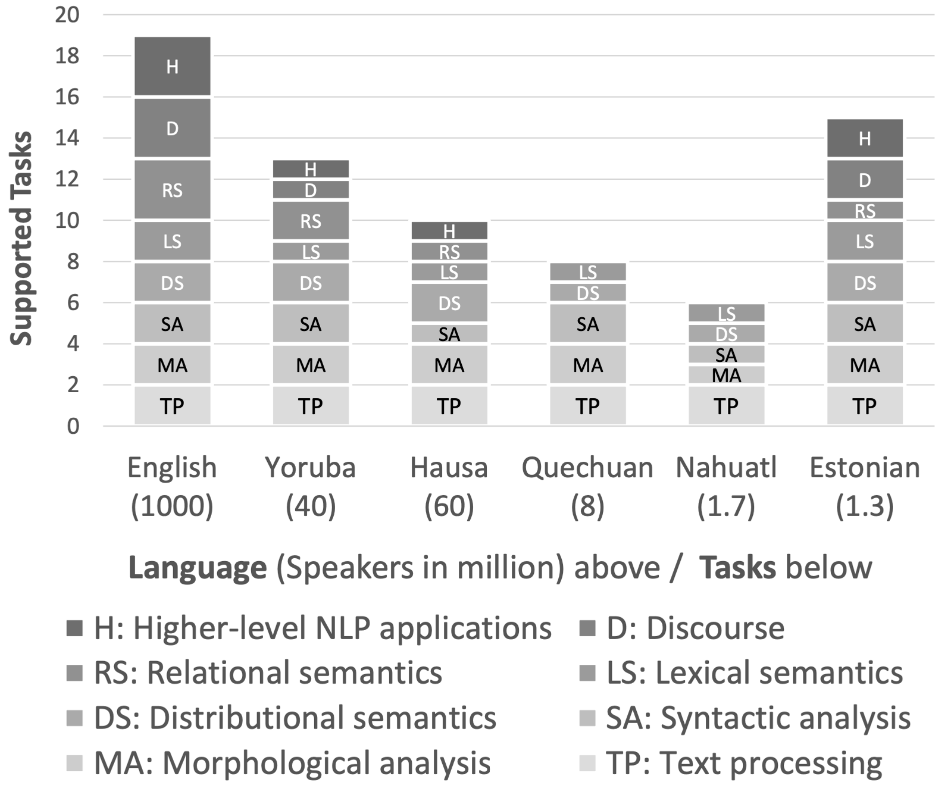
https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/

Data



7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
Articles in Wikipedia
1
15
29
43
57
71
85
99
113
127
141
155
169
183
197
211
225
239
253
267
281
295
Language Rank

NLP Beyond English



NLP for low resource languages

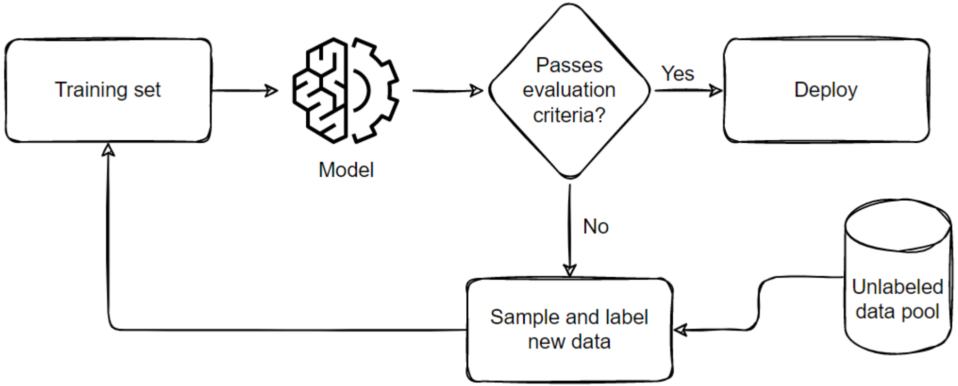
- 310 languages that have at least 1M speakers each (Eberhard et al 2019)
- **Goal:** supporting tech development ⇨ increasing participation in a digital world
- The low-resource setting can be applied for non main-stream domains of high resource languages too.
- Bender rule: clarifying the language of focus in publications.

