Assessing Large LMs and their Multilingualism

François Yvon

LISN — CNRS and Université Paris-Saclay

ALPS winter school

january 2023

Evaluating LMs, a serious matter

In recent years, the AI technology that has arguably advanced the most is foundation models (Bommasani et al., 2021), headlined by the rise of language models (LMs; Peters et al., 2018; Devlin et al., 2019; Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022). At its core, a language model is a box that takes in text and generates text (...) Yet the immense surface of model capabilities, limitations, and risks remains poorly understood. The rapid development, rising impact, and inadequate understanding demand that we benchmark language models holistically. [Liang et al., 2022]

The multiple dimensions of LM evaluation

As text generators	As representation learners
Good models of actual texts?	Recovering realistic word associations?
Discriminating well-formed utterances?	Useful for downstream applications?
Generating realistic texts?	- in 'unstructured' learners?
Enabling controlled generation?	- in 'structured' learners ?

Evaluating LMs, a serious matter

In recent years, the AI technology that has arguably advanced the most is foundation models (Bommasani et al., 2021), headlined by the rise of language models (LMs; Peters et al., 2018; Devlin et al., 2019; Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022). At its core, a language model is a box that takes in text and generates text (...) Yet the immense surface of model capabilities, limitations, and risks remains poorly understood. The rapid development, rising impact, and inadequate understanding demand that we benchmark language models holistically. [Liang et al., 2022]

The multiple dimensions of LM evaluation

As text generators	**As representation learners**
Good models of actual texts?	Recovering realistic word associations?
Discriminating well-formed utterances?	Useful for downstream applications?
Generating realistic texts?	- in 'unstructured' learners?
Enabling controlled generation?	- in 'structured' learners ?

Algorithms for Text Generation

Greedy search

$$w_0 = \texttt{< s >}$$

$$w_t = \underset{w \in \bar{\mathcal{V}}}{\operatorname{argmax}} \ \log \mathrm{P}(w \,|\, w_{<t}) \qquad \text{(for } t > 0\text{)}$$

$\bar{\mathcal{V}} = \mathcal{V} \cup \{\texttt{< s >}, \texttt{< /s >}\}$. Generation stops with the < /s > symbol or when some maximum time step T is reached.

Algorithms for Text Generation

Ancestral sampling

$$w_0 = \text{< s >}$$

$$w_t \sim \mathrm{P}(w \mid w_{<t}) \qquad \text{(for } t > 0\text{)}$$

Recursion stops with the < /s > symbol or when some maximum time step T is reached. Add a temperature parameter α to flatten / sharpen the distribution with $\mathrm{P}(w \mid w_{<t})^{\alpha}$

Algorithms for Text Generation

Top-k sampling

$$w_0 = \texttt{<s>}$$

$$Q(w_t \mid w_{<t}) \propto \begin{cases} P(w_t \mid w_{<t}) \text{ if } w \in \text{top-k}(P(W \mid w_{<t})) \\ 0 \text{ otherwise} \end{cases}$$

$$w_t \sim Q(w \mid w_{<t}) \qquad (\text{for } t > 0)$$

Sample from a "truncated" distribution containing the k most likely symbols. Generation stops with the < /s > symbol or when some maximum time step T is reached.

Algorithms for Text Generation

Nucleus sampling (top p, with variable p)

$$w_0 = \texttt{<s>}$$

$$Q(w_t \,|\, w_{<t}) \propto \begin{cases} P(w_t \,|\, w_{<t}) & \text{if } w \in \text{top-p}(P(W \,|\, w_{<t})) \\ 0 & \text{otherwise} \end{cases}$$

$$w_t \sim Q(w \,|\, w_{<t}) \qquad (\text{for } t > 0)$$

Where p is the smallest integer such that $\sum_{w \in \text{top-p}} P(w \,|\, w_{<t}) > \alpha$. Sample from a "truncated" distribution for the p most likely symbols, with variable p (α typically $\in [0.7; 0.9]$).

Evaluating Language Models with Perplexity

Perplexity of a test sequence $w_{[1:T]}$ [Brown et al., 1992]

$$\mathrm{PPL}(M) = 2^{\frac{-1}{T}\log_2 \mathrm{P}(w_{[1:T]}|M)} = \mathrm{P}(w_{[1:T]}|M)^{-\frac{1}{T}}$$

- The cross-entropy between the source (S) and model M:

$$H(S,M) = \lim_{T\to\infty} \frac{-1}{T} \log_2 \mathrm{P}(w_{[1:T]}|M)$$

  $H(S,M)$ upper bounds $H(S)$

- PLL() homogeneous to a vocabulary size

$$\mathrm{PPL}(\mathrm{Unif}) = 2^{\frac{-1}{T}\log_2 \mathrm{P}(w_{[1:T]}|M)} = 2^{\frac{-1}{T}T\log_2(1/n_W)} = n_W$$

Evaluating Language Models with Perplexity

Comparing LMs with different support or segmentations?

1. closed-world LMs assume a fixed vocabulary size n_W - models with different n_W cannot be compared.
2. open-world models with different segmentations can be compared, must use the same normalizer ($\frac{1}{T}$)
3. typical normalizers when using subwords
   - number of chars
   - number of bytes

Evaluating with linguistic contrasts

Do LSTMs learn long-range structural dependencies ? Results from [Linzen et al., 2016]

Linguistic phenomena

Subject/Verb agreement in English is a structural constraint

The keys to the cabinet [are] on the table

(*) *The keys to the cabinet [is] on the table*

Subject can be arbitrarily remote from the main verb

The keys to the cabinet where I used to store my very precious mother's book [are] ...

Evaluating with linguistic contrasts

Do LSTMs learn long-range structural dependencies ? Results from [Linzen et al., 2016]

Experimental design : supervised training

Train a LM-RNN to predict the verb number

The key to the cabinet VRB-SING

The keys to the cabinet VRB-PLUR

Main findings

- performance is very good (error < 1 %)
- performance (slowly) drops with subject-verb distance
- performance (slowly) drops with intervening distractors

The keys/PLUR to the cabinet/SING in my boat/SING ...

Evaluating with linguistic contrasts

Do LSTMs learn long-range structural dependencies ? Results from [Linzen et al., 2016]

Experimental design : unsupervised training

Train a LM-RNN as a language model to predict next word. Compare scores $\log \mathrm{P}(w \mid w_{<t})$ of two alternatives $w \in \{$ *is*, *are* $\}$:

The key to the cabinet → is

The keys to the cabinet → are

Main findings

- 10-fold loss in performance (error > 6.5 %)
- for difficult cases, performance is worst than random guess

In summary, we conclude that while the LSTM is capable of learning syntax-sensitive agreement dependencies under various objectives, the language-modeling objective alone is not sufficient for learning such dependencies and a more direct form of training signal is required.

Evaluating with linguistic contrasts

Do LSTMs learn long-range structural dependencies ? Results from [Linzen et al., 2016]

Generalized linguistic evaluation Warstadt et al. [2020]

12 families of linguistic patterns, 67,000 contrastive pairs.

