



ANITA
B.ORG **20th**
GRACE HOPPER CELEBRATION
V I R T U A L



/ANITA
B.ORG

20th

GRACE HOPPER CELEBRATION

V I R T U A L

Bonjour
mademoiselle
!

Wǒ
hěn
hǎo.

Wie geht
es dir?

Dhanyavaad!

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

Artificial Intelligence | How Multilingual is Your NLP Model?

01

Multilingual Data

02

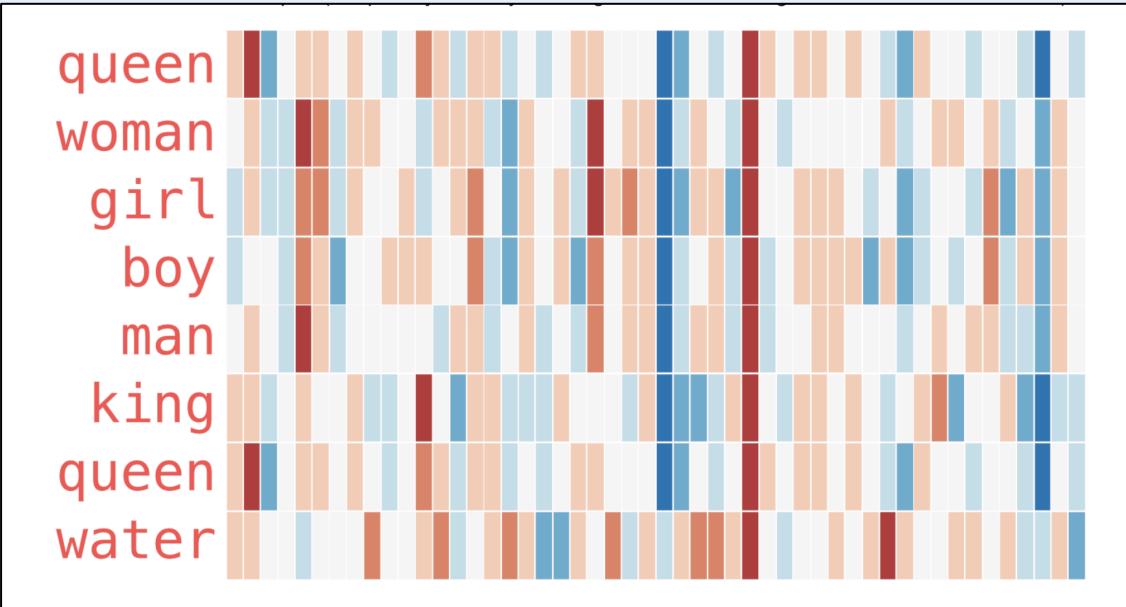
Language Identification

03

BERT

Off-the-shelf Models

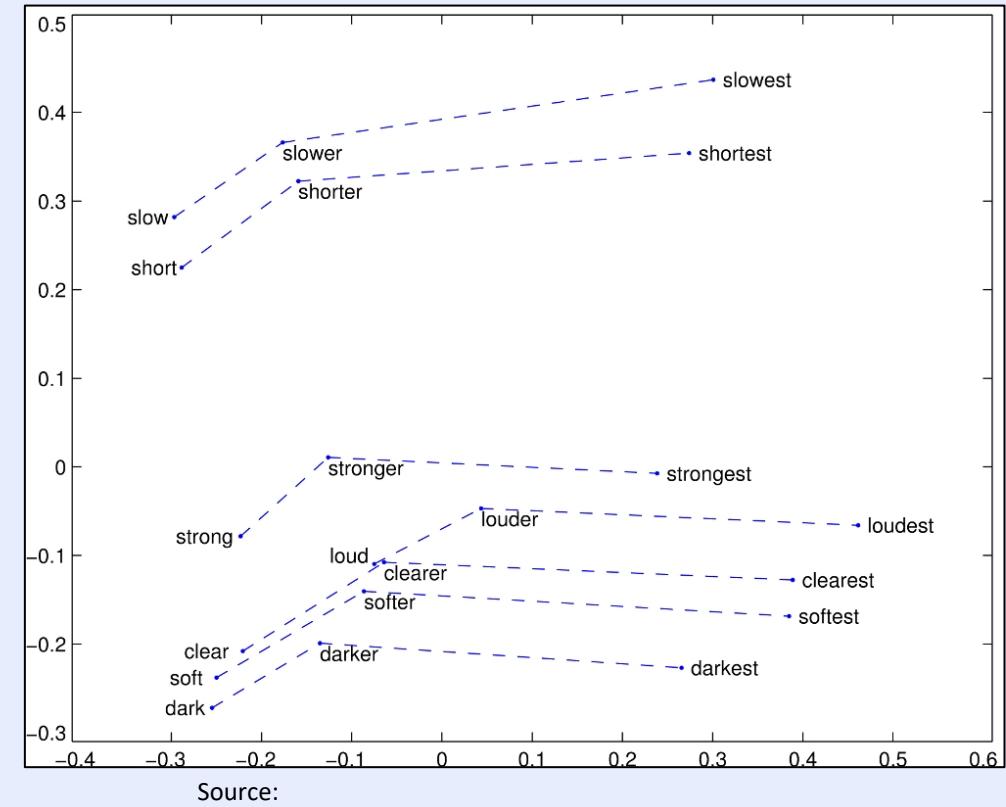
Word2Vec



Source:

