# Multilingual Language Models: NLP Beyond English

Berkeley NLP

Eric Wallace

CS 288

## NLP Beyond English

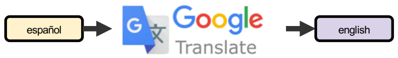
- An overwhelming majority of NLP research focuses on English!

How to build non-English NLP systems?

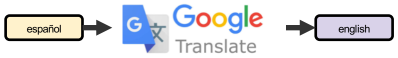
- translate baseline
- monolingual LMs for each language
- multilingual LMs

## Translate Baseline

## Translate Baseline



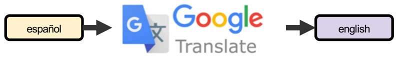
## Translate Baseline



Pros:

- Straightforward to implement
- Strong baseline, especially for classification tasks

## Translate Baseline



Pros:

- Straightforward to implement
- Strong baseline, especially for classification tasks

Cons:

- Suffers from cascading errors
- Limited to languages that translation systems support
- Can be slow and computationally expensive
- Translation is fundamentally lossy?

## Monolingual LMs

## Monolingual LMs

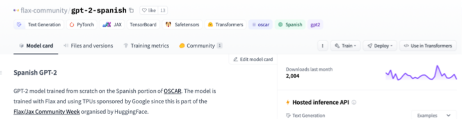
- Can we just repeat the LM pre-training pipeline for other languages?

## Monolingual LMs

- Can we just repeat the LM pre-training pipeline for other languages?
  - Sort of!

## Monolingual LMs

- Can we just repeat the LM pre-training pipeline for other languages?
  - Sort of!


flax-community/gpt-2-spanish

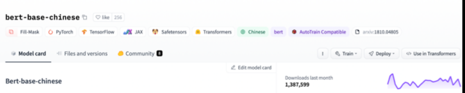
Spanish GPT-2

GPT-2 model trained from scratch on the Spanish portion of OSCAR. The model is trained with Flax and using TPUs sponsored by Google since this is part of the Flax/Jax Community Week organised by HuggingFace.

Downloads last month
2,004

## Monolingual LMs

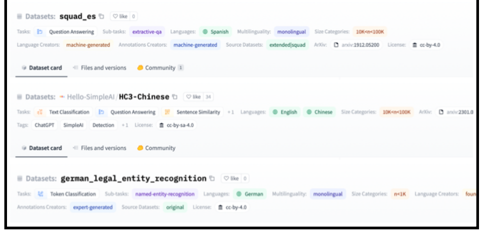
- Can we just repeat the LM pre-training pipeline for other languages?
  - Sort of!


bert-base-chinese

Bert-base-chinese

Downloads last month
1,387,599

## Evaluating LMs in Other Languages


Datasets: squad_es

Tasks: Question Answering Sub-tasks: extractive-qa Languages: Spanish Multilinguality: monolingual Size Categories: 10K<n<100K

Language Creators: machine-generated Annotations Creators: machine-generated Source Datasets: extended|squad ArXiv: arxiv:1912.05200 License: cc-by-4.0

Datasets: Hello-SimpleAI/HC3-Chinese

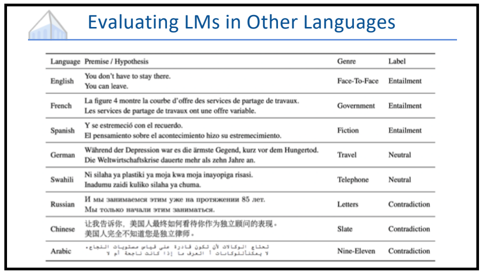
Tasks: Text Classification Question Answering Sentence Similarity +1 Languages: English Chinese Size Categories: 10K<n<100K ArXiv: arxiv:2301.0