- Predictions for agreement constraints (det+N, subj-verb) > 90% already for GPT-2
- Other patterns are more difficult, eg. `Island constraints`

Whose hat should Tonya wear vs. Whose should Tonya wear hat

Evaluating with linguistic contrasts

Do LSTMs learn long-range structural dependencies ? Results from [Linzen et al., 2016]

Generalized linguistic evaluation Warstadt et al. [2020]

12 families of linguistic patterns, 67,000 contrastive pairs.

- Predictions for agreement constraints (det+N, subj-verb) > 90% already for GPT-2
- Other patterns are more difficult, eg. `Island constraints`

Whose hat should Tonya wear vs. Whose should Tonya wear hat

Caveats

- watch out for confounding factors
- does not really test generation

Language model (de)generation

The promise ☞ OpenAi Website https://openai.com/blog/better-language-models/

GPT-2 generates synthetic text samples in response to the model being primed with an arbitrary input. The model is chameleon-like—it adapts to the style and content of the conditioning text. This allows the user to generate realistic and coherent continuations about a topic of their choosing, as seen by the following select samples.

GPT-2 displays a broad set of capabilities, including the ability to generate conditional synthetic text samples of unprecedented quality, where we prime the model with an input and have it generate a lengthy continuation.

Language model (de)generation

The truth about language model generation

Prefix	Lyrically the song has excerpts of different languages including French , Spanish ,
GPT2-greedy	German , Italian , Portuguese , Spanish , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , ...
GPT2-top3	German , Italian , Portuguese , Spanish , Portuguese , Portuguese, Italian and Spanish. It is also available in English, French and Spanish. In addition to its lyrics, the album features a number of original songs, ...
GPT2-nucleus-0.3	German , Italian , Portuguese , Spanish , Portuguese , Italian , Portuguese , Spanish , Portuguese , Spanish , Portuguese , Spanish , Portuguese , Spanish , Portuguese , Spanish , Portuguese , Spanish , Portuguese , Spanish , Portuguese , ...
Prefix	The first pair of pereiopods is armed with a large , asymm
GPT2-greedy	etrical, and long-range laser cannon. The second pair is armed with a large , asymmetrical, and long-range laser cannon. The third pair is armed with a large , asymmetrical, and long-range laser cannon. The fourth pair is armed with a large ...
GPT2-top3	etrical, and highly mobile head, and the second pair has a small , asymmetrical, and highly mobile head. The second pair has a large and highly mobile head, and the third pair is armed with a large and highly mobile head. The first pair ...
GPT2-nucleus-0.3	etrical head and a large body. The first pair of pereiopods is armed with a large , asymmetrical head and a large body. The first pair of pereiopods is armed with a large , asymmetrical head and a large body. The first pair of pereiopods is armed ...

Examples from [Welleck et al., 2019].

Language model (de)generation

High probability sentences do not resemble human productions

- too many repetitions
- high frequency tokens over-represented, low frequency tokens under-represented
- lack of lexical diversity
- lack of global consistency
- posterior distribution poorly calibrated

Evaluating LMs with distributional properties

rep(l): a repetition metric

Given a set $\mathcal{D}$ of length-T sequences,

$$\mathrm{rep}/\ell = \frac{1}{|\mathcal{D}|T} \sum_{\mathbf{x} \in \mathcal{D}} \sum_{t=1}^{T} \mathbb{I}\left[\arg\max \mathrm{P}(x_t \,|\, x_{<t}) \in \mathbf{x}_{t-\ell-1:t-1}\right].$$

Generalizes to repeated n-gram sequences

Global distributional properties [Meister and Cotterell, 2021]

- Zipfian behavior, power-law distribution

$$\mathrm{P}_{\mathrm{zipf}}(W = w_k) \propto k^{-s}, s \approx 1$$

- type-token ratios (TTR) (depends on length)
- proportion of frequency 1 words (hapax legomena)
- proportion specific of token classes (punctuation, stopwords etc)
- consistency metrics ?

Linguistic Probes

A common methodology to answer questions about the structure of models is to associate internal representations with external properties, by training a classifier on said representations that predicts a given property. This framework, known as probing classifiers, has emerged as a prominent analysis strategy in many studies of NLP models. [Belinkov, 2022]

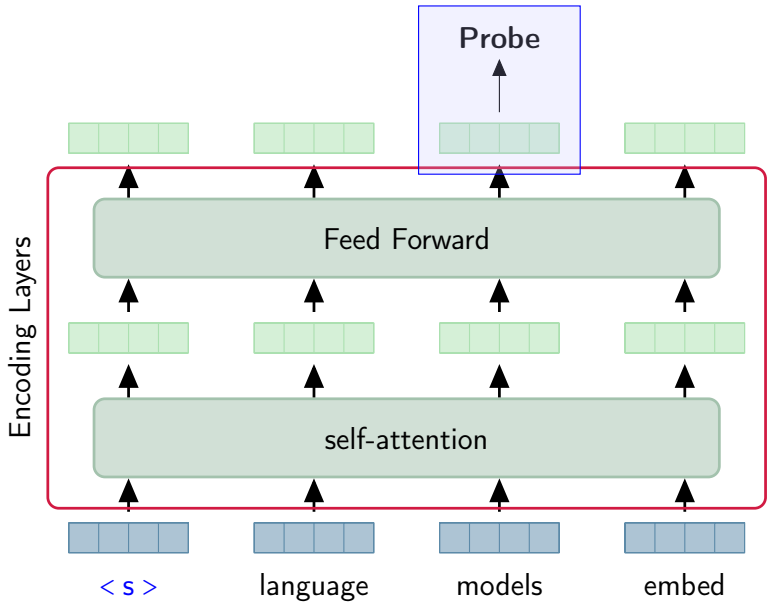


Figure (C) G. Wisniewski

Probing representations

- encoder outputs contextualized representation of input tokens
- linguistic probes: supervised classifier to predict linguistic properties from the token representation
- small error rate ⇒ linguistic property encoded in representation

Linguistic Probes

A common methodology to answer questions about the structure of models is to associate internal representations with external properties, by training a classifier on said representations that predicts a given property. This framework, known as probing classifiers, has emerged as a prominent analysis strategy in many studies of NLP models. [Belinkov, 2022]

