Analyzing NLP model performance on high-resource and low-resource language pairs

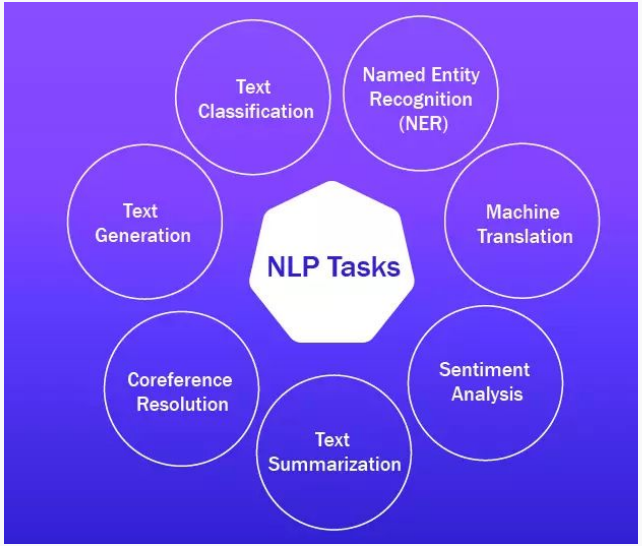
Angelic McPherson

01

Background

What is NLP?

- NLP stands for **Natural Language Processing**, a field of machine learning that allows computers to learn, manipulate, and generate language
- Examples: ChatGPT, Siri, Grammarly, Google Translate



Training a model

Data Collection

- Type? (audio, text, etc.)
- Representative and diverse?
- Annotated

Data processing

- Clean dataset (filter words, repeated sentences, etc.)
- Create test and train sets

Feature Extraction

- Find aspects of the data the model will learn

Training a model

Training

- Choose an appropriate model
- Train it on train dataset
- Select hyperparameters

Evaluation

- Test model on test set
- Compute metrics

Experimentation

- Tune hyperparameters based on results
- Find peek metrics

Key Points

- **Data** is the most important aspect of training a model
- All models learn **their dataset** for training and testing

- **Question: What if we do not have the data for our task?**

02

Research problem

Research Problem

High Resource vs. Low resource Languages

- Majority of NLP models are trained on **high-resource languages**; languages with rich datasets (English, Chinese, Spanish)
- **Low-resource languages** do not have sufficient data available to run analysis
- **Issue:** Models have bias toward high-resource languages, but may need to process input from low-resource languages
    - Ex: Siri or Alexa

Research Problem

High Resource vs. Low resource Languages

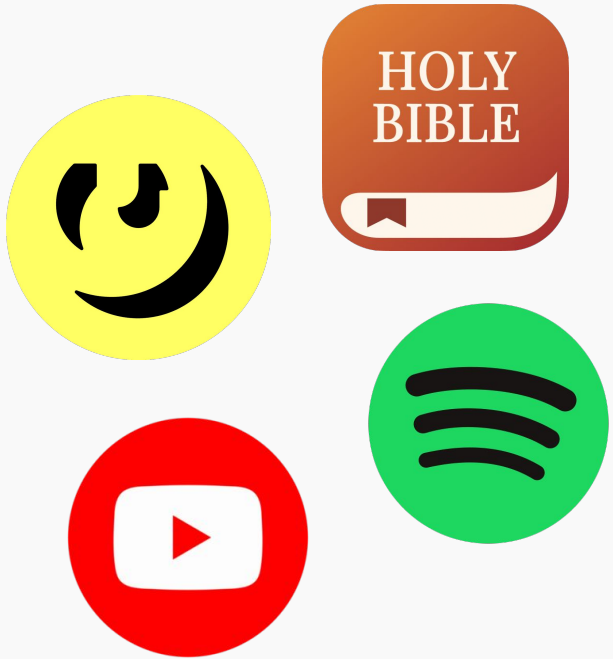
- **Hypothesis:** If we have a pair of linguistically similar high-resource and low-resource languages, can we find a connection between NLP model performance?
- **My language pair:** Jamaican Patois (Creole) and English

03

Methods

Dataset

- Dancehall and reggae lyrics from Spotify and YouTube
- Jamaican Bible
- Approx 200 sentences
- Range in topic and complexity
- Processing: cleaning (no repeated lines) and tokenizing



Tasks

POS Tagging

Parts of speech: nouns, adjectives, verbs, etc.