Tags: ChatGPT SimpleAI Detection +1 License: cc-by-sa-4.0

Datasets: german_legal_entity_recognition

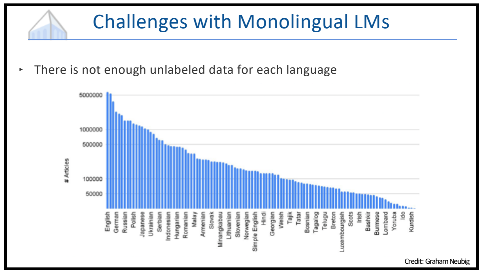
Tasks: Token Classification Sub-tasks: named-entity-recognition Languages: German Multilinguality: monolingual Size Categories: n<1K Language Creators: foun

Annotations Creators: expert-generated Source Datasets: original License: cc-by-4.0



## Evaluating LMs in Other Languages

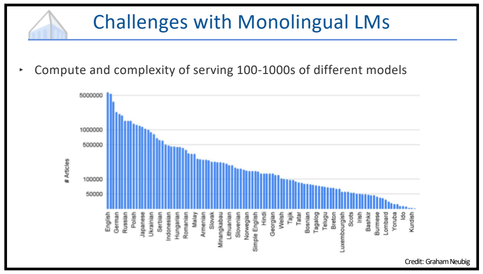
| Language | Premise / Hypothesis | Genre | Label |
|---|---|---|---|
| English | You don't have to stay there.<br>You can leave. | Face-To-Face | Entailment |
| French | La figure 4 montre la courbe d'offre des services de partage de travaux.<br>Les services de partage de travaux ont une offre variable. | Government | Entailment |
| Spanish | Y se estremeció con el recuerdo.<br>El pensamiento sobre el acontecimiento hizo su estremecimiento. | Fiction | Entailment |
| German | Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod.<br>Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an. | Travel | Neutral |
| Swahili | Ni silaha ya plastiki ya moja kwa moja inayopiga risasi.<br>Inadumu zaidi kuliko silaha ya chuma. | Telephone | Neutral |
| Russian | И мы занимаемся этим уже на протяжении 85 лет.<br>Мы только начали этим заниматься. | Letters | Contradiction |
| Chinese | 让我告诉你，美国人最终如何看待你作为独立顾问的表现。<br>美国人完全不知道您是独立律师。 | Slate | Contradiction |
| Arabic | تحتاج الوكالات لأن تكون قادرة على قياس مستويات النجاح.<br>لا يمكنالوكالات أ العرف ما إذا كانت ناجحة أم لا | Nine-Eleven | Contradiction |



## Challenges with Monolingual LMs

- There is not enough unlabeled data for each language

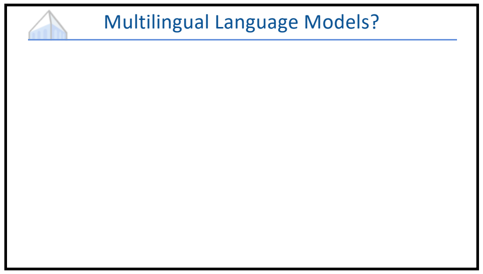
Credit: Graham Neubig



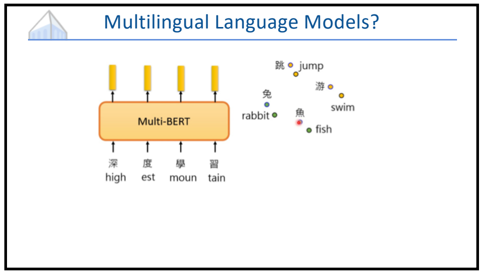
## Challenges with Monolingual LMs

- Compute and complexity of serving 100-1000s of different models

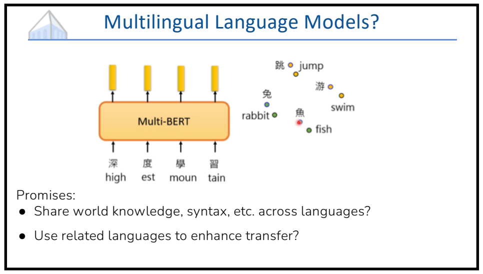
Credit: Graham Neubig



## Multilingual Language Models?



## Multilingual Language Models?



## Multilingual Language Models?

Promises:

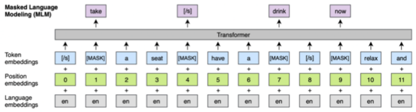
- Share world knowledge, syntax, etc. across languages?
- Use related languages to enhance transfer?

