/ANITA B.ORG 20th

GRACE HOPPER CELEBRATION

VIRTUAL

/ ANITA B.ORG 20th

GRACE HOPPER CELEBRATION

VIRTUAL

How Multilingual is Your NLP Model?

Shreya Khurana

whoami

- Sr. Data Scientist at GoDaddy
- Work with language data and ML models

Follow this presentation:
https://github.com/ShreyaKhurana/ghc2020

Outline

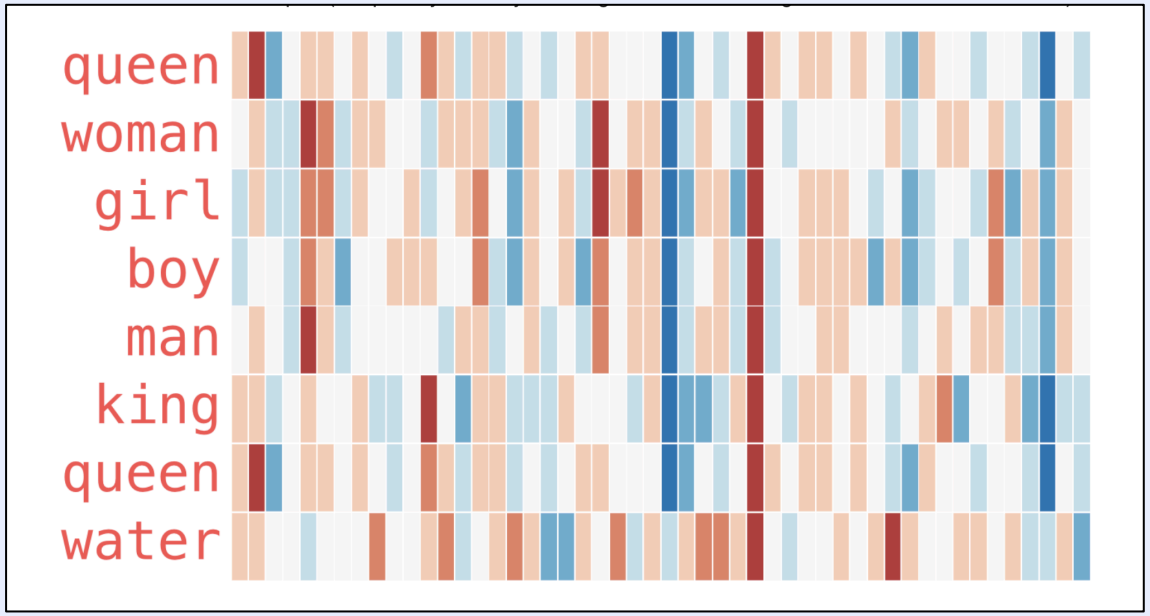
01 Multilingual Data

02 Language Identification

03 BERT

Off-the-shelf Models

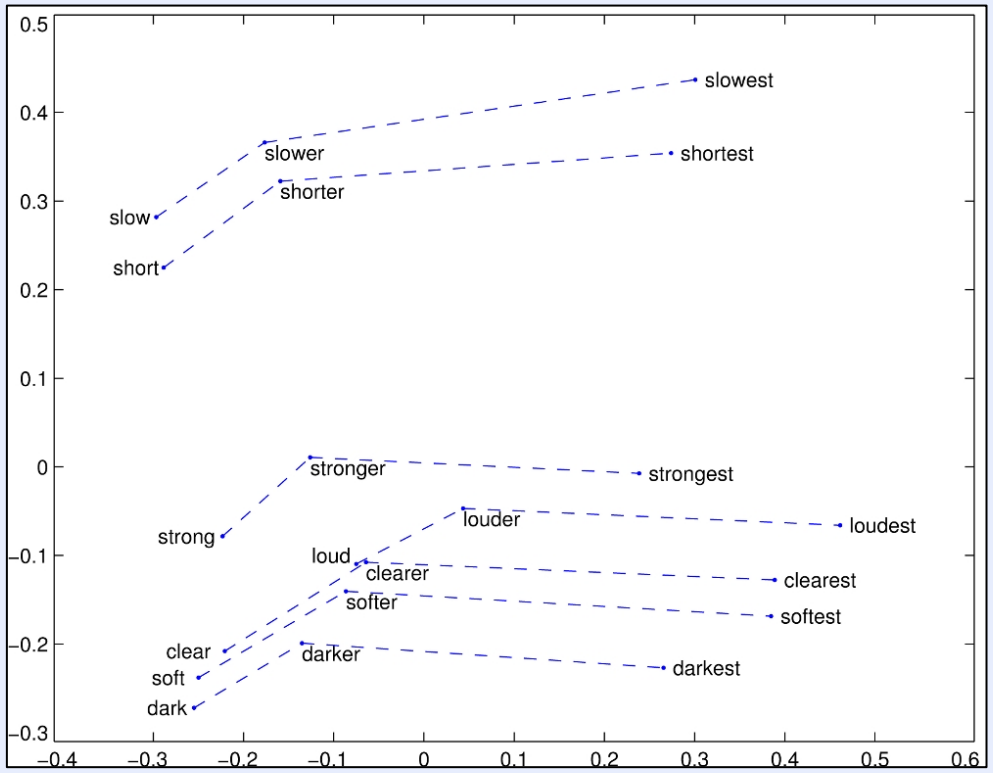
Word2Vec



Source:

- Pre-trained embeddings available for English words
- Works on monolingual corpus
- Need large corpus to train for new languages

GloVe



Source:

Multilingual Data

Code-Switched Data

Alternating between two languages

November 9, 2017 ·

Chalo dollars bhejo is khushi me... let me celebrate...