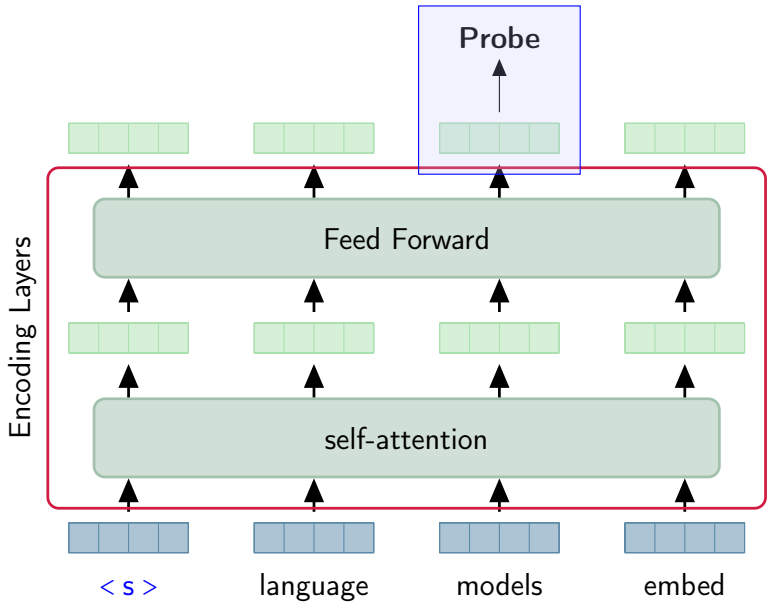


Figure (C) G. Wisniewski

Probing words

- what is the (gender, number, tense, mode, etc) of current token?
- what is the syntactic head of current token?
- what is the (gender, number, tense, mode, etc) of current token's head ?

Linguistic Probes

A common methodology to answer questions about the structure of models is to associate internal representations with external properties, by training a classifier on said representations that predicts a given property. This framework, known as probing classifiers, has emerged as a prominent analysis strategy in many studies of NLP models. [Belinkov, 2022]

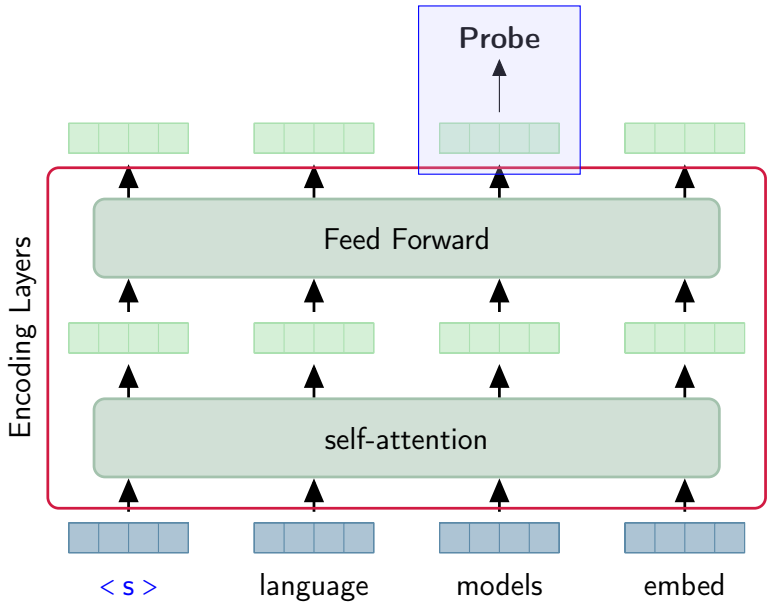


Figure (C) G. Wisniewski

Probing sentences via < s >

- what is the sentence length?
- does sentence contain word X?
- what is the (gender, number, tense, mode, etc) of sentence head?

Linguistic Probes

A common methodology to answer questions about the structure of models is to associate internal representations with external properties, by training a classifier on said representations that predicts a given property. This framework, known as probing classifiers, has emerged as a prominent analysis strategy in many studies of NLP models. [Belinkov, 2022]

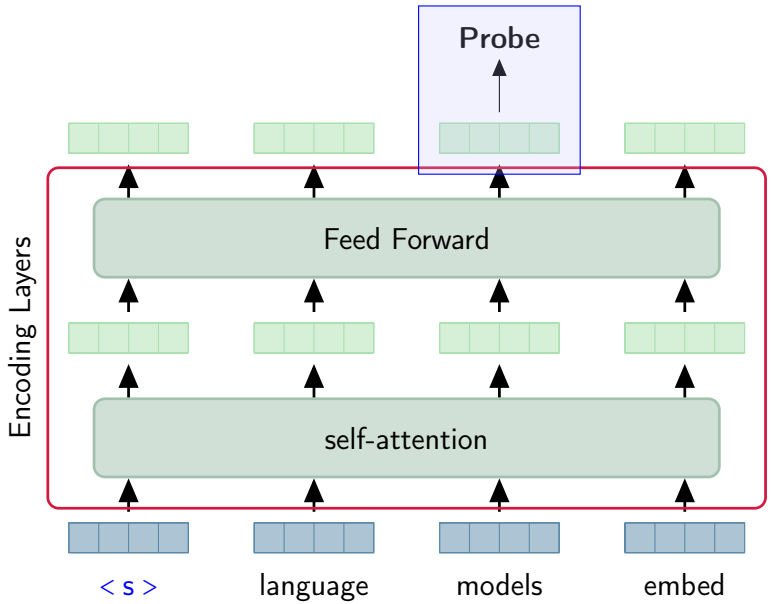
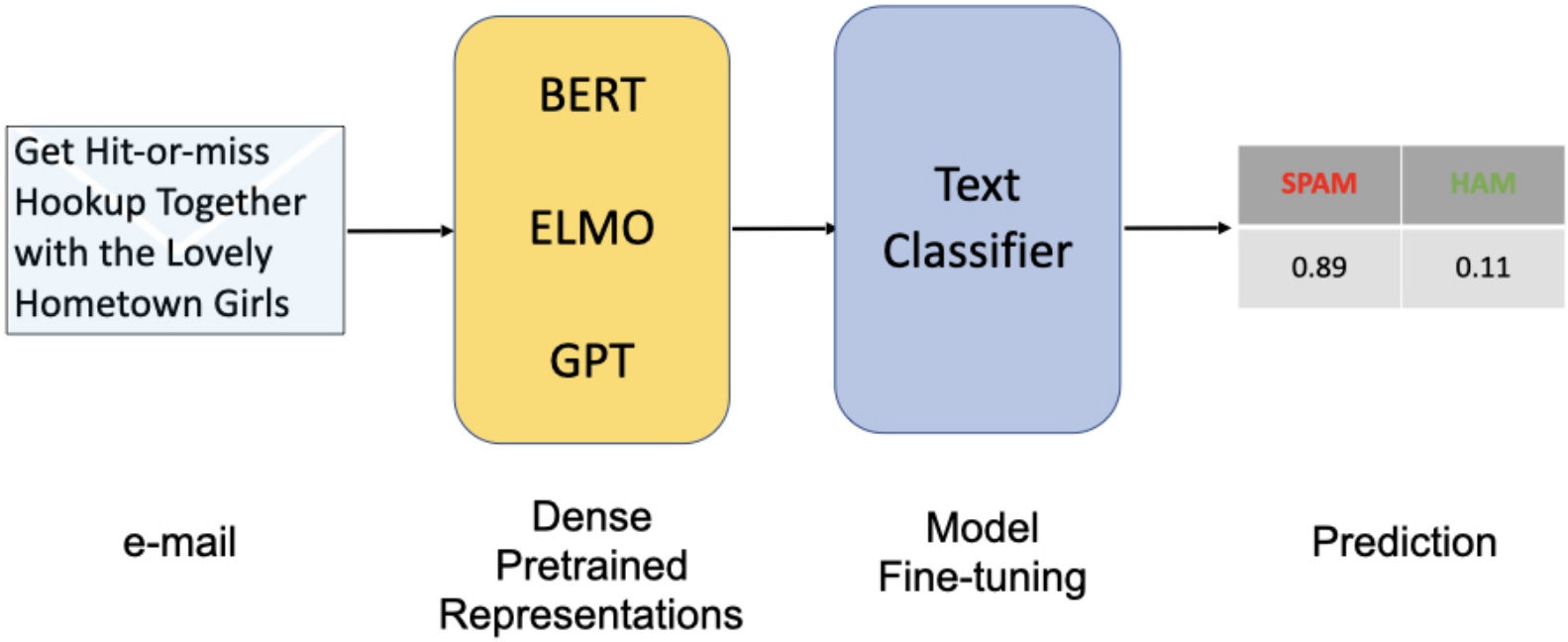