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

GloVe



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 नहीं 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

```
{  
    "ਕ": "k",  
    "ਚ": "ch",  
    "ਟ": "t",  
    "ਤ": "t",  
    "ਪ": "p",  
    "ਖ": "kh",  
    "ਛ": "chh",  
    "ਠ": "th",  
    "ਥ": "th",  
    "ਫ": "ph",  
    "ਗ": "g",  
    "ਯ": "j",  
    "ਡ": "d",  
    "ਦ": "d",  
    "ਭ": "b",  
}
```



Akshar
@AksharPathak

aayega
apna time aayega
pani puri bhel puri
sab do do plate khayega
doston se ek baar
fir se tu mil paayega
jaake apne baal
professionally katwayega
jitna tune khaaya hai
gym jaa ke sab pighlayega
aisa mera khwaab
ki traffic bhi sataayega
kyun
kyunki apna time aayega

الْأَرْضُ قَدْ مَدَتْ بِسِاطًا أَخْضَرْ
وَالْأَقْحَوْنَ يَفْتَحُ، وَالْدُّنْيَا تَزَهَرْ

حَدَّثَ عَنِ السَّوْسَنِ وَامْدَحَ حَمَالَةَ
وَالْوَرْدُ لَا تَنْسَاهُ وَامْدَحْ دَحَالَةَ
وَجَلَسَ النَّرْجِسُ عَلَى شِمَالَةَ
وَأَغْفِلْ عَنِ الْيَاسِمِينِ حَتَّى يَنْوَرْ

Transliteration:

Al-ardu qad maddat bisātan ahdar
Wa ɻ-aqhuwan yaftah, wa ɻ-dunya tazhar:

Haddat ɻan as-sūsān w̑amdaḥ ḡamālu
Wa al-w̑ard lā tansāh w̑amdaḥ diḡālu
Wa ḡalas an-narḡas ɻalā simālu

Wa ɻ-ḡfil ɻuni ɻ-yasmin ɻatfā yanawwar!

English Translation:

The earth spreads out a green carpet
The daisies open up and the world blossoms:

Speak of the white lily and praise its beauty,
And forget not the magnificence of the rose,
And place the narcissus on the left.

And mention nhot the jasmine, until it blooms.

Language Identification: cld3

```
"eo", "su", "uz",
"zh-Latin", "ne",
"n\u0101", "sw", "sq",
"hm\u0101", "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-
Latin", "et", "ms",
"gd", "bg-Latin",
"ha", "is", "ur",
"mi", "hi", "bn",
"hi-Latin", "fr",
"hu", "lb", "el-
Latin", "st", "ceb",
"pl", "ja-Latin"
```

Transliterated examples

Russian

```
3 cld3.get_frequent_languages("Privet, kak tebya zovut?", num_langs=2)
4
5 [LanguagePrediction(language='ru-Latin', probability=0.8417865633964539, is_reliable=True, proportion=1.0),
6 LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]
7
8 cld3.get_frequent_languages("Я иду на рынок сегодня", num_langs=2)
9
10 [LanguagePrediction(language='ru', probability=0.995514452457428, is_reliable=True, proportion=1.0),
11 LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]
```

Hindi

```
33 # This is a text piece in Hindi transliterated to Roman characters, should support it, but doesn't do a good job
34 cld3.get_frequent_languages("Main Madhuri Dixit banna chahti hoon", num_langs=2)
35 # Predicts Finnish
36
37 [LanguagePrediction(language='fi', probability=0.47270092368125916, is_reliable=False, proportion=1.0),
38 LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]
39
40 # This is transliterated Hindi for "How are you? I'm good"
41 cld3.get_frequent_languages("Kya haal hai? Main achhi hoon", num_langs=2)
42 # Gaelic is the prediction
43
44 [LanguagePrediction(language='gd', probability=0.4288159906864166, is_reliable=False, proportion=1.0),
45 LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]
```

Language Identification: cld3

```
"eo", "su", "uz",
"zh-Latin", "ne",
"n\u00f1", "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-
Latin", "et", "ms",
"gd", "bg-Latin",
"ha", "is", "ur",
"mi", "hi", "bn",
"hi-Latin", "fr",
"hu", "lb", "el-
Latin", "st", "ceb",
"pl", "ja-Latin"
```

Code-switched examples

Spanish/French

```
15 cld3.get_frequent_languages("Pero like", num_langs=2)
16 # Predicts Maori - a language used by an indigenous group in NZ
17 [
18     LanguagePrediction(language='mi', probability=0.8353496193885803, is_reliable=True, proportion=1.0),
19     LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]
20
21 cld3.get_frequent_languages("Cojelo con take it easy", num_langs=2)
22 [
23     LanguagePrediction(language='en', probability=0.41589194536209106, is_reliable=False, proportion=1.0),
24     LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]
25
26 cld3.get_frequent_languages("Ce week-end va \u00eatre super cool.", num_langs=2)
27 # Predicts catalan- western Romance language derived from Latin
28 Out[20]:
29 [
30     LanguagePrediction(language='ca', probability=0.5457612872123718, is_reliable=False, proportion=1.0),
31     LanguagePrediction(language='und', probability=0.0, is_reliable=False, proportion=0.0)]
```

Language Identification: Langid

Transliterated example

Hindi: Original script/transliterated

```
48 # Does well on original script
49 identifier.classify("अगर आप हिंदी में जानकारी पढ़ा पसंद करते हैं ")
50 ('hi', 0.999999749526877)
51
52 # Let's try it on a Hindi transliterated text
53 identifier.classify("Main Madhuri Dixit banna chahti hoon")
54 # Predicts Irish
55 ('ga', 0.5746522148102053)
56
57 identifier.set_languages(['hi','en'])
58
59 identifier.classify("Main Madhuri Dixit banna chahti hoon")
60 ('en', 0.99999999991774)
61
62
63 identifier.set_languages(['ru', 'en', 'it', 'sk'])
64 identifier.classify("Privet, kak tebya zovut?")
65 ('sk', 0.66767735005164)
66
```

Russian transliterated

Underrepresented language

English vs Swahili

```
68 identifier.classify("Hello! How are you this fine day?")
69 ('en', 0.9999998136419443)
70
71 # Expected: Swahili, Predicted: Malay
72 identifier.classify("Naenda kwa alama leo")
73 ('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!