मेरे घर आयी मेरी नन्ही परी... still remember those happy moments ... happy birthday Shreya .. love you Mera bachcha ❤️

Richa Gandhi
@beingtweet

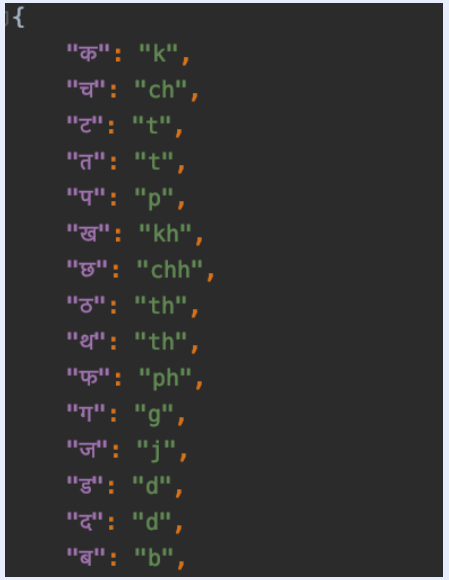
When your Mom says "Aadha apne bhai/ behen ko dena"

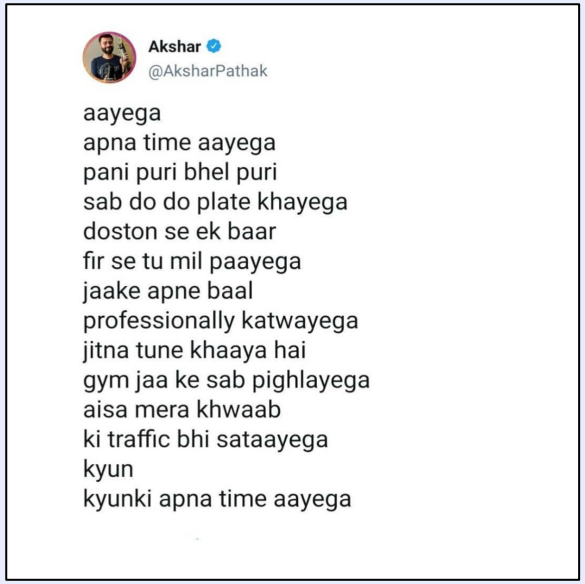
#GrowingUpWithSiblings

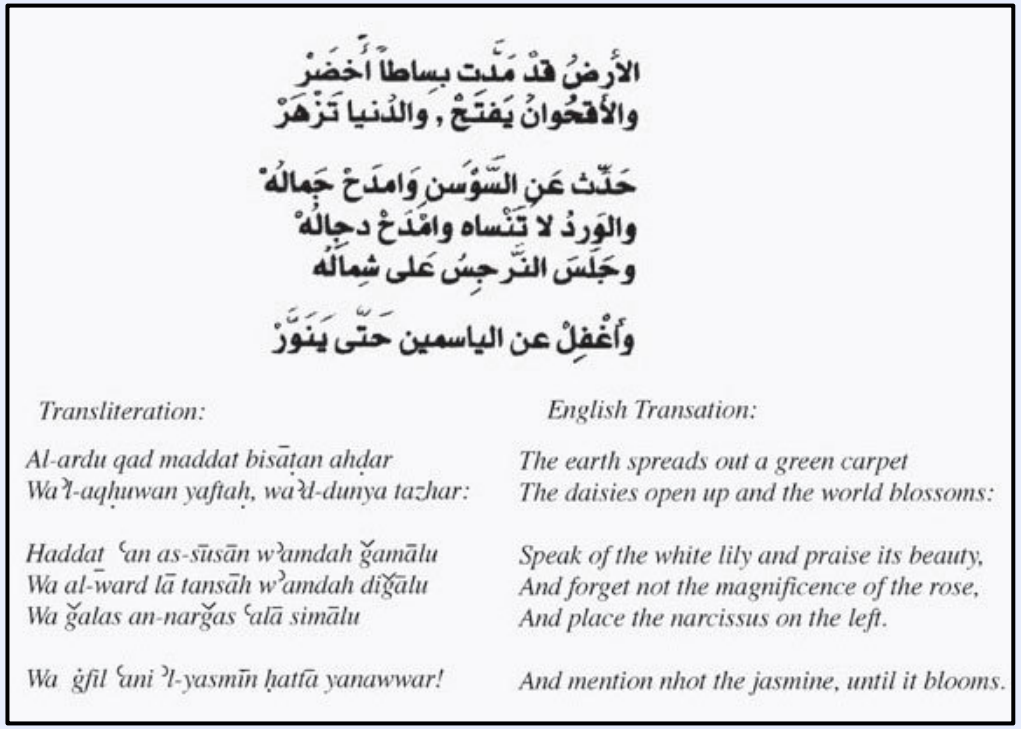
Source: My Facebook timeline

Transliterated Data

Converting words written in alphabet of one language to another







Source : Twitter (topleft), Google images (bottom)

Language Identification: cld3

```
"eo", "su", "uz",
"zh-Latn", "ne",
"nl", "sw", "sq",
"hmn", "ja", "no",
"mn", "so", "ko",
"th", "kk", "sl",
"ig", "mr", "zu",
"ml", "pt", "yi",
"lv", "iw", "cs",
"vi", "jv", "be",
"km", "mk", "tr",
"la", "id", "fil",
"sm", "ca", "el",
"ka", "sr", "it",
"sk", "ru", "ru-
Latn", "et", "ms",
"gd", "bg-Latn",
"ha", "is", "ur",
"mi", "hi", "bn",
"hi-Latn", "fr",
"hu", "lb", "el-
Latn", "st", "ceb",
"pl", "ja-Latn"
```

Transliterated examples

Russian

```
cld3.get_frequent_languages("Privet, kak tebya zovut?", num_langs=2)

[LanguagePrediction(language='ru-Latn', probability=0.8417865633964539, is_reliable=True, proportion=1.0),
 LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]

cld3.get_frequent_languages("Я иду на рынок сегодня", num_langs=2)

[LanguagePrediction(language='ru', probability=0.995514452457428, is_reliable=True, proportion=1.0),
 LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]
```

Hindi

```
# This is a text piece in Hindi transliterated to Roman characters, should support it, but doesn't do a good job
cld3.get_frequent_languages("Main Madhuri Dixit banna chahti hoon", num_langs=2)
# Predicts Finnish

[LanguagePrediction(language='fi', probability=0.47270092368125916, is_reliable=False, proportion=1.0),
 LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]

# This is transliterated Hindi for "How are you? I'm good"
cld3.get_frequent_languages("Kya haal hai? Main achhi hoon", num_langs=2)
# Gaelic is the prediction

[LanguagePrediction(language='gd', probability=0.4288159906864166, is_reliable=False, proportion=1.0),
 LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]
```