Dependency Parsing

Grammatical dependencies: clauses, indirect/direct objects, root verbs, etc.

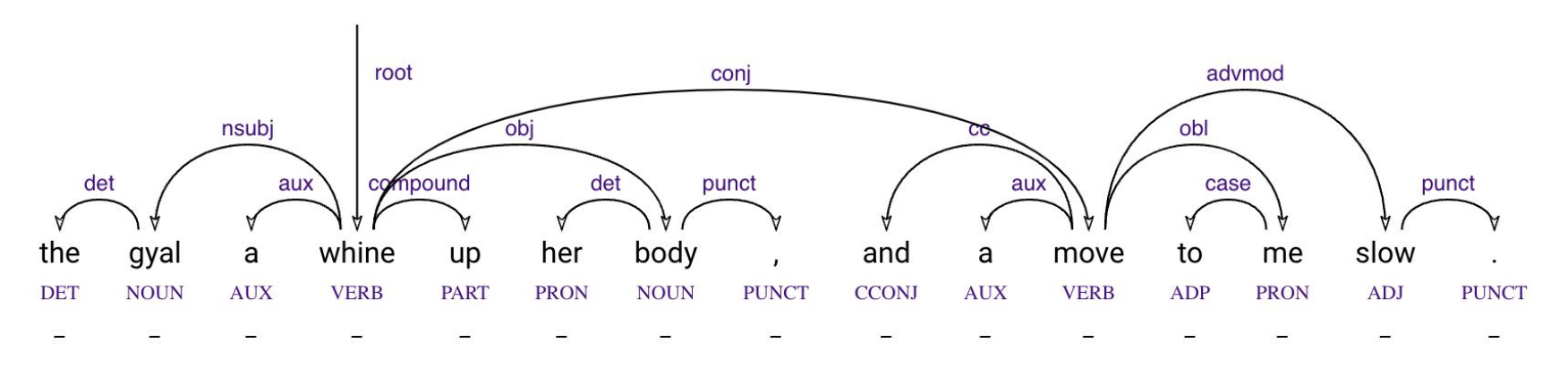
Experimentation

Train models for each task

Analysis

Analyze results to draw conclusions

POS Tagging & Dependency Parsing



37 Dependency Labels

17 POS Labels

04

Experimentation and Results

Testing the models

Cross-lingual Testing

Assumption: JP and English are linguistically similar
Hypothesis: Models should be able to perform cross-lingually
Procedure: Establish baselines and perform testing across languages

1. Test JP Model on JP dataset for baseline
2. Test JP Model on English
3. Test English Model on English dataset for baseline
4. Test English Model on JP
5. Compare and Analyze results

POS Model Details

Infrastructure and Hyperparameters

- Pretrained model with base of RoBERTa infrastructure
  - NLP model released by Facebook

- Learning rate: 0.0001
- LR Scheduler: cosine_with_restarts
- Num epochs: 10
- Batch size: 8
- Weight decay: 0.01

Dependency Model Details

Infrastructure and Hyperparameters

- Pretrained model with base of RoBERTa infrastructure
  - NLP model released by Facebook

- Learning rate: 0.001
- LR Scheduler: Cosine with restarts
- Num epochs: 12
- Batch size: 8
- Weight decay: 0.1
- Warm-up steps: 3000

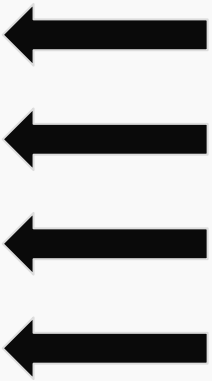
Metrics

- **True positive:** Label correctly predicted where it exists
- **False positive:** Label predicted where it doesn't exist
- **True negative:** Label correctly not predicted where it doesn't exist
- **False negative:** Label not predicted where it should exist

- **Precision:** %of predicted positive labels that were actually correct
- **Recall:** % of actual positive labels that the model successfully found
- **F1:** Balances both false positives and false negatives.
- **Confusion Matrix:** Table describing predicted and actual labels. The diagonal shows TPs