```
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:0.28571326273782804,  
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]
```

Russian: Original script/transliterated

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

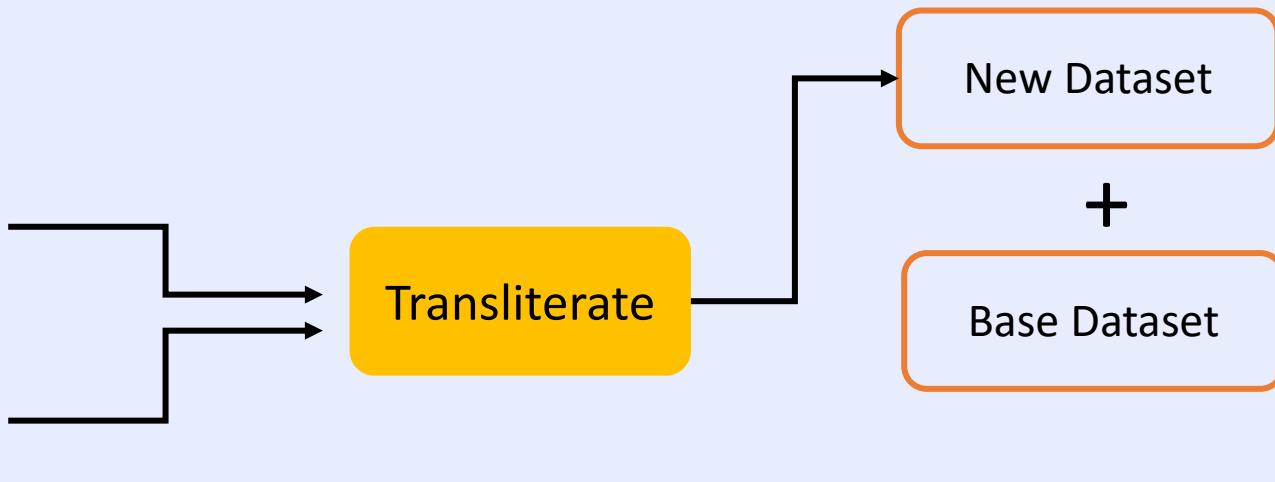
HOW?

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

EXAMPLE

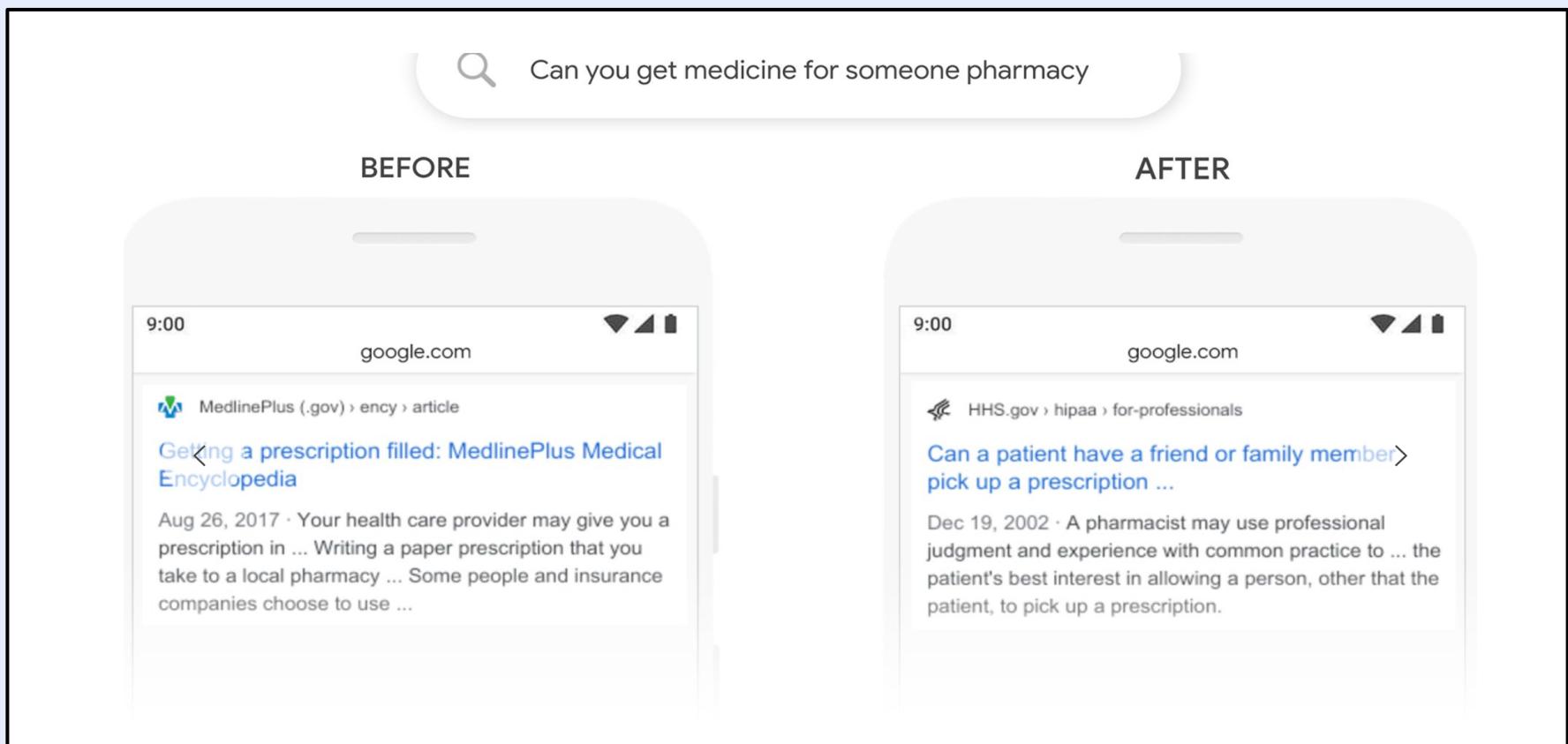
Dump of Hindi
Wiki articles

Rule based