Language Identification: cld3

```
"eo", "su", "uz",
"zh-Latn", "ne",
"nl", "sw", "sq",
"hmn", "ja", "no",
"mn", "so", "ko",
"th", "kk", "sl",
"ig", "mr", "zu",
"ml", "pt", "yi",
"lv", "iw", "cs",
"vi", "jv", "be",
"km", "mk", "tr",
"la", "id", "fil",
"sm", "ca", "el",
"ka", "sr", "it",
"sk", "ru", "ru-
Latn", "et", "ms",
"gd", "bg-Latn",
"ha", "is", "ur",
"mi", "hi", "bn",
"hi-Latn", "fr",
"hu", "lb", "el-
Latn", "st", "ceb",
"pl", "ja-Latn"
```

Code-switched examples

Spanish/French

```
cld3.get_frequent_languages("Pero like", num_langs=2)
# Predicts Maori - a language used by an indigenous group in NZ

[LanguagePrediction(language='mi', probability=0.8353496193885803, is_reliable=True, proportion=1.0),
 LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]

cld3.get_frequent_languages("Cojelo con take it easy", num_langs=2)

[LanguagePrediction(language='en', probability=0.41589194536209106, is_reliable=False, proportion=1.0),
 LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]

cld3.get_frequent_languages("Ce week-end va être super cool.", num_langs=2)
# Predicts catalan- western Romance language derived from Latin
Out[20]:
[LanguagePrediction(language='ca', probability=0.5457612872123718, is_reliable=False, proportion=1.0),
 LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]
```

Language Identification: Langid

Transliterated example

Hindi: Original script/transliterated

```
# Does well on original script
identifier.classify("अगर आप हिंदी में जानकारी पढ़ना पसंद करते हैं ")
('hi', 0.9999999749526877)

 # Let's try it on a Hindi transliterated text
identifier.classify("Main Madhuri Dixit banna chahti hoon")
# Predicts Irish
('ga', 0.5746522148102053)

identifier.set_languages(['hi','en'])

identifier.classify("Main Madhuri Dixit banna chahti hoon")
('en', 0.999999999991774)
```

Russian transliterated

```
identifier.set_languages(['ru', 'en', 'it', 'sk'])
identifier.classify("Privet, kak tebya zovut?")
('sk', 0.66767735005164)
```

Underrepresented language

English vs Swahili

```
identifier.classify("Hello! How are you this fine day?")
('en', 0.9999998136419443)

# Expected: Swahili, Predicted: Malay
identifier.classify("Naenda kwa alama leo")
('ms', 0.5526980270071106)
```

97 languages

af, am, an, ar, as, az, cs, cy, da, de, dz, el, en, eo, es, et, hi, hr, ht, hu, hy, id, is, it, ja, ml, mn, mr, ms, mt, nb, ne, nl, nn, no, oc, or, pa, pl, ps, pt, qu, ro, ru, rw, se, si, sk, sl, sq, sr, sv, sw, ta, zh, zu,...

Not trained for transliterated languages!

Language Identification: langdetect

af, ar, bg, bn, ca, cs, cy, da, de, el, en, es, et, fa, fi, fr, gu, he, hi, hr, hu, id, it, ja, kn, ko, lt, lv, mk, ml, mr, ne, nl, no, pa, pl, pt, ro, ru, sk, sl, so, sq, sv, sw, ta, te, th, tl, tr, uk, ur, vi, zh-cn, zh-tw

Not trained for transliterated languages!

Russian: Original script/transliterated

```
77  detect_langs("Я иду на рынок сегодня")
78  [ru:0.9999959546567648]
79
80  detect_langs("Privet, kak tebya zovut?")
81  [hr:0.2857142151330604,
82   sk:0.285713262737828O4,
83   sq:0.2857128308459763,
84   hu:0.1428565021344586]
85
86
87
88  detect_langs("Kya haal hai? Main achhi hoon")
89  # Hindi transliterated is detected as Somali
90  [so:0.9999979650989328]
```

Hindi: transliterated

Challenges

Detecting text with Romanized script

Small text length

Slang/ borrowed words

Different transliteration schemes

Overlapping vocabulary

Limited Data

Building Annotated Datasets

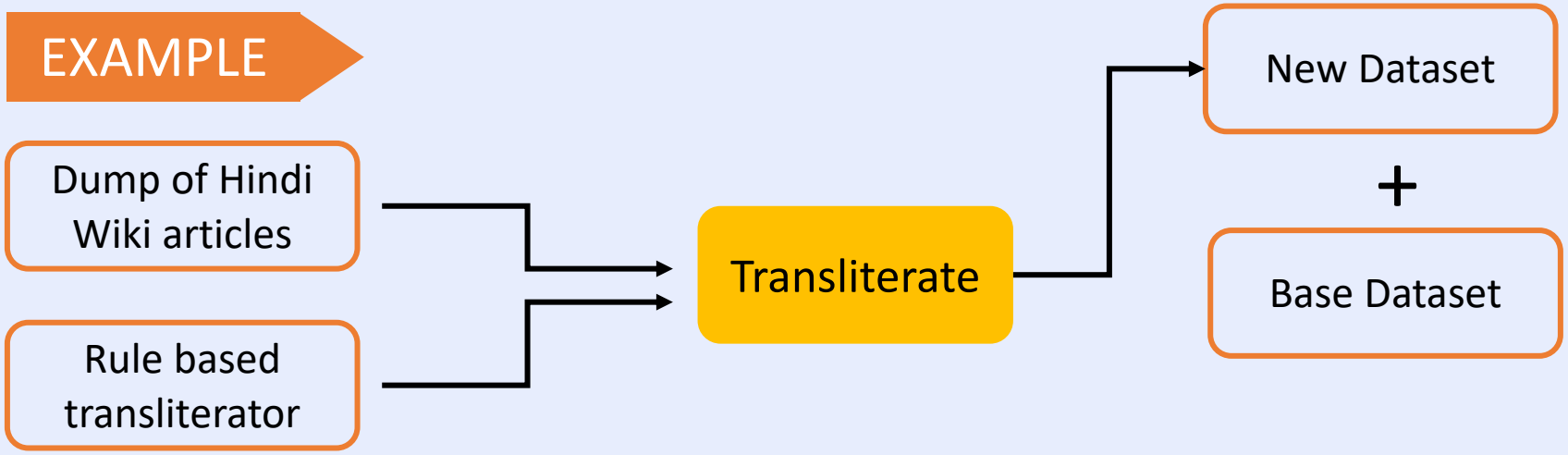
WHY?