## Multilingual BERT

- Simply rerun BERT with 100+ language's Wikipedia and new BPE

## Multilingual BERT

- Simply rerun BERT with 100+ language's Wikipedia and new BPE

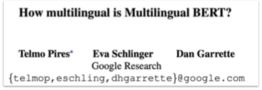


## Multilingual BERT

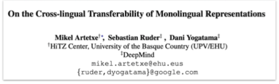
- This works surprisingly well!

## Multilingual BERT

- This works surprisingly well!


**How multilingual is Multilingual BERT?**

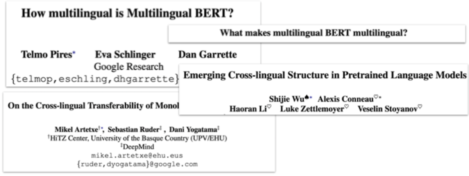
**Telmo Pires*** **Eva Schlinger** **Dan Garrette**
Google Research
{telmop,eschling,dhgarrette}@google.com


**On the Cross-lingual Transferability of Monolingual Representations**

**Mikel Artetxe**[†*], **Sebastian Ruder**[‡], **Dani Yogatama**[‡]
[†]HiTZ Center, University of the Basque Country (UPV/EHU)
[‡]DeepMind
mikel.artetxe@ehu.eus
{ruder,dyogatama}@google.com

## Multilingual BERT

- This works surprisingly well!


**How multilingual is Multilingual BERT?**

**Telmo Pires*** **Eva Schlinger** **Dan Garrette**
Google Research
{telmop,eschling,dhgarrette}

**What makes multilingual BERT multilingual?**

**Emerging Cross-lingual Structure in Pretrained Language Models**

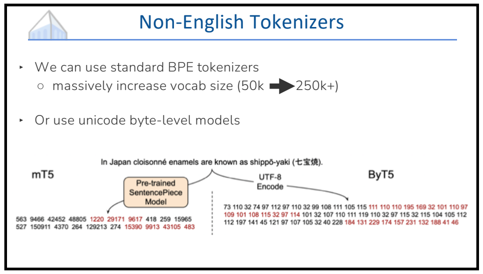
**Shijie Wu**[♠*] **Alexis Conneau**[♡*]
**Haoran Li**[♡] **Luke Zettlemoyer**[♡] **Veselin Stoyanov**[♡]

**On the Cross-lingual Transferability of Monol**

**Mikel Artetxe**[†*], **Sebastian Ruder**[‡], **Dani Yogatama**[‡]
[†]HiTZ Center, University of the Basque Country (UPV/EHU)
[‡]DeepMind
mikel.artetxe@ehu.eus
{ruder,dyogatama}@google.com

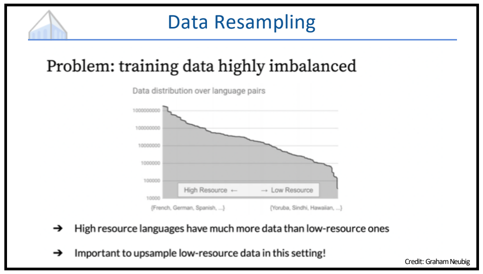
## Non-English Tokenizers

- We can use standard BPE tokenizers
  - massively increase vocab size (50k ➡ 250k+)



## Non-English Tokenizers

- We can use standard BPE tokenizers
  - massively increase vocab size (50k ➡ 250k+)
- Or use unicode byte-level models



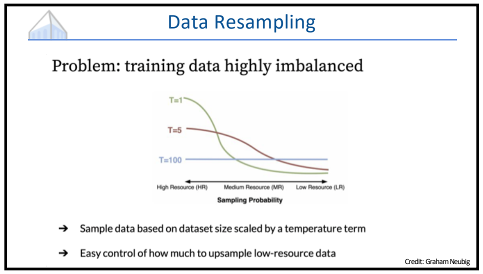
## Data Resampling

Problem: training data highly imbalanced

➔ High resource languages have much more data than low-resource ones

➔ Important to upsample low-resource data in this setting!