Generating Additional Data

- Shortage of labeled data for supervised learning is the most prevalent challenge
    - Annotation with Active Learning
    - Data Augmentation
    - Cross-lingual projection

Annotation by Active Learning

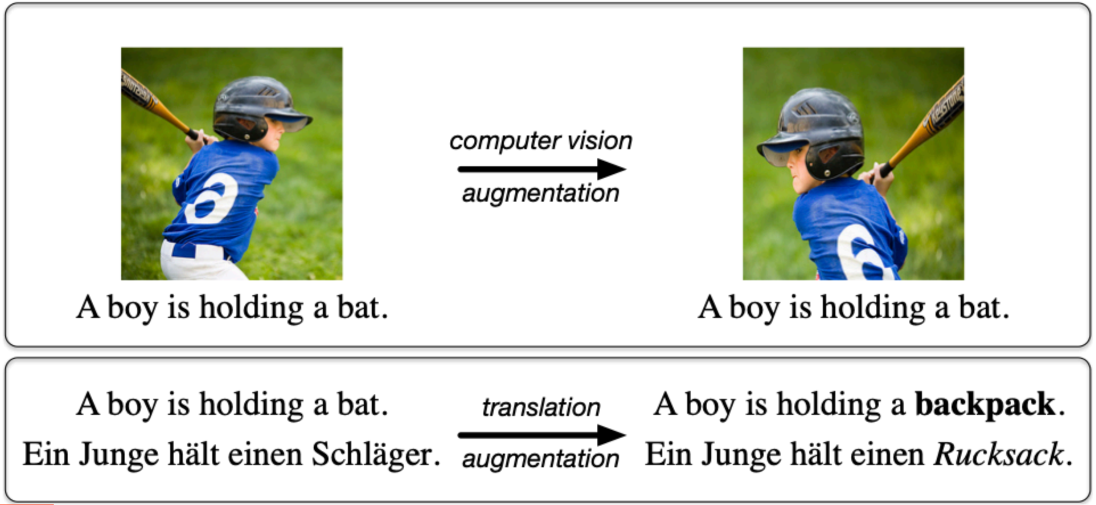
- Optimizing the new annotation iteratively



https://keras.io/examples/nlp/active_learning_review_classification/

Data Augmentation

- Expand your data by augmenting the (small) existing ones.

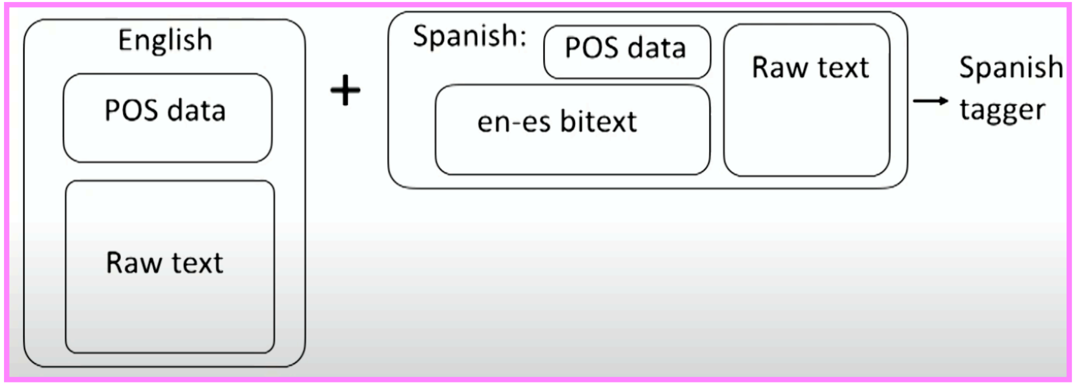


Fadaee et al 2017

Challenge: Scaling can result in noisy data.