- Not enough data available in your desired language
- Do not want models trained on multiple languages to bring in noise

HOW?

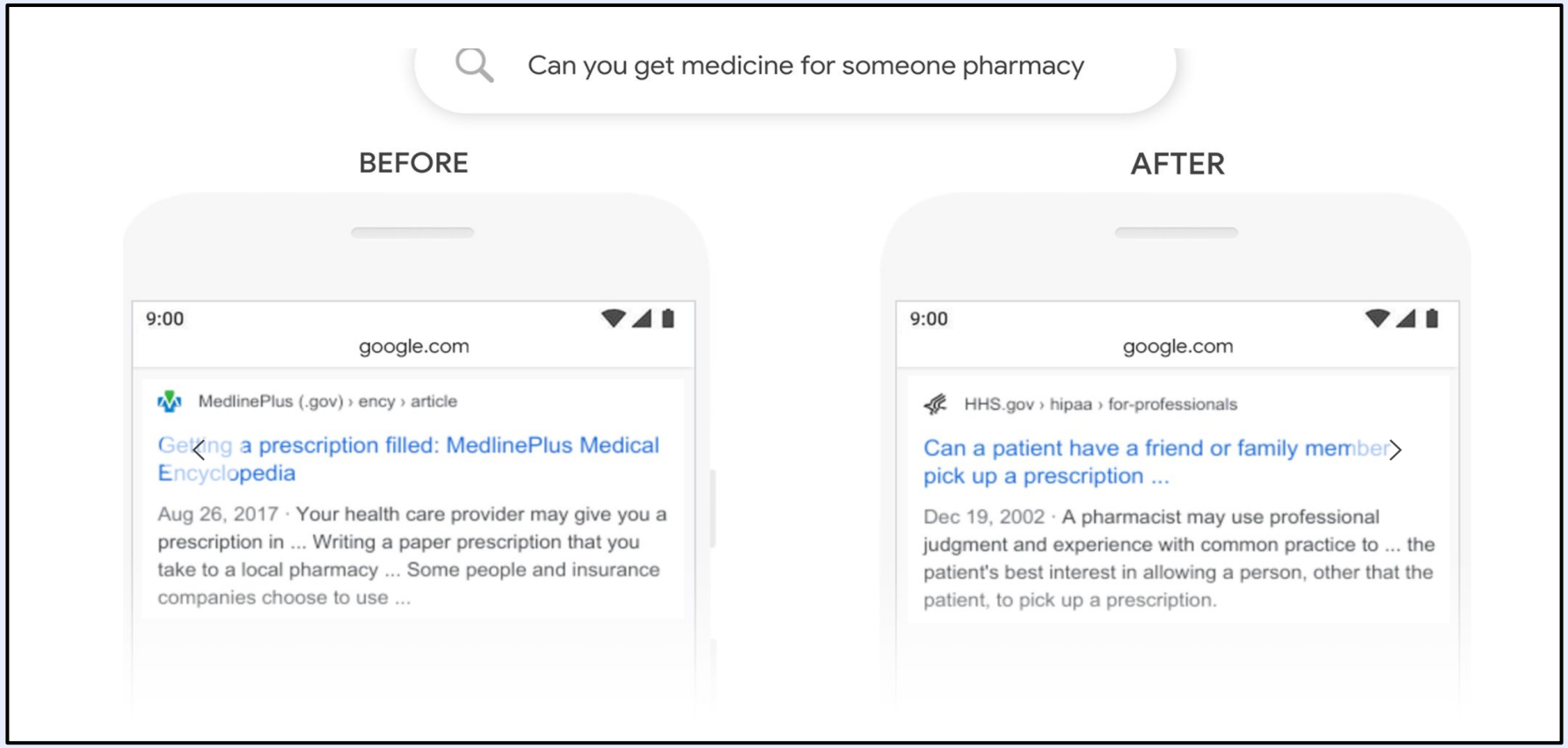
- Identify data source in any language that can be translated to desired language by rules/ machine
- Use this machine generated data to augment