Figure (C) G. Wisniewski

Caveats

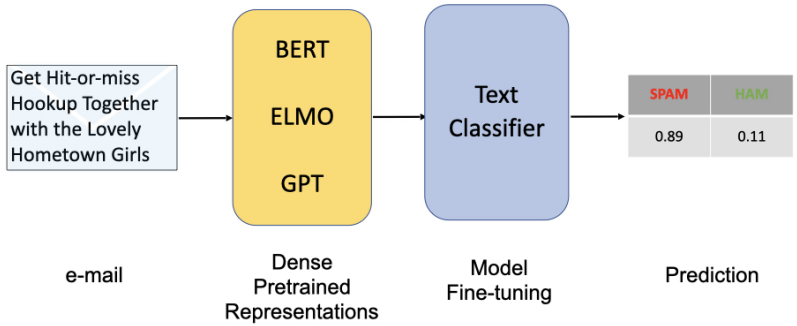
- use weak classifiers and contrast with random labels
- how good should classification accuracy be ?
- a feature is learned does not mean it is actually used

Using Representations in Downstream Tasks



Improving spam filtering with pre-trained representations

Using Representations in Downstream Tasks



Task Families, structured and unstructured

- text mining: text classification, sentiment analysis, hate speech detection, textual entailment, etc
- linguistic tasks: POS tagging, dependency parsing, named entities recognition (NER), etc

Using Representations in Downstream Tasks

Larger models, larger benchmarks

- SuperGlue: a monolingual dataset for representation learning (10 tasks) Wang et al. [2019]
- BigBench (204 tasks) *(...) drawing problems from linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development, and beyond.*

Evaluating zero-shot / few-shot behaviour

Reduce NLP tasks to text generation with appropriate instructions in NL as prompts

Prompts = instructions in Natural Language + [tricks] (from [Brown et al., 2020])

Specifically, we evaluate GPT-3 on over two dozen NLP datasets,(...) For each task, we evaluate GPT-3 under 3 conditions:

- "zero-shot" learning, where no demonstrations are allowed and only an instruction in natural language is given to the model.
  "Evaluate 125 + 12 ="
- "one-shot learning", where we allow only one demonstration, and
  "Evaluate 17 + 301 = 318 < /s >Evaluate 125 + 12 = "
- "few-shot learning", or in-context learning, where we allow as many demonstrations as will fit into the model's context window,
  "Evaluate 17 + 301 = 318 < /s >Evaluate 48 + 67 = 105 < /s >Evaluate 125 + 12 = "

Tricks: *"On tasks with free-form completion, we use beam search with the same parameters as [RSR+19]: a beam width of 4 and a length penalty of $\alpha = 0.6$."* (+ stopping criterion)

Evaluating zero-shot / few-shot behaviour

Reduce NLP tasks to text generation with appropriate instructions in NL as prompts

Task types and their evaluation

Assuming prompt / instruction: $w_1 \dots w_T$.

- Yes / No answers

  Question: [`Question`] True or false? *[prediction]*

  Correct if $\mathrm{P}(\text{True} \,|\, \text{prompt}) > \mathrm{P}(\text{False} \,|\, \text{prompt})$.

- Multiple choice answers.

  Question: Which factor will most likely cause a person to develop a fever?

Correct Answer	*a bacterial population in the bloodstream*
Incorrect Answer	*a leg muscle relaxing after exercise*
Incorrect Answer	*several viral particles on the skin*
Incorrect Answer	*carbohydrates being digested in the stomach*

  Correct if $\mathrm{P}(\text{Correct answer} \,|\, \text{prompt}) > \mathrm{P}(\text{Alternative} \,|\, \text{prompt})$

- One word continuation. Correct if $w_{T+1} == w^*$
- Multiple word continuation. Measure $\Delta(w_{T+1} \dots w_{T+S}; w_1^* \dots w_L^*)$ with $\Delta()$ task-dependent distance (ROUGE for summarization, BLEU for MT, etc)

Evaluating zero-shot / few-shot behaviour

Reduce NLP tasks to text generation with appropriate instructions in NL as prompts

Understanding "instruction learning" results

Should pay attention to:

- how much effort went into prompting ?
- free generation or text infilling or multi-choice answers ?
- how were alternatives selected / generated ?
- how was search performed (greedy or beam) ?
- how does generation stops
- how many shots is few shots?

Open issues

- generating / optimizing discrete / continuous prompts
- training with prompts and meta-learning

Towards Multilingual LMs

Monolingual pre-trained Language Models

- Generate texts, one token at a time
- Compute dense representations

Towards Multilingual LMs

Monolingual pre-trained Language Models

- Generate texts, one token at a time
- Compute dense representations

Multilingual pre-trained Language Models

- Generate texts in multiple languages
- Compute dense multilingual representations

Towards Multilingual LMs

Multilingual pre-trained Language Models

- Generate texts in multiple languages
- Compute dense multilingual representations

Training multilingual models

Basic requirements: [Conneau et al., 2020]

- language independent representations (multilingual BPEs or wordpieces)
- multilingual corpora