transliterator

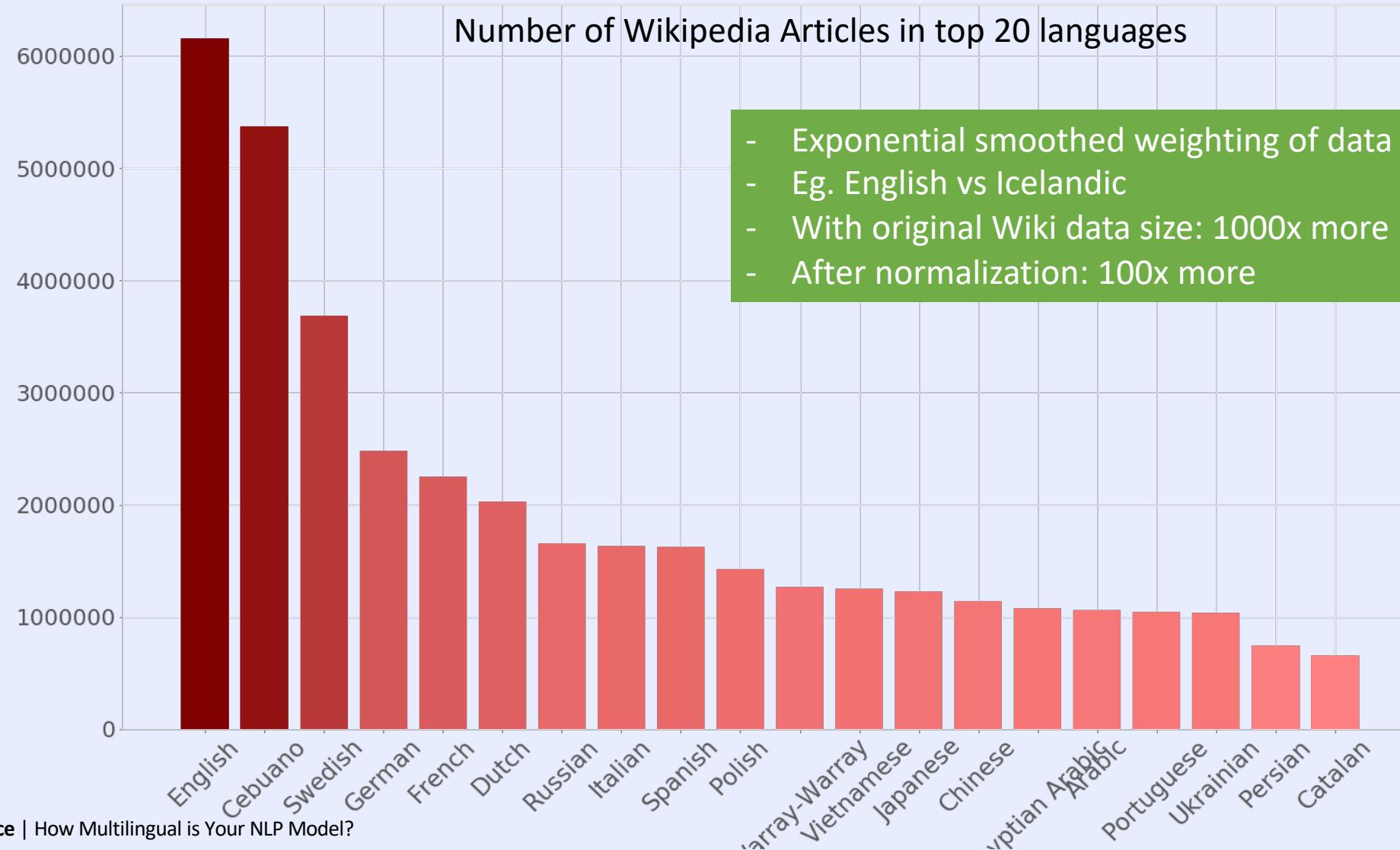


BERT: Bidirectional Encoder Representations from Transformers

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



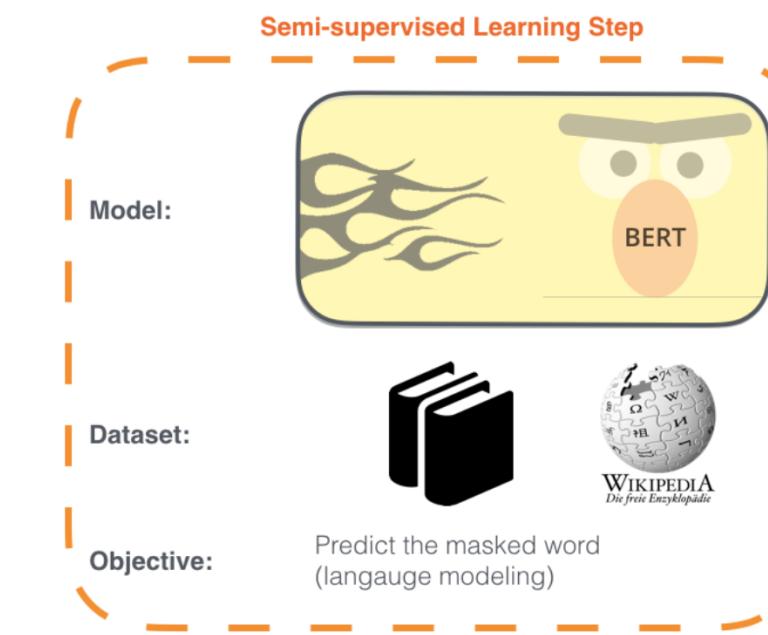
BERT – Languages distribution



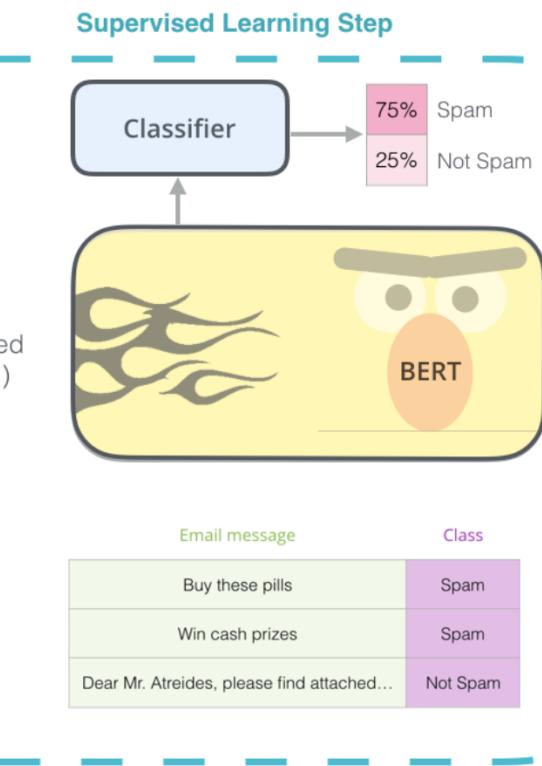
BERT: How

Pre-training: Train on unlabeled data

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

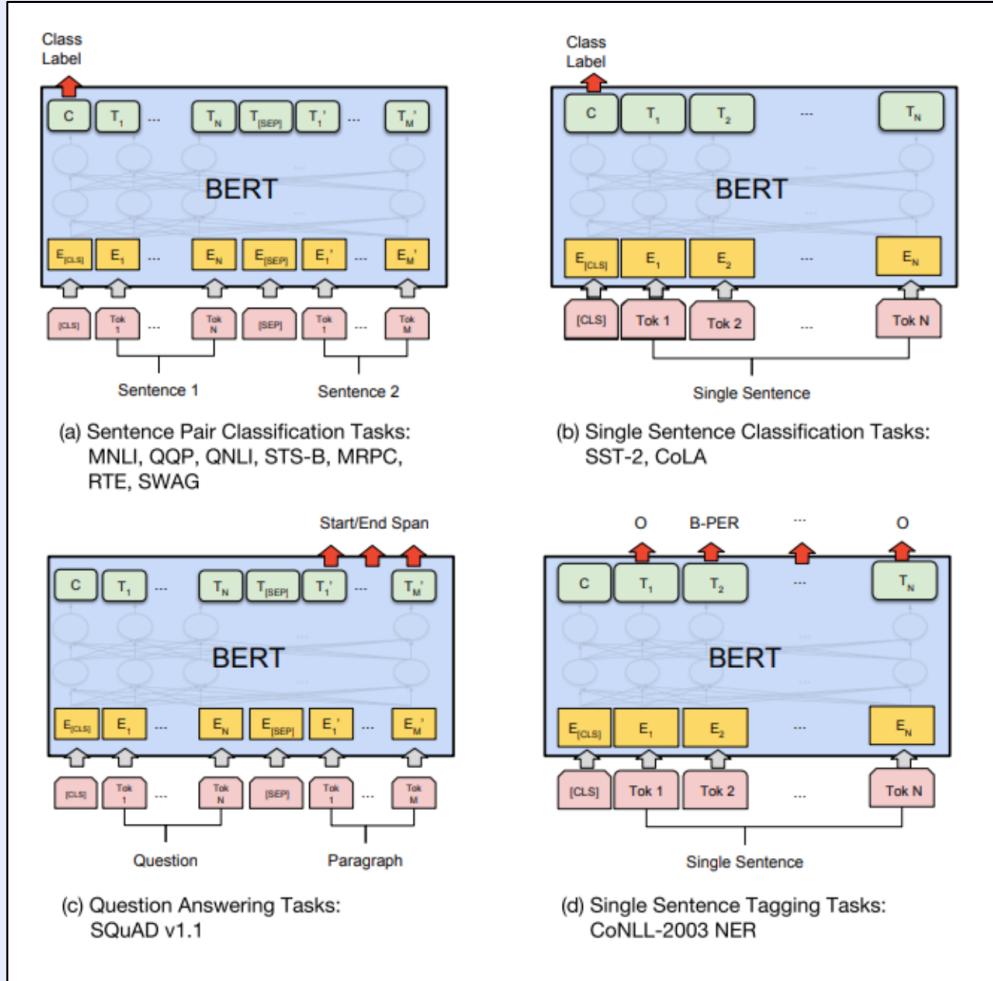