EXAMPLE

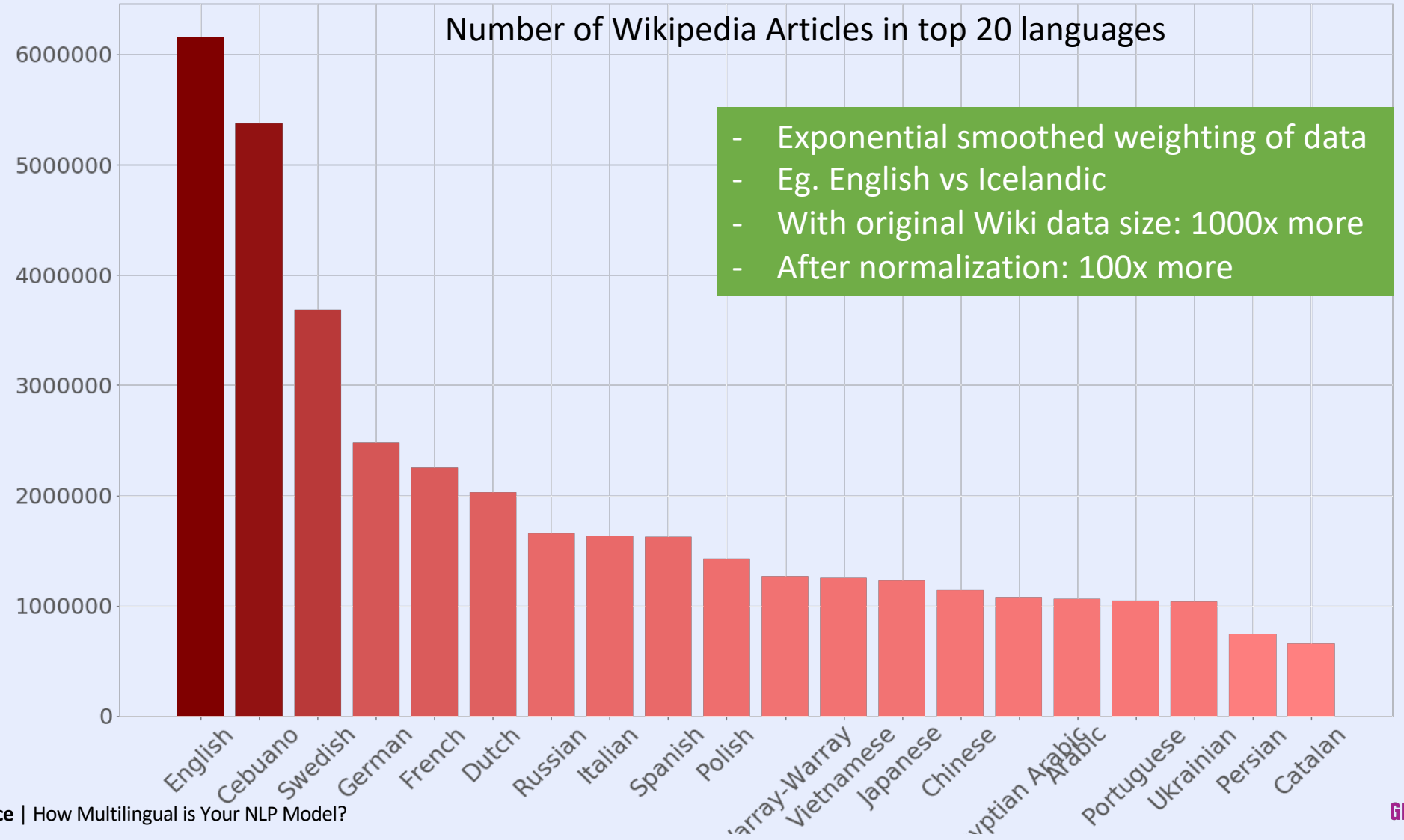


BERT: Bidirectional Encoder Representations from Transformers

- Helps in language understanding!
- Bidirectional
- Looks at all words in the input text



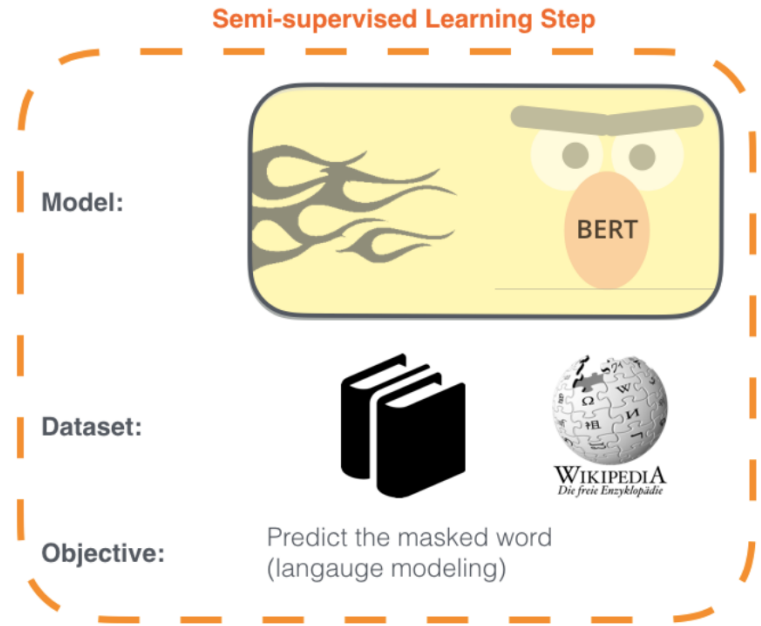
BERT – Languages distribution



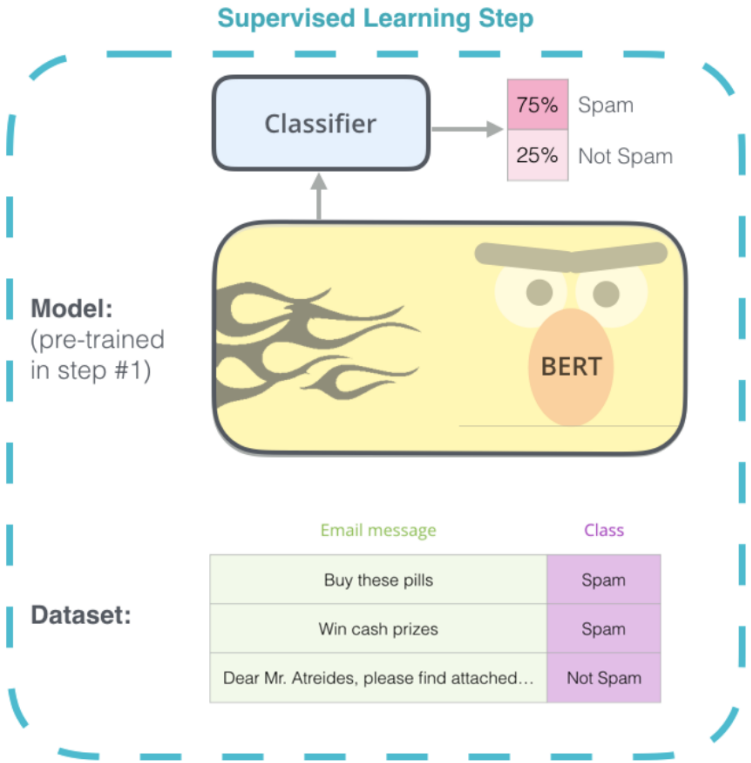
BERT: How

Pre-training: Train on unlabeled data

- Next sentence prediction
  - True label: IsNext 50% of the time
  - False Label: as NotNext
- Masked LM