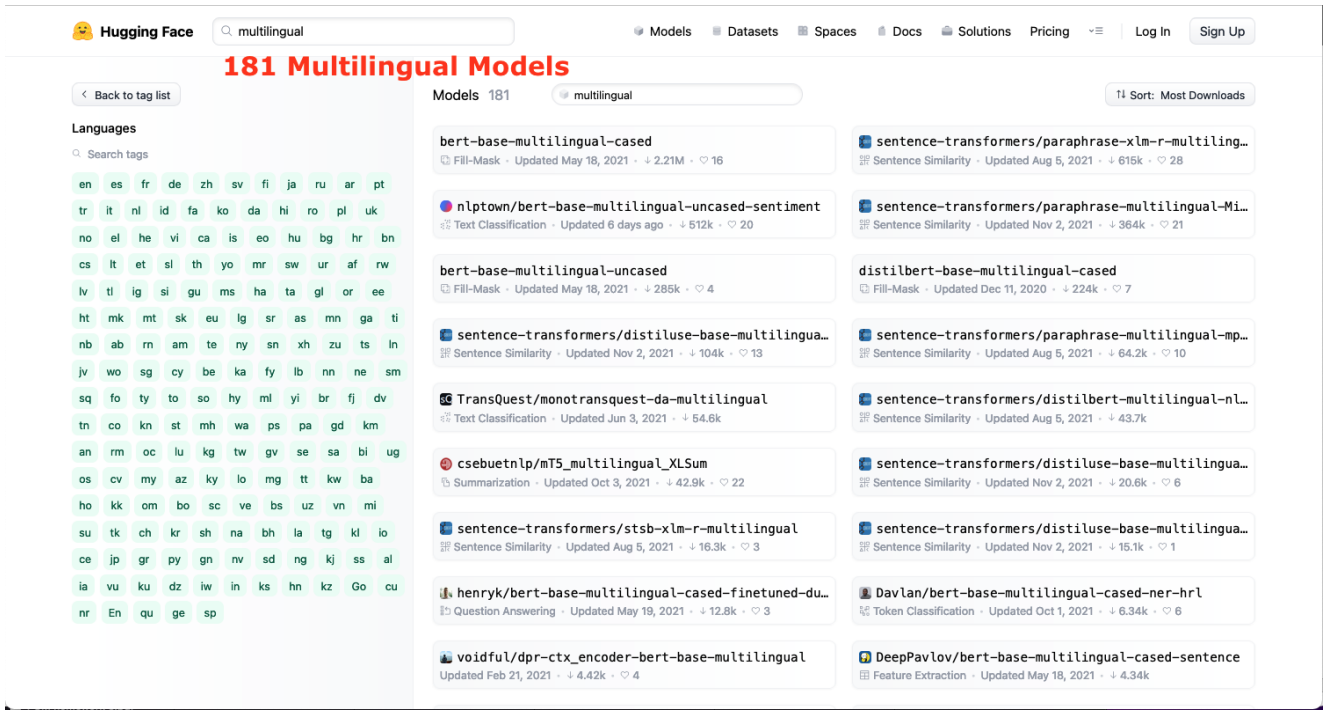
Training is business as usual (NWP, MLM),

- pay attention to the data distribution
- parallel data and dedicated losses help a bit [Ouyang et al., 2021, Chi et al., 2021]

Towards Multilingual LMs

Multilingual pre-trained Language Models

- Generate texts in multiple languages
- Compute dense multilingual representations



The bloom of multilingual models in NLP

Towards Multilingual LMs

Multilingual pre-trained Language Models

- Generate texts in multiple languages
- Compute dense multilingual representations

The landscape of multilingual LMs

- “pure representations learners”: mBERT (100 languages), XLM-R (100 languages), Serengeti (517 languages)
- "causal" LMs: mGPT (60 languages), BLOOM (46 languages), XGLM (30 languages), PALM (100 languages)
- others: mBART (50+ languages), mT5 (105 languages), YaLM (Russian-English), GLM (Chinese-English), AlexaTM (12 languages), etc
- + multilingual translation models M2M (100 languages), NLLB (200 languages), etc

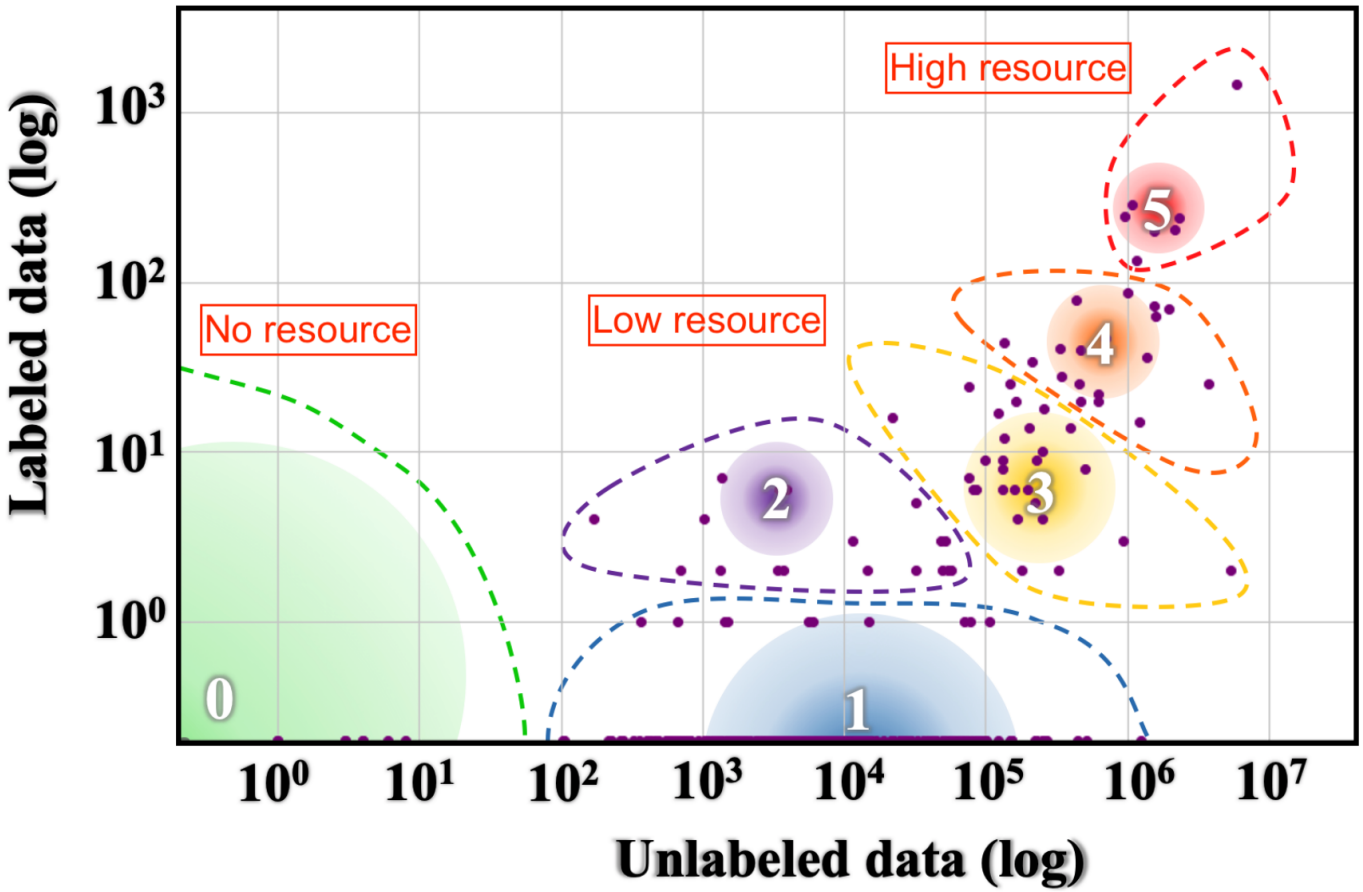
Next target - 1000 languages (if only we could count them) ?

