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

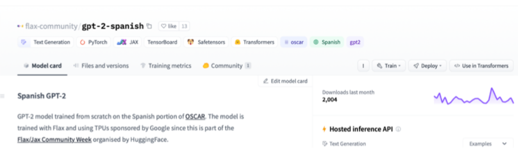
Google Translate

Pros:
- Straightforward to implement
- Surprisingly strong baseline, especially for classification tasks

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

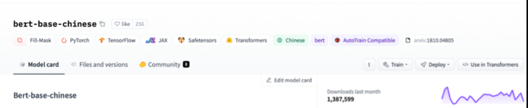
Monolingual LMs

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



Monolingual LMs

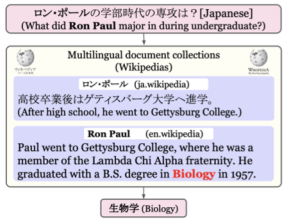
- Can we just repeat the LM pre-training pipeline for other languages?
  - Sort of!



Few-shot Learning in Other Languages

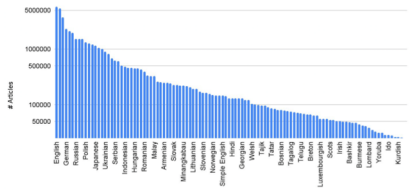
Language	Premise / Hypothesis	Genre	Label
English	You don't have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. 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。 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. لا يمكنللوكالات أ التعرف ما إذا كانت ناجحة أم لا	Nine-Eleven	Contradiction

Few-shot Learning in Other Languages



Challenges with Monolingual LMs

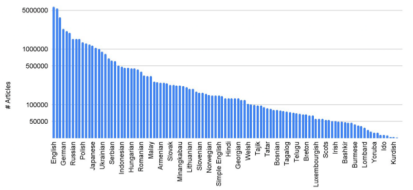
- There is not enough unlabeled data for each language



Credit: Graham Neubig

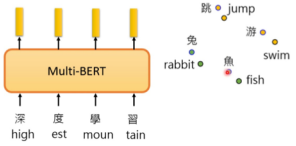
Challenges with Monolingual LMs

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



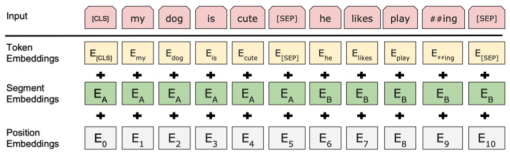
Credit: Graham Neubig

Multilingual Language Models?



Multilingual BERT

- Simply rerun BERT, except use 100+ Wikipedias and new BPE

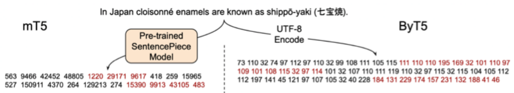


Non-English Tokenizers

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

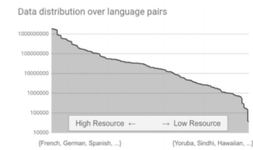
Non-English Tokenizers

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



Multilingual Language Models

Problem: training data highly imbalanced

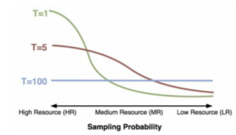


- ➔ High resource languages have much more data than low-resource ones
- ➔ Important to upsample low-resource data in this setting!

Credit: Graham Neubig

Multilingual Language Models

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

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

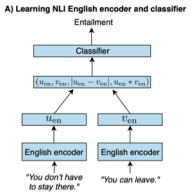
Language	Premise / Hypothesis	Genre	Label
English	You don't have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. 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。 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. لا يمكننا أن نعرف ما إذا كانت ناجحة أم لا	Nine-Eleven	Contradiction

Cross-lingual Transfer

- If I have English supervised data, can transfer to low-resource language

Cross-lingual Transfer

- If I have English supervised data, can transfer to low-resource language



Cross-lingual Transfer

- If I have English supervised data, can transfer to low-resource language