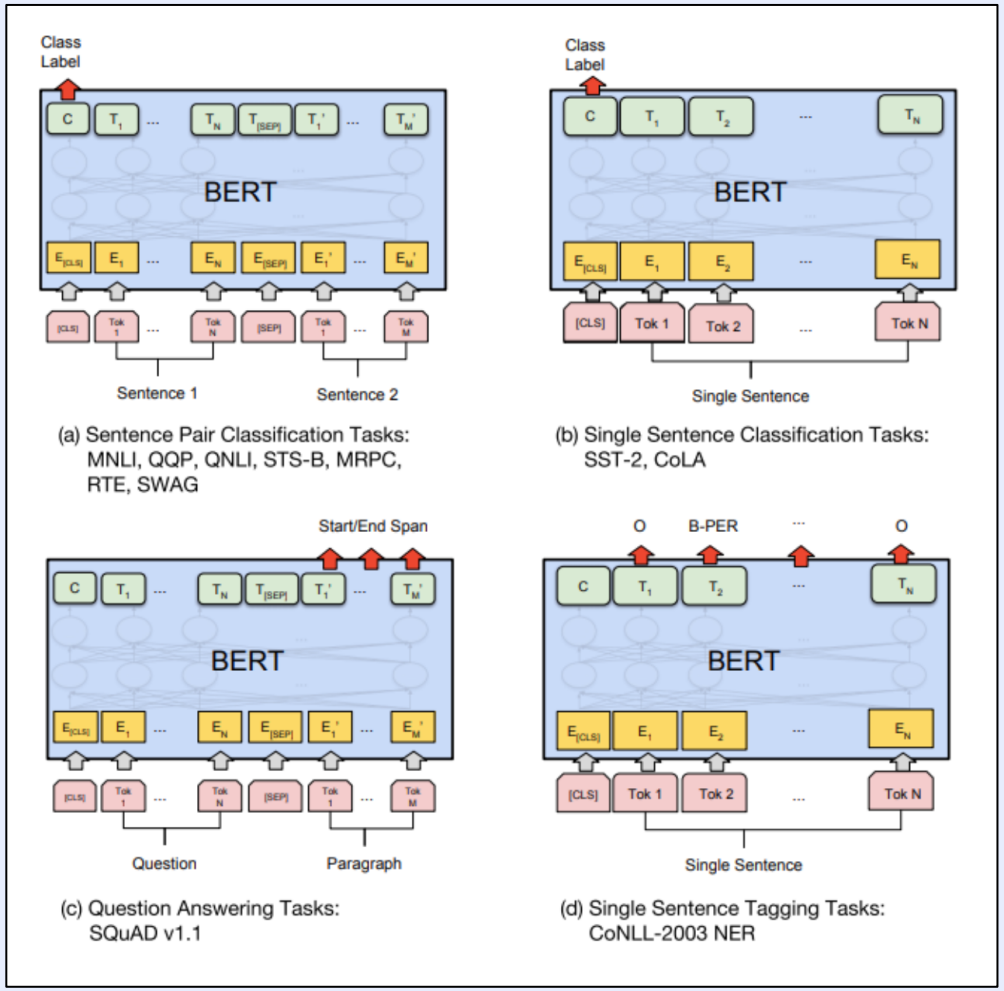


Fine-tuning: Initialize with pre-trained weights and fine-tune for task with labeled data



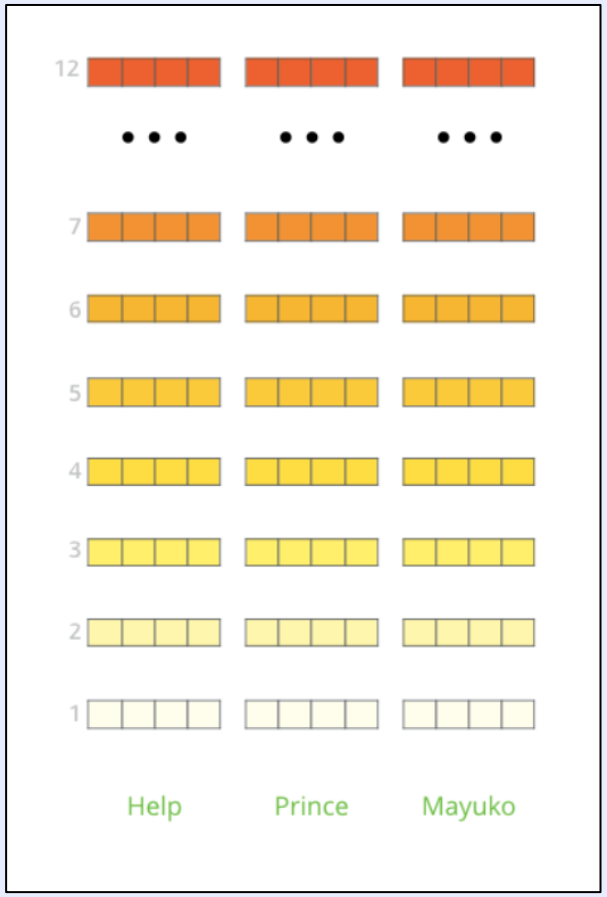
BERT: Tasks

Sentence pairs in paraphrasing
Hypothesis-premise pairs
Question-passage pairs in question answering
A degenerate text-Ø pair in text classification or sequence tagging.