Issues in multilingual LMs

Diversity of phonological, morphological and grammatical systems

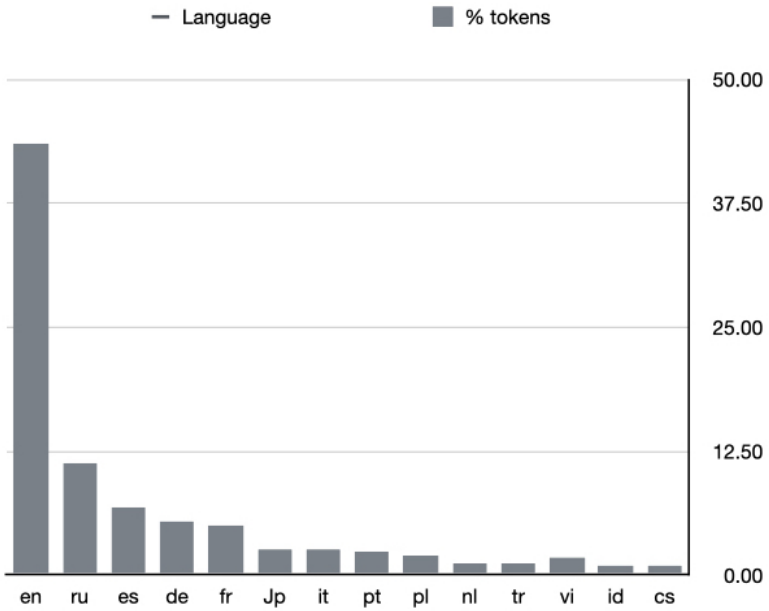
Issues in multilingual LMs

Resource unbalance

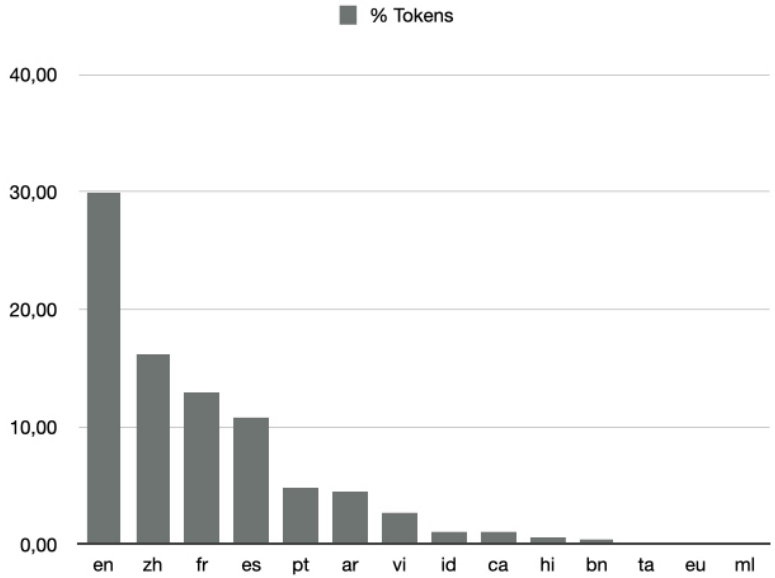


☞ (Joshi et al, 2020) https://arxiv.org/pdf/2004.09095.pdf

Issues in multilingual LMs



Languages in mT5
(104 languages)



Languages in Bloom
(46 languages)

Evaluating mLMs, a very serious matter

The multiple dimensions of mLM evaluation

As text generators	As representation learners
Good models of actual texts? [1]	Recovering realistic word associations? [3]
Discriminating well-formed sentences? [1]	Useful for downstream applications? [4]
Generating realistic texts? [2]	- in 'unstructured' learners ?
Enabling controlled generation?	- in 'structured' learners ?

[0] Evaluate mLMs as monolingual language models

New questions:

[1] equally accurate for all languages ? proportional to the training datasize ?

[2] does this include code-switched text ?

[3] how about cross-lingual word associations?

[4] including cross-lingual transfer learning?

[*] improving over monolingual LMs? For all languages?

Evaluating mLMs, a very serious matter

The multiple dimensions of mLM evaluation

As text generators	As representation learners
Good models of actual texts? [1]	Recovering realistic word associations? [3]
Discriminating well-formed sentences? [1]	Useful for downstream applications? [4]
Generating realistic texts? [2]	- in 'unstructured' learners ?
Enabling controlled generation?	- in 'structured' learners ?

[0] Evaluate mLMs as monolingual language models

New questions:

[1] equally accurate for all languages ? proportional to the training datasize ?

[2] does this include code-switched text ?

[3] how about cross-lingual word associations?

[4] including cross-lingual transfer learning?

[*] improving over monolingual LMs? For all languages?

The Perplexing Perplexity of Multilingual LMs

Questions that you can answer

Comparing open world models M_1 and M_2 perplexity-wise is fair if:

1. performed with the exact same text $w_{1:T} = w_1 \dots w_T$
2. computed with same normalizer (eg. T) irrespective of underlying segmentation (*)
3. $PPL(M_i) = \mathrm{P}(w_{1:T} \,|\, M_i))^{-\frac{1}{T}} = 2^{-\frac{1}{T}\sum_{s=1}^{S} \mathrm{P}(u_s \,|\, u_{<s}, M_i)}$

then

$$PPL(M_1) < PPL_2(M_2) \Rightarrow M_1 \text{ better model than } M_2$$

(*) caveat: as we should sum over segmentations but usually do not, we should control this approximation works equally well for all languages.

The Perplexing Perplexity of Multilingual LMs

Questions that you can answer

For multilingual models,

1. $w_{1:T} = w_1 \ldots w_T$ should be a multilingual text
2. and perplexity computed with the same normalizer irrespective of segmentation

Fair design of $w_{1:T}$? Use multi-parallel (Bible, Europarl) or comparable corpora to make all languages equally difficult (*).