Fine-tuning: Initialize with pre-trained weights and fine-tune for task with labeled data



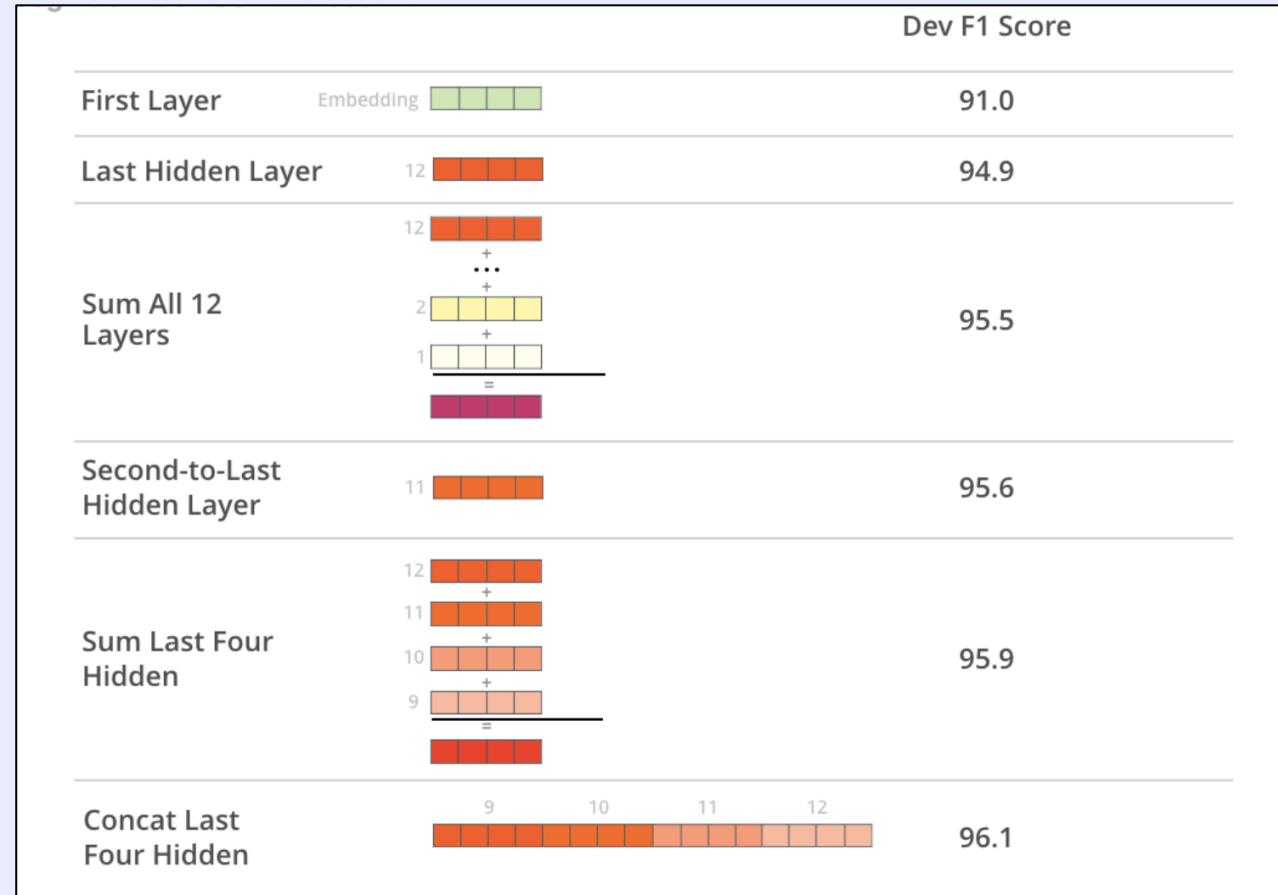
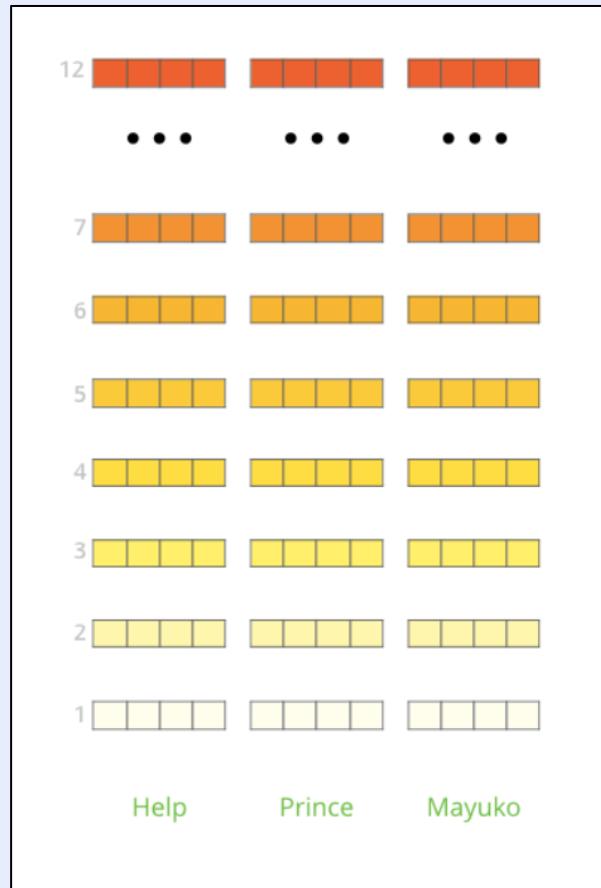
BERT: Tasks

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



BERT – Language Model

Masked
Input text

Bert-base-
multilingual-cased

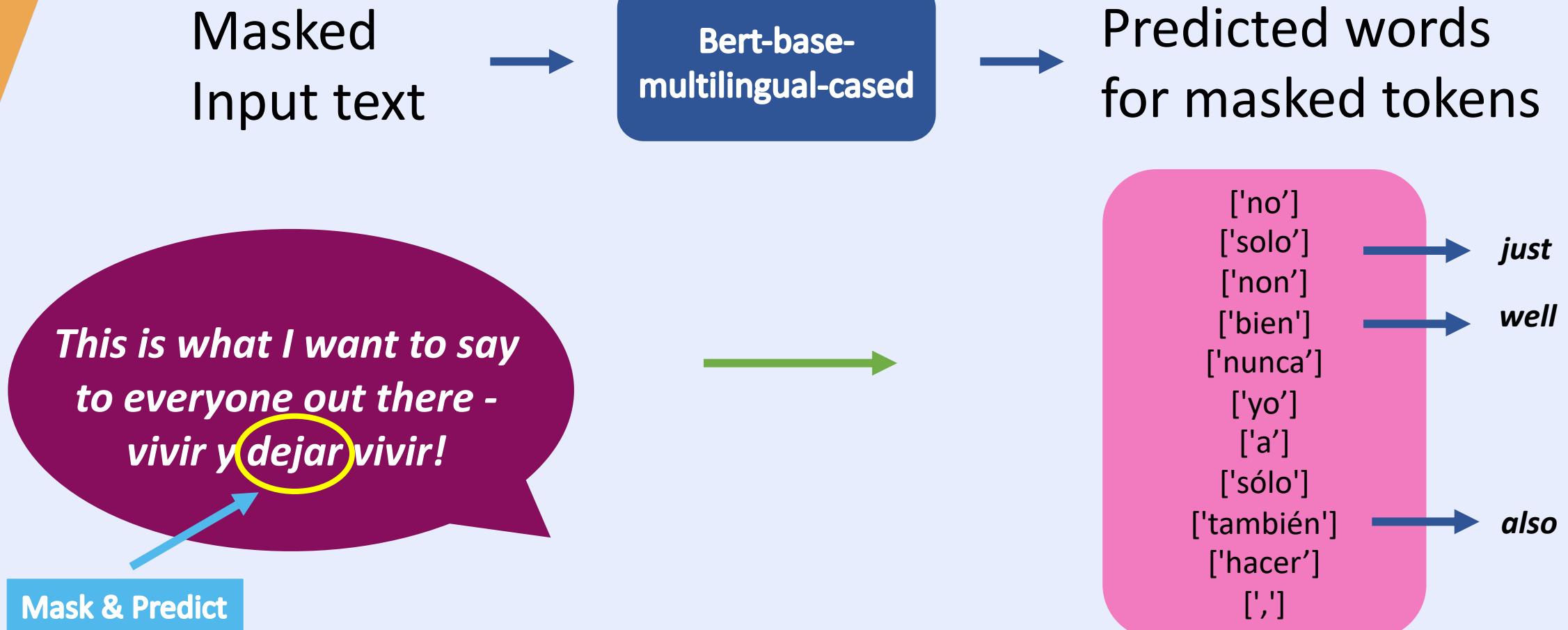
Predicted words
for masked tokens

*This is what I want to say