Credit: Graham Neubig



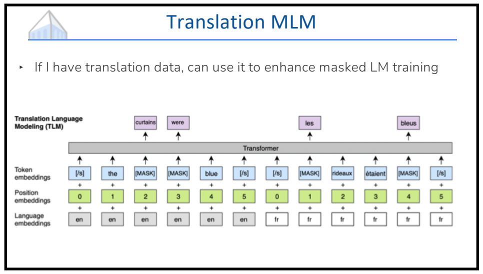
## Data Resampling

Problem: training data highly imbalanced

➔ Sample data based on dataset size scaled by a temperature term

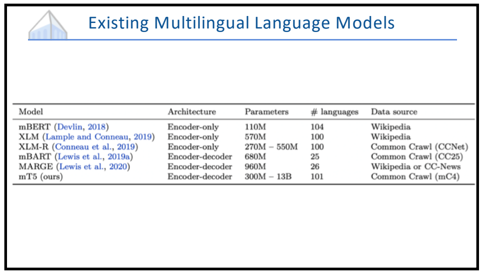
➔ Easy control of how much to upsample low-resource data

Credit: Graham Neubig



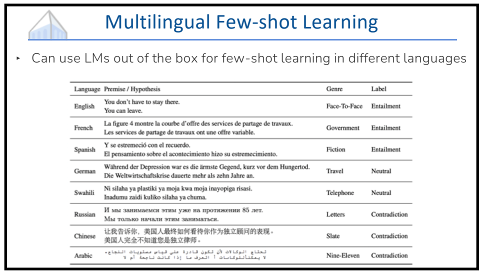
## Translation MLM

- If I have translation data, can use it to enhance masked LM training



## Existing Multilingual Language Models

| Model | Architecture | Parameters | # languages | Data source |
|---|---|---|---|---|
| mBERT (Devlin, 2018) | Encoder-only | 110M | 104 | Wikipedia |
| XLM (Lample and Conneau, 2019) | Encoder-only | 570M | 100 | Wikipedia |
| XLM-R (Conneau et al., 2019) | Encoder-only | 270M – 550M | 100 | Common Crawl (CCNet) |
| mBART (Lewis et al., 2019a) | Encoder-decoder | 680M | 25 | Common Crawl (CC25) |
| MARGE (Lewis et al., 2020) | Encoder-decoder | 960M | 26 | Wikipedia or CC-News |
| mT5 (ours) | Encoder-decoder | 300M – 13B | 101 | Common Crawl (mC4) |



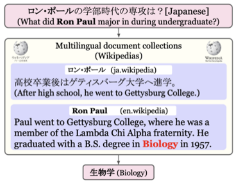
## Multilingual Few-shot Learning

- Can use LMs out of the box for few-shot learning in different languages

| Language | Premise / Hypothesis | Genre | Label |
|---|---|---|---|
| English | You don't have to stay there.<br>You can leave. | Face-To-Face | Entailment |
| French | La figure 4 montre la courbe d'offre des services de partage de travaux.<br>Les services de partage de travaux ont une offre variable. | Government | Entailment |
| Spanish | Y se estremeció con el recuerdo.<br>El pensamiento sobre el acontecimiento hizo su estremecimiento. | Fiction | Entailment |
| German | Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod.<br>Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an. | Travel | Neutral |
| Swahili | Ni silaha ya plastiki ya moja kwa moja inayopiga risasi.<br>Inadumu zaidi kuliko silaha ya chuma. | Telephone | Neutral |
| Russian | И мы занимаемся этим уже на протяжении 85 лет.<br>Мы только начали этим заниматься. | Letters | Contradiction |
| Chinese | 让我告诉你，美国人最终如何看待你作为独立顾问的表现。<br>美国人完全不知道您是独立律师。 | Slate | Contradiction |
| Arabic | لتحتاج الوكالات لأن تكون قادرة على قياس مستويات النجاح.<br>لا يمكنأللوكالات أ العرف ما إذا كانت ناجحة أم لا ؟ | Nine-Eleven | Contradiction |

## Multilingual Few-shot Learning

- Can use LMs out of the box for few-shot learning in different languages



ロン・ポールの学部時代の専攻は？[Japanese]
(What did **Ron Paul** major in during undergraduate?)