(*) caveat - this is notwithstanding translationese issues - translated texts are simpler - are they ? [Mielke et al., 2019]

The Perplexing Perplexity of Multilingual LMs

Questions that you cannot answer

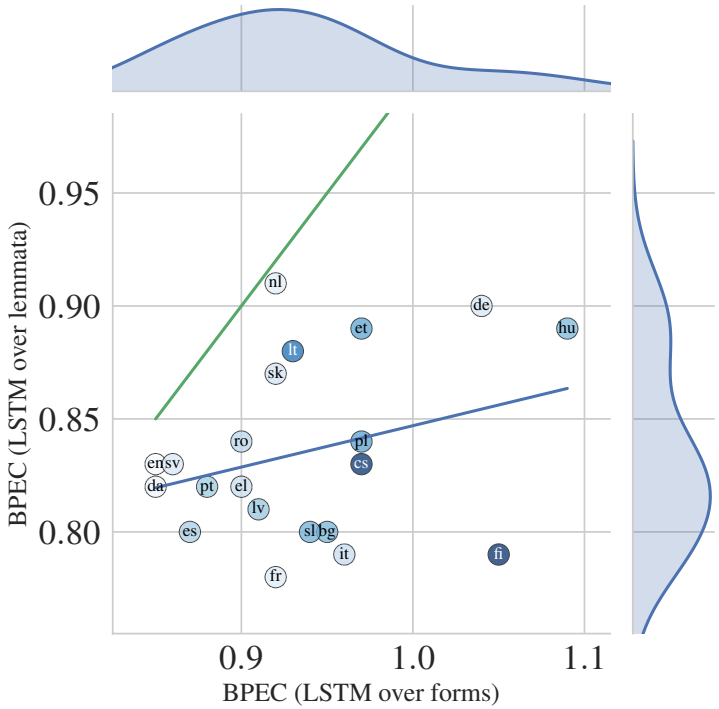
Comparing perplexity across languages is tricky

1. morphological divergences make lexical variation more or less severe: 5 forms / verb in English, ≈ 50 in French, thousands in Georgian
   morphologically simple languages are easier to predict
2. syntactic divergences make word order more or less fixed
   fixed word order languages are easier to predict
3. using the same text is not possible - use translations instead?

Not all languages are equal, some are intrinsically "difficul-to-language-model" languages

The Perplexing Perplexity of Multilingual LMs

Questions that you cannot answer



Comparing language perplexity normalized per English characters for LSTM based language models. [Mielke et al., 2019]

The Perplexing Perplexity of Multilingual LMs

Questions that you cannot answer

How well is a language modeled in a multilingual LM ?

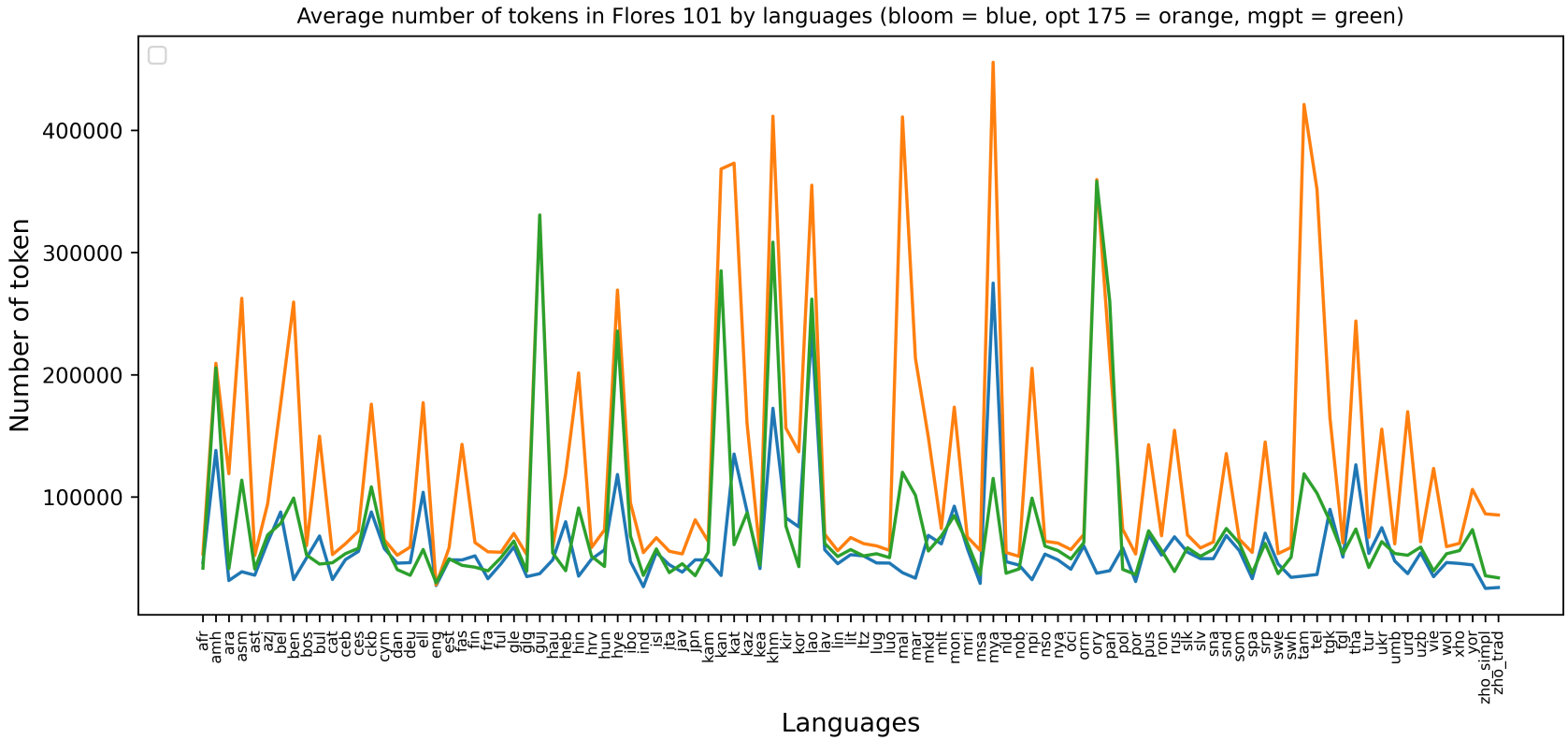
Experimental protocol

- 101 languages
- Flores multiparallel test set, Wikipedia, translations from English [Goyal et al., 2022]
- 3 multilingual models:
  - Bloom : 46 languages, byte level BPEs, 176B
  - mGPT: 60 languages, char level BPE, 13B
  - OPT: only English, byte level BPE, 175B

The Perplexing Perplexity of Multilingual LMs

Questions that you cannot answer

How well is a language modeled in a multilingual LM?