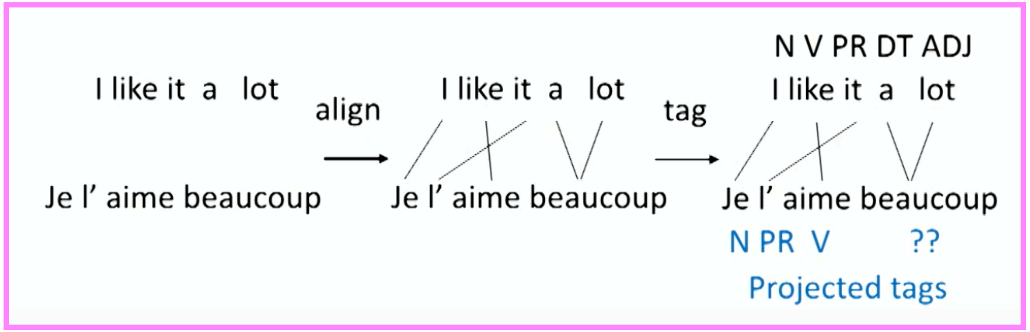
Weak Supervision

- Leveraging from MT data to create labeled data for other tasks.



Cross Lingual Projection

- Use word alignments to project the labels across.
- Partially noisy data, better than no data.



Xi & Hwa 2005, Das & Petrov 2011

Cross Lingual Projection

- Use machine translation (and its word/phrase alignments) to project the labels across.
- Partially noisy data, better than no data.

Khalil et al 2019, Amjad et al 2020

Cross Lingual Projection

- Use machine translation (and its word/phrase alignments) to project the labels across.
- Partially noisy data, better than no data.

Challenge: Availability of parallel data/MT

Khalil et al 2019, Amjad et al 2020

Transfer Learning

- A lot of neural-based methodologies for dense representation and modeling are supposedly language agnostic.
- Word-piece tokenization, Byte-pair-encoding, etc. address a lot of morphological differences —> pre-trained embedding for 270+ languages
- Monolingual BERT has been applied successfully to many languages

Transfer Learning

- A lot of neural-based methodologies for dense representation and modeling are supposedly language agnostic.

- Word-piece tokenization, Byte-pair-encoding, etc. address a lot of morphological differences —> pre-trained embedding for 270+ languages

- Monolingual BERT has been applied successfully to many languages

Challenge: Availability and diversity of unlabeled data for low resource languages. Word embeddings quality can vary.

Transfer Learning

- A lot of neural-based methodology for dense representation and modeling are language agnostic
- Word-piece tokenization, Byte-pair-encoding, etc. address a lot of morphological differences —> pre-trained embedding for 270+ languages
- Monolingual BERT has been applied successfully to many languages
- What about pre-training a shared pre-trained model?
    - **Multi-lingual models**

Multilingual Models

- Combining data into one multilingual model
    - Multilingual BERT, XLM-RoBERTa

Cross-lingual Zero Shot Learning

- **Goal:** We have labeled data for task **X** in **high resource language**. We want a model for task **X** in a **low resource language**.

- **Idea:** Leverage the resources for the high resource language

-

Cross-lingual Zero Shot Learning

- **Goal:** We have labeled data for task **X** in **high resource language**. We want a model for task **X** in a **low resource language**.
- **Idea:** Leverage the resources for the high resource language.
- **Zero-shot:** Fine-tune the multilingual backbone with the task X with the high resource language data (and flexible prompts/instructions) towards generalizing for the low resource languages.
    - NER (lin et al, 2019), reading comprehension (Hsu et al 2019), Parsing (Muller et al 2020)
- Few shot: Add small set (10-100) of low-resource labeled data

Transfer Learning

- Low resource languages in multi-lingual pre-trained language models.
- **Challenge**: Availability of diversity of data for low resource languages
    - Word embedding quality can vary a lot.