BERT: Representations

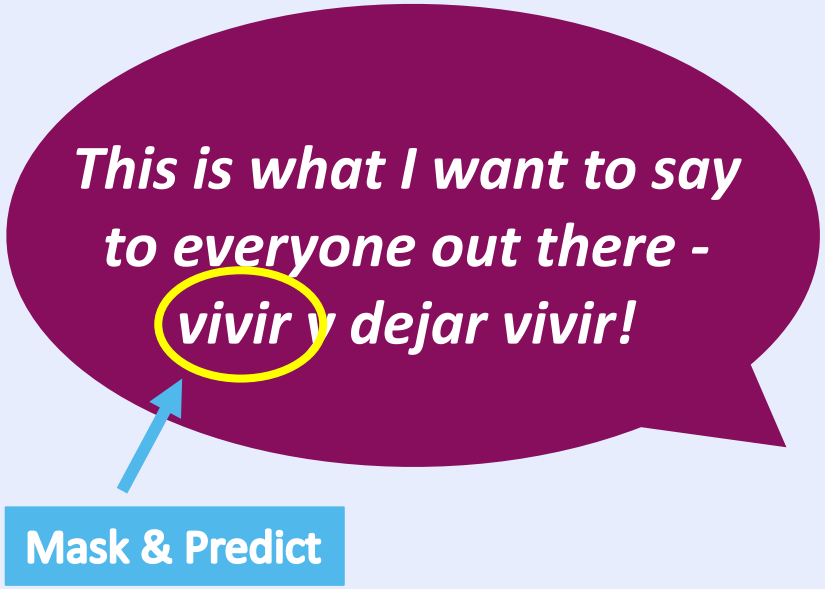
Choose vectors learned from any of these layers



		Dev F1 Score
First Layer	Embedding	91.0
Last Hidden Layer	12	94.9
Sum All 12 Layers	12 + ... + 2 + 1 =	95.5
Second-to-Last Hidden Layer	11	95.6
Sum Last Four Hidden	12 + 11 + 10 + 9 =	95.9
Concat Last Four Hidden	9 10 11 12	96.1

BERT – Language Model

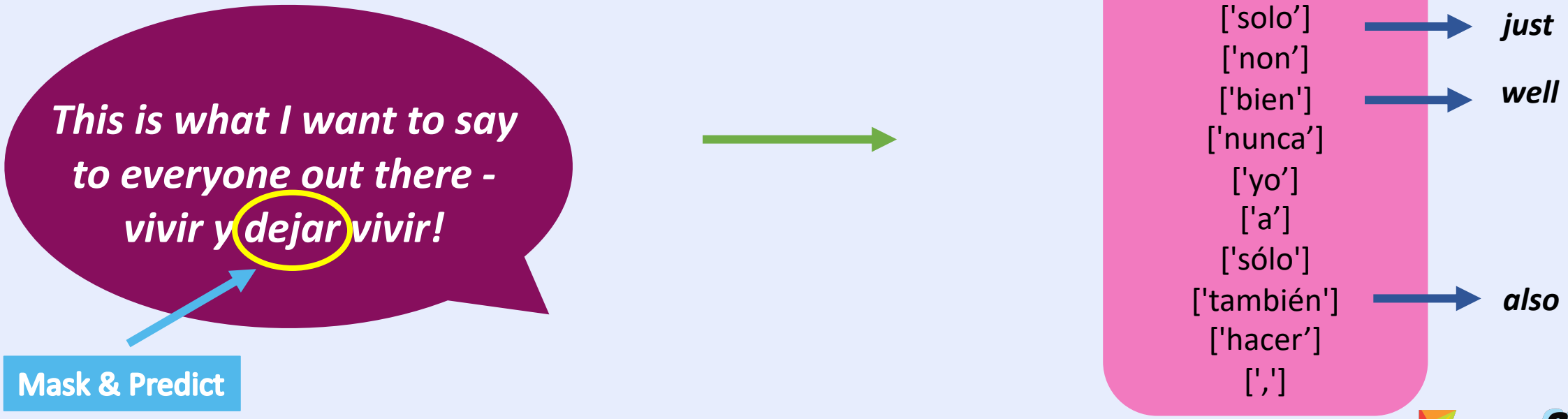
Masked Input text → **Bert-base-multilingual-cased** → Predicted words for masked tokens



['i’]
['...’]
['Viva’] → *live*
['Vive’]
['Hasta'] → *until*
['dejar’]
['Dar']
['empezar'] → *start*
['Gracias’]
['volver’] → *return*
['Por']

BERT – Language Model

Masked Input text → **Bert-base-multilingual-cased** → Predicted words for masked tokens



BERT – Language Model

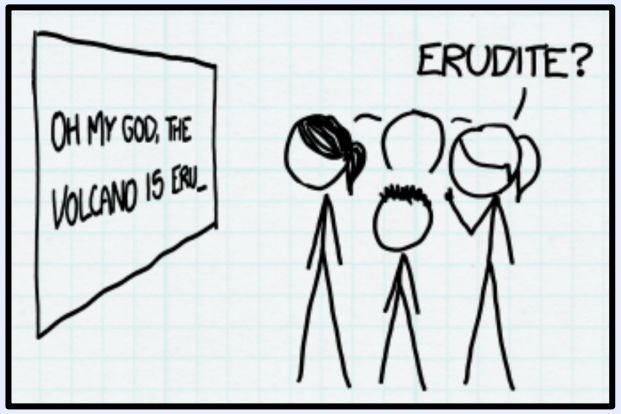
Masked Input text → **Bert-base-multilingual-cased** → Predicted words for masked tokens

Mask & Predict

['blog']
['Blog']
['publisher']
['forum']
['tag']
['poster']
['photographer]
['peer']
['reporter']
['guru']
['pseudonym']

Evaluation

Language modeling: Perplexity



$$PP = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_N)}}$$

Translation: BLEU Score

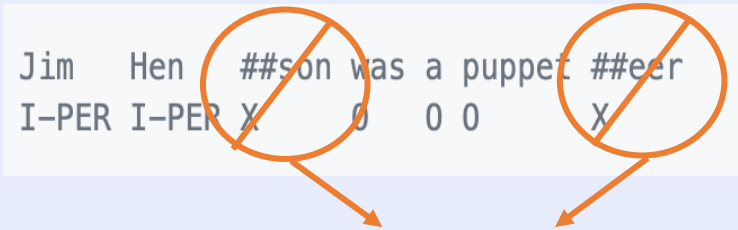
BLEU score

Here, c = candidate length
r = reference length
w_n = *weights of n-gram*
p_n = *modified precision of n-gram*