to everyone out there -
vivir y dejar vivir!*

Mask & Predict

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

BERT – Language Model



BERT – Language Model

Masked
Input text

Bert-base-
multilingual-cased

Predicted words
for masked tokens

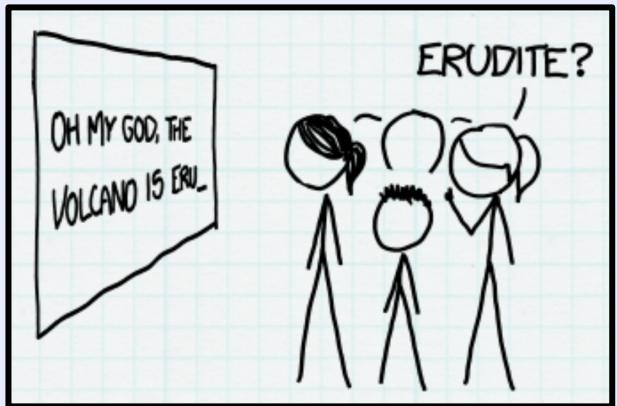
अगर आप हिंदी में जानकारी पढ़ना पसंद करते हैं तो आपको best hindi **blogs** यानी भारत में popular Hindi bloggers कौन-कौन से हैं ... हिंदी