POS Model Results

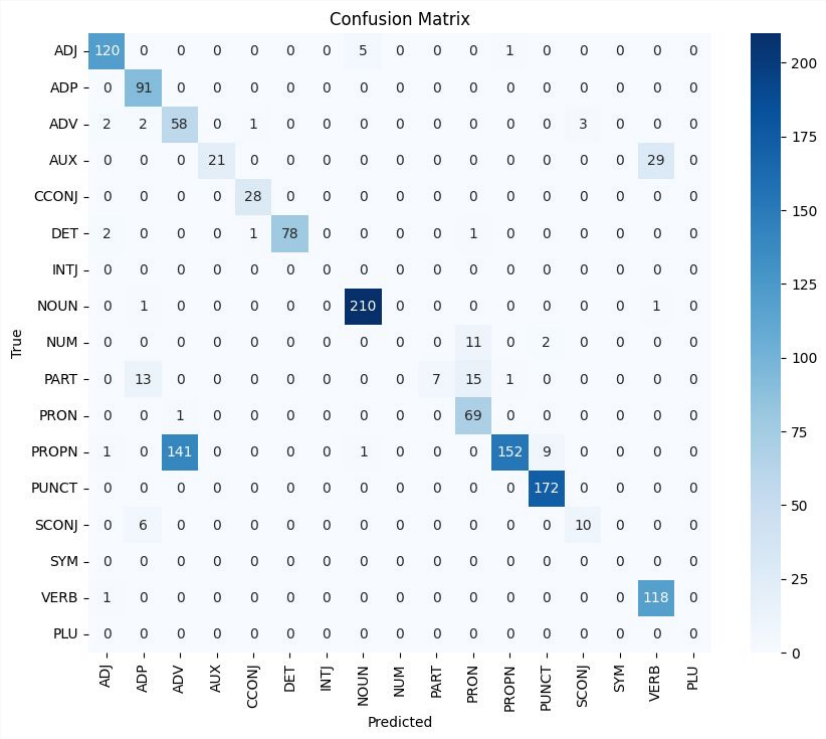
Model	Dataset	Accuracy	Precision	Recall	F1
JP	JP	0.8539	0.85	0.85	0.85
JP	English	0.8188	0.89	0.82	0.82
English	English	0.9906	0.99	0.99	0.99
English	JP	0.4905	0.54	0.49	0.48



POS Confusion Matrix

How does the JP model process english sentences?

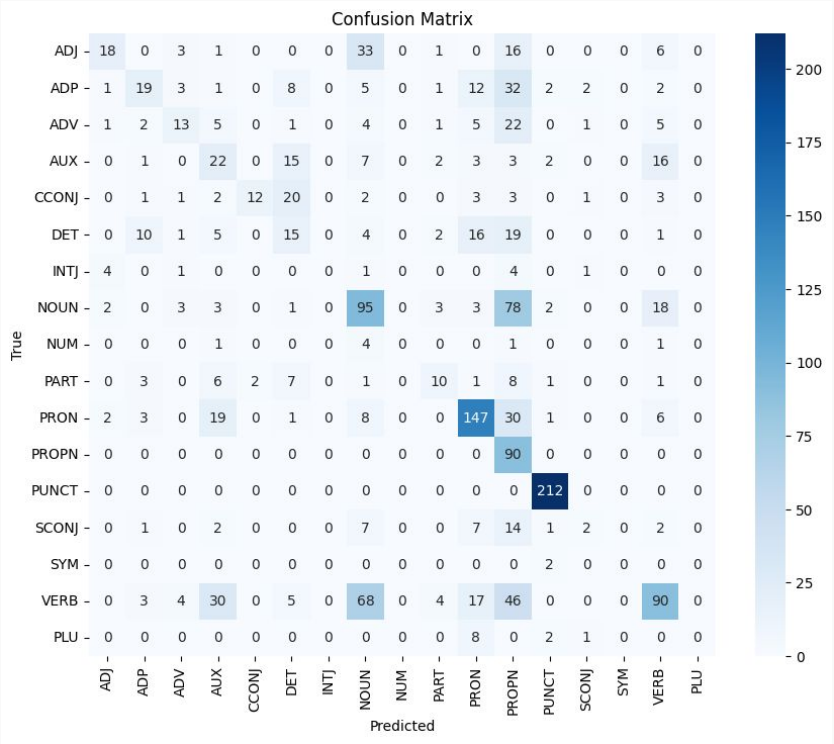
- Learned well:
  - noun, det, cconj, adj
  - F1 from 98% to 95%
- Learned poorly:
  - adverbs, parts
  - F1 from 44% to 33%



POS Confusion Matrix

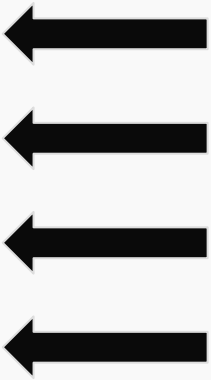
How does the English model process JP sentences?