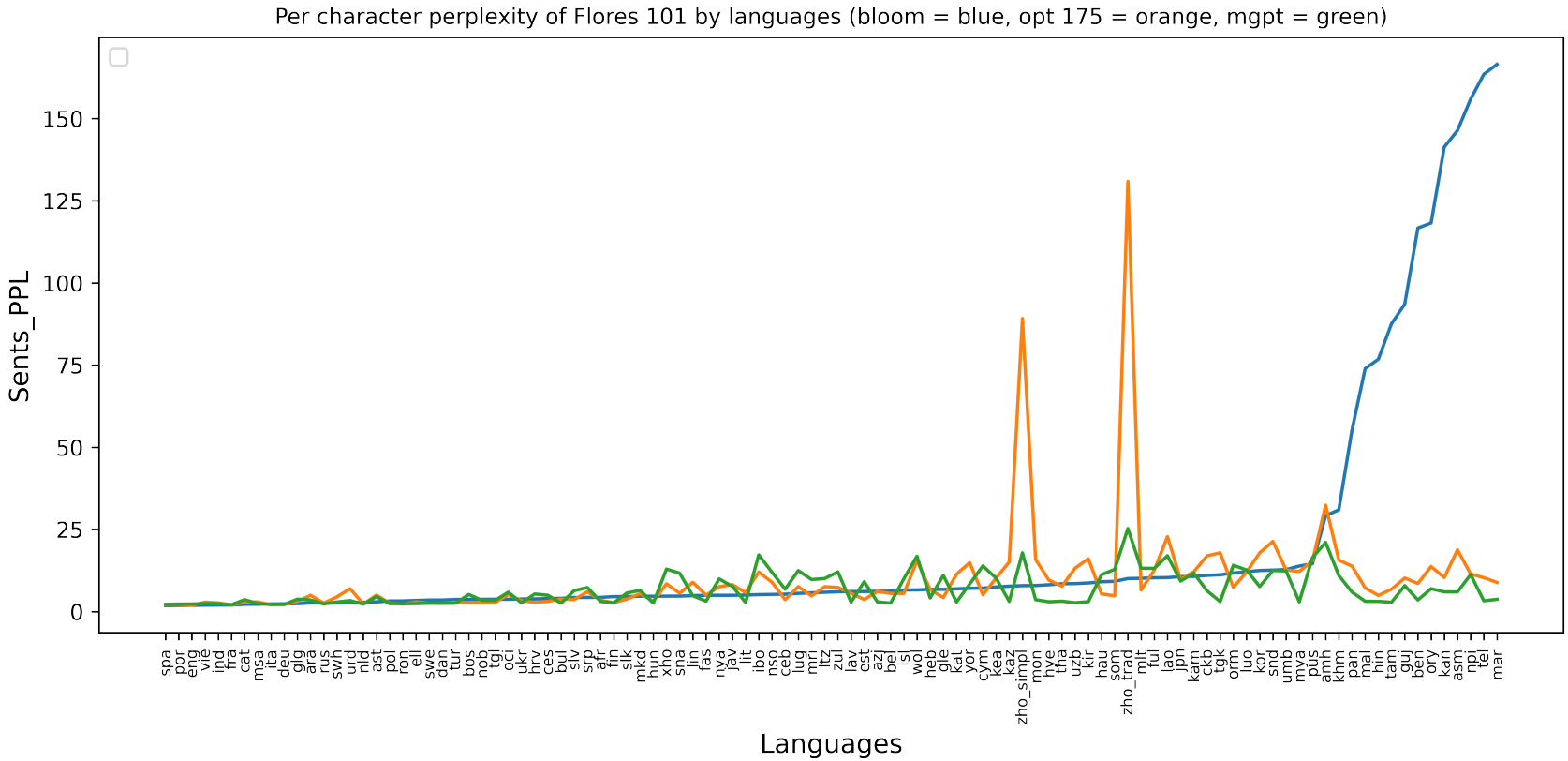


The variance of length: do not normalize per token.

The Perplexing Perplexity of Multilingual LMs

Questions that you cannot answer

How well is a language modeled in a multilingual LM?

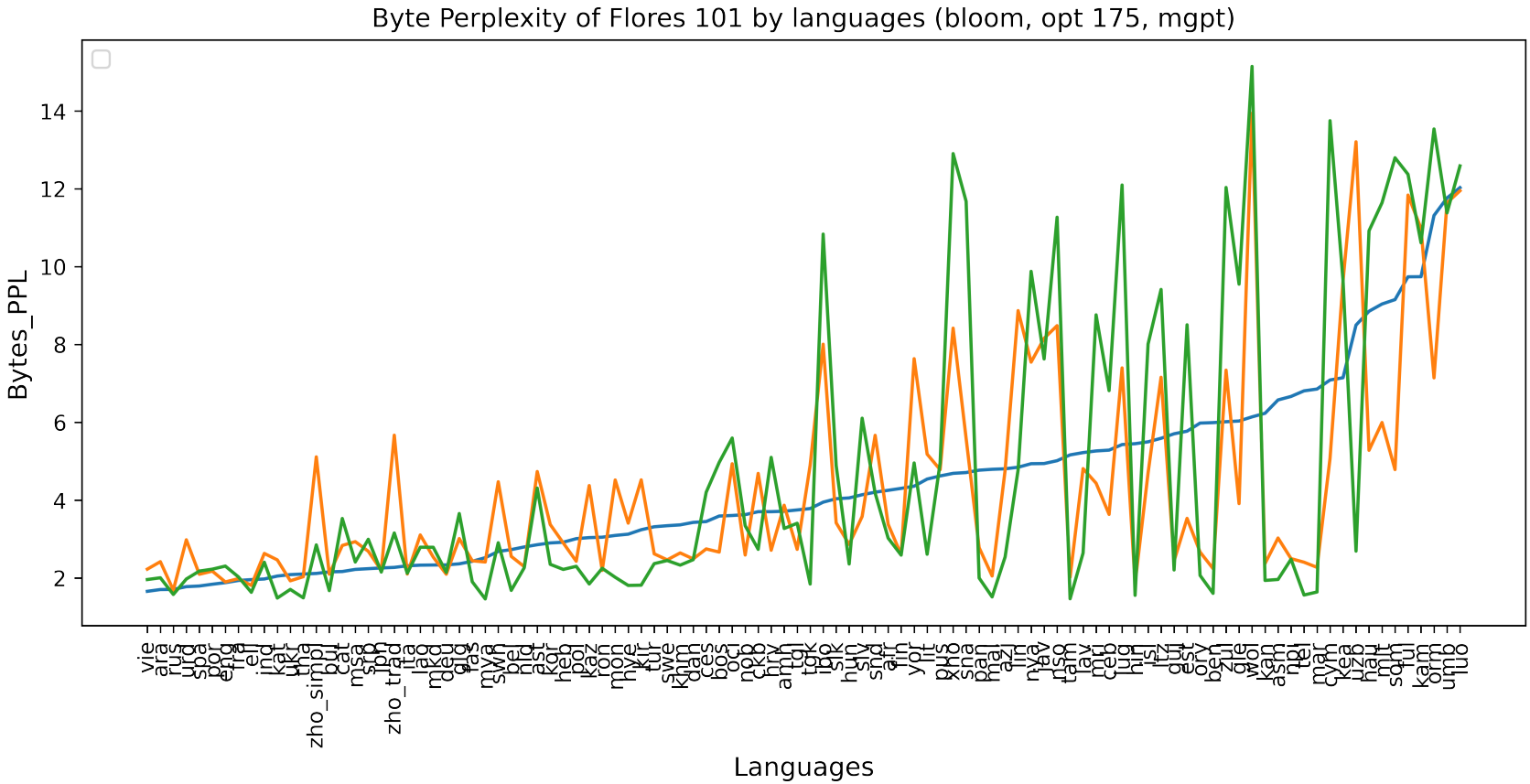


Normalizing per characters: improves PPL of languages with long words

The Perplexing Perplexity of Multilingual LMs

Questions that you cannot answer

How well is a language modeled in a multilingual LM?