Multilingual document collections
(Wikipedias)

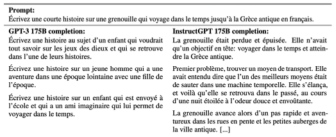
ロン・ポール (ja.wikipedia)
高校卒業後はゲティスバーグ大学へ進学。
(After high school, he went to Gettysburg College.)

**Ron Paul** (en.wikipedia)
Paul went to Gettysburg College, where he was a member of the Lambda Chi Alpha fraternity. He graduated with a B.S. degree in **Biology** in 1957.

生物学 (Biology)

## Cross-lingual Supervised Transfer

- If I have supervised data, can transfer to languages w/o supervised data

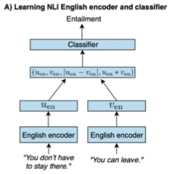


**Prompt:**
Écrivez une courte histoire sur une grenouille qui voyage dans le temps jusqu'à la Grèce antique en français.

| **GPT-3 175B completion:** | **InstructGPT 175B completion:** |
|---|---|
| Écrivez une histoire au sujet d'un enfant qui voudrait tout savoir sur les jeux des dieux et qui se retrouve dans l'une de leurs histoires.<br>Écrivez une histoire sur un jeune homme qui a une aventure dans une époque lointaine avec une fille de l'époque.<br>Écrivez une histoire sur un enfant qui est envoyé à l'école et qui a un ami imaginaire qui lui permet de voyager dans le temps. | La grenouille était perdue et épuisée. Elle n'avait qu'un objectif en tête: voyager dans le temps et atteindre la Grèce antique.<br>Premier problème, trouver un moyen de transport. Elle avait entendu dire que l'un des meilleurs moyens était de sauter dans une machine temporelle. Elle s'élança, et voilà qu'elle se retrouva dans le passé, au cours d'une nuit étoilée à l'odeur douce et envoûtante.<br>La grenouille avance alors d'un pas rapide et aventureux dans les rues en pente et les petites auberges de la ville antique. [...] |

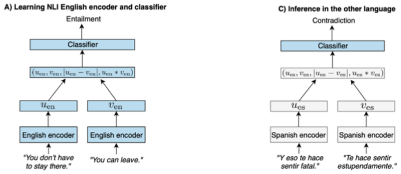
## Cross-lingual Supervised Transfer

- If I have supervised data, can transfer to languages w/o supervised data


A) Learning NLI English encoder and classifier
Entailment
Classifier
$(u_{\mathrm{en}}, v_{\mathrm{en}}, |u_{\mathrm{en}} - v_{\mathrm{en}}|, u_{\mathrm{en}} * v_{\mathrm{en}})$
$u_{\mathrm{en}}$
$v_{\mathrm{en}}$
English encoder
English encoder
"You don't have to stay there."
"You can leave."

## Cross-lingual Supervised Transfer

- If I have supervised data, can transfer to languages w/o supervised data


A) Learning NLI English encoder and classifier
Entailment
Classifier
$(u_{\mathrm{en}}, v_{\mathrm{en}}, |u_{\mathrm{en}} - v_{\mathrm{en}}|, u_{\mathrm{en}} * v_{\mathrm{en}})$
$u_{\mathrm{en}}$
$v_{\mathrm{en}}$
English encoder
English encoder
"You don't have to stay there."
"You can leave."
C) Inference in the other language
Contradiction
Classifier
$(u_{\mathrm{es}}, v_{\mathrm{es}}, |u_{\mathrm{es}} - v_{\mathrm{es}}|, u_{\mathrm{es}} * v_{\mathrm{es}})$
$u_{\mathrm{es}}$
$v_{\mathrm{es}}$
Spanish encoder
Spanish encoder
"Y eso te hace sentir fatal."
"Te hace sentir estupendamente."