- Learned well:
  - pron
  - F1 of 67%
- Learned poorly:
  - det, sconj
  - F1 from 21% to 9%



Dependency Model Results

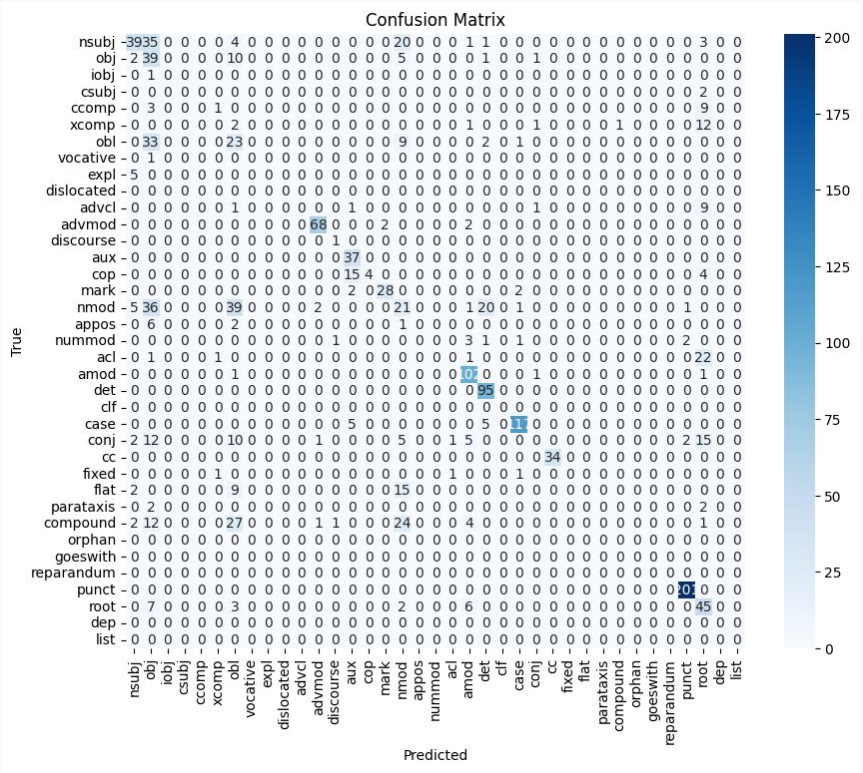
Model	Dataset	Accuracy	Precision	Recall	F1
JP	JP	0.6318	0.64	0.63	0.62
JP	English	0.6117	0.57	0.61	0.57
English	English	0.9212	0.92	0.92	0.92
English	JP	0.4242	0.44	0.42	0.42



Dependency Confusion Matrix

How does the JP model process english sentences?

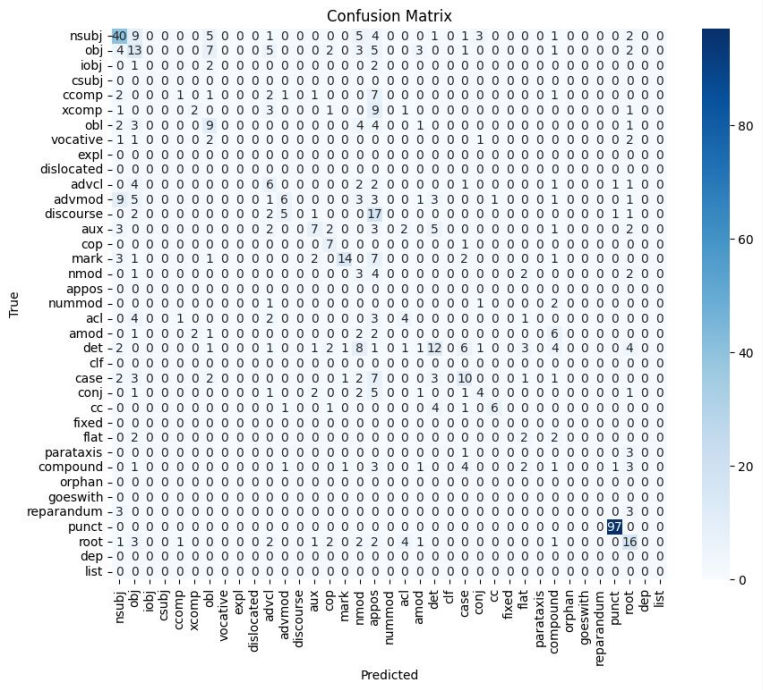
- Learned well:
  - advmod, case, mark, amod, det
  - F1 from 94% to 86%
- Learned poorly:
  - ccomp
  - F1 of 12%