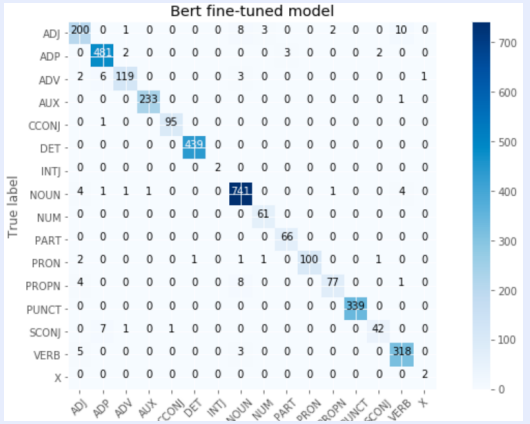
$$BP = \begin{cases} 1, & c > r \\ e^{(1-\frac{r}{c})}, & c \leq r \end{cases}$$

$$BLEU = BP.exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

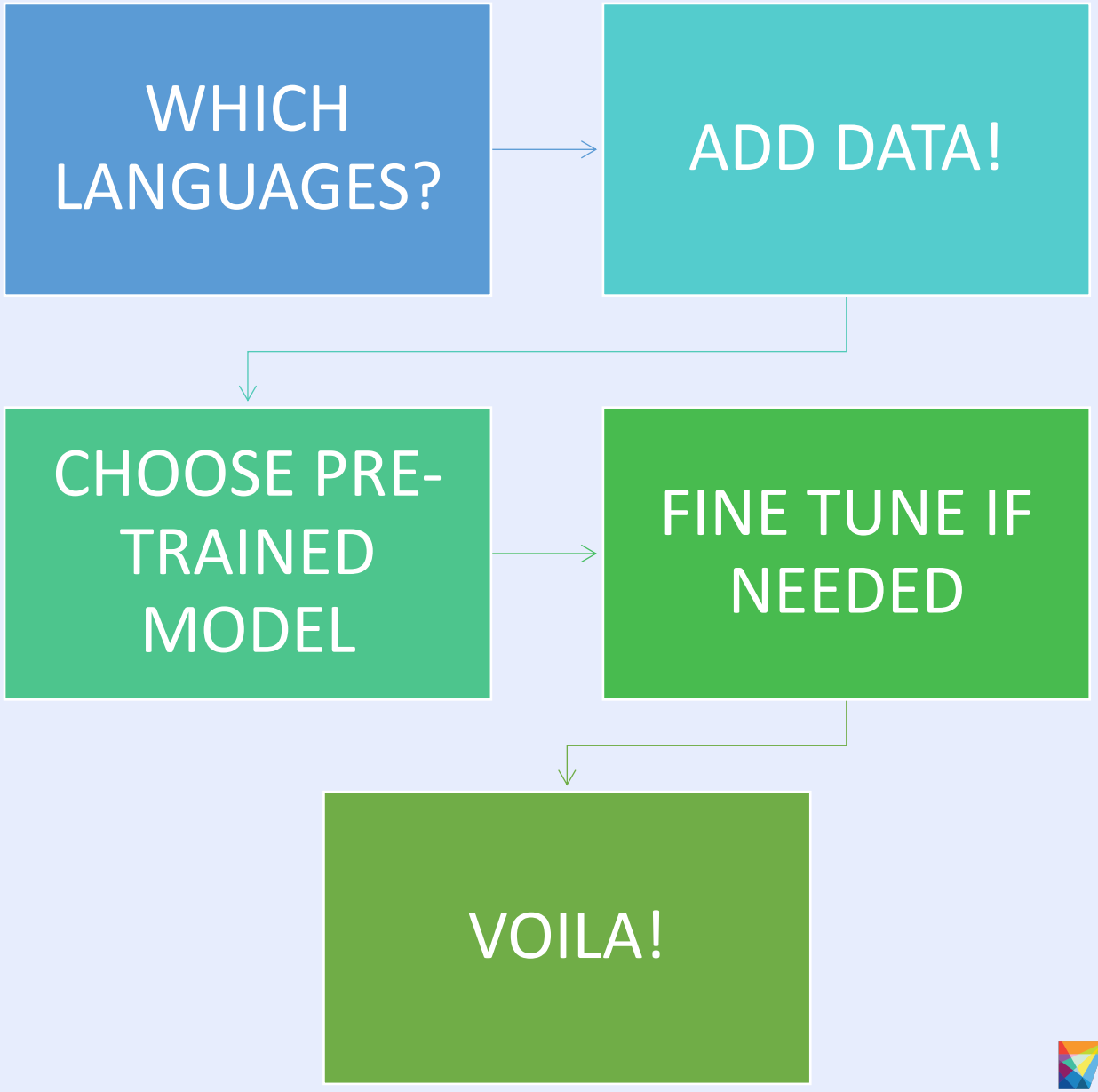
POS Tagging/ Classification



NOT PREDICTED!

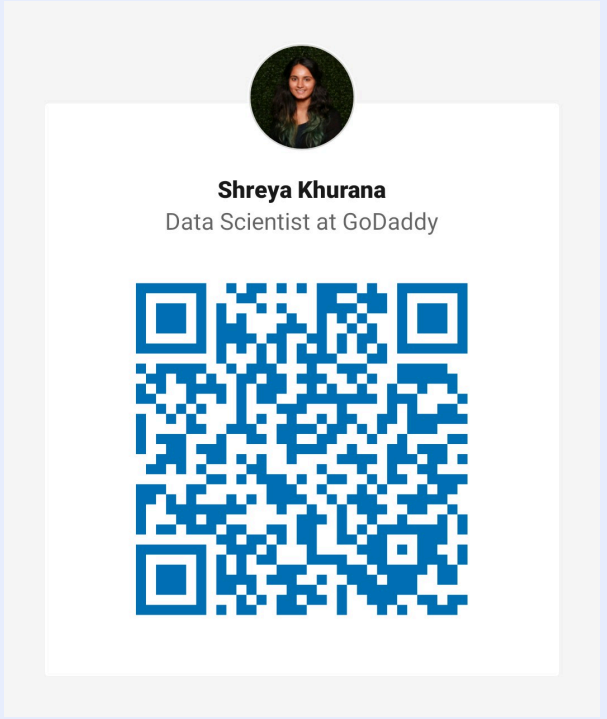


Adding Multilingual Support



Thank you!

Questions?



https://github.com/ShreyaKhurana