में blog पढ़ना पसंद करते हैं लेकिन ... कुछ popular Hindi blogs list होने चाहिए।

Mask & Predict

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

Evaluation

Language modeling: Perplexity



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

BLEU score

Here, c = candidate length

r = reference length

w_n = weights of n -gram

p_n = modified precision of n -gram

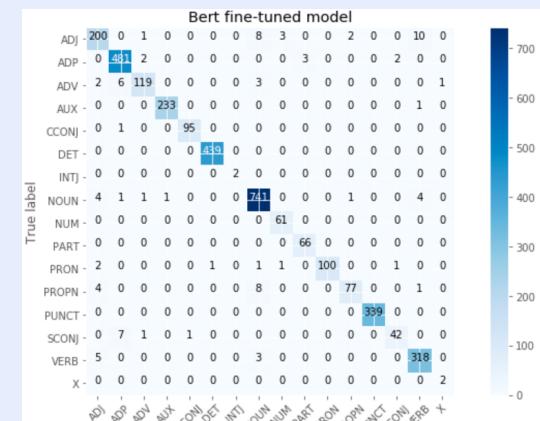
$$BP = \begin{cases} 1, & c > r \\ e^{(1-\frac{r}{c})}, & c \leq r \end{cases}$$

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

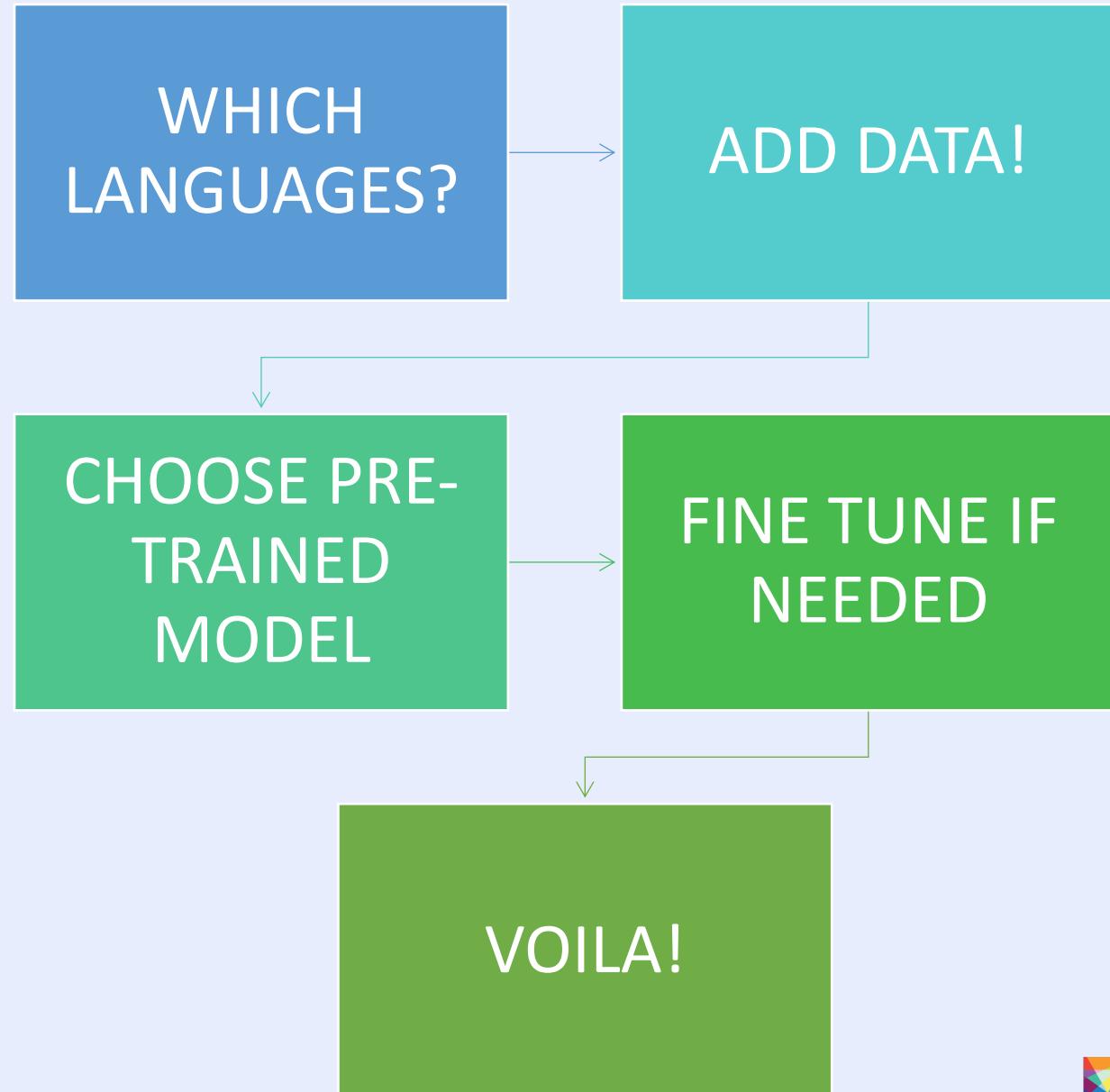
Translation: BLEU Score

POS Tagging/ Classification

Jim Hen ##son was a puppet ##eer
I-PER I-PER X 0 0 0
NOT PREDICTED!



Adding Multilingual Support





Thank you!

Questions?



Shreya Khurana
Data Scientist at GoDaddy

<https://github.com/ShreyaKhurana>