Dependency Confusion Matrix

How does the English model process JP sentences?

- Learned well:
  - nsubj, cc, cop
  - F1 from 65% to 58%
- Learned poorly:
  - xcomp, ccomp, compound
  - F1 from 14% to 12%



05

Discussion

Results Analysis

- While the JP model demonstrated cross-lingual learning capabilities, the English model had poor performance in **both** tasks
- **Linguistic Differences:**
  - Creolization
- **Infrastructure**:
  - Model generalization

Issue 1: Creolization

Creole Languages: Languages resulting from a mixture of cultures, usually under dominant/subordinate conditions (i.e. colonization)

- **Grammaticalization:** the process where words that originally had concrete, referential meanings evolve into grammatical structures
- **Characteristics**: “simpler” language, limited grammatical morphology
- **Issues**: Orthographic mapping mismatch

Examples

- **JP:** “Roun dem siem taim de, Saal stil did a chretn fi kil aaf di biliiva dem.”
- **Translation:** “Around those times, Saul was still threatening to kill the believers”

- **JP:** “Im uopm op im mout an staat fi tiich dem ”
- **Translation:** “He opened up his mouth and started to teach them”

Issue 2: Model Generalization

Model Generalization: Ability of a model to perform well on unseen data

- Out of vocabulary words (OOVs)
- Loaner words
- Poor tokenization
- Small dataset

Examples

Example 4

Token	True Label	Predicted Label
_Af	ADP	PROPN
ta	ADP	PROPN
_ko	DET	PROPN
pl	DET	PROPN
_die	NOUN	PROPN
_Ji	PROPN	PROPN
iza	PROPN	PROPN
s	PROPN	PROPN
_go	VERB	VERB
_baka	ADP	PROPN
_Kya	PROPN	PROPN
porn	PROPN	PROPN
iyo	PROPN	PROPN
m	PROPN	PROPN
_an	CCONJ	DET
_ny	NOUN	PROPN
u	NOUN	PROPN
uz	NOUN	PROPN
_pred	VERB	ADP
_a	DET	NOUN
al	DET	NOUN
_bout	ADP	NOUN
_se	SCONJ	SCONJ
_im	PRON	PRON
_kom	VERB	PROPN
_u	NOUN	PROPN
om	NOUN	PROPN
_	PUNCT	PUNCT
.	PUNCT	PUNCT

Example 4

Token	True Label	Predicted Label
_So	cc	advmod
_no	advmod	det
_bada	root	compound
_joj	xcomp	compound
_no	obj	root
badi	obj	root
_til	mark	case
_di	det	nmod
_tai	nsubj	obl
m	nsubj	nmod
_	advcl	obl
rait	advcl	obl
_	punct	punct
.	punct	punct

Conclusion

- In cross-lingual transfer, models trained on **low-resource languages** have superior performance to **high-resource** language models
- This suggests that the **lexical base** of low-resource languages affords these models more **generalization capabilities**
- Despite advances in NLP models, they are **unable to detect** parallels between **closely related** languages.
- This is evident in **foundational NLP tasks** (POS and dependency tagging) that rely on models' ability to understand linguistic structures

Future Work

1. Curating more diverse, representative, and annotated datasets
2. Exploring other high-resource, low-resource language pairs
3. Experimenting with other NLP tasks

Thank you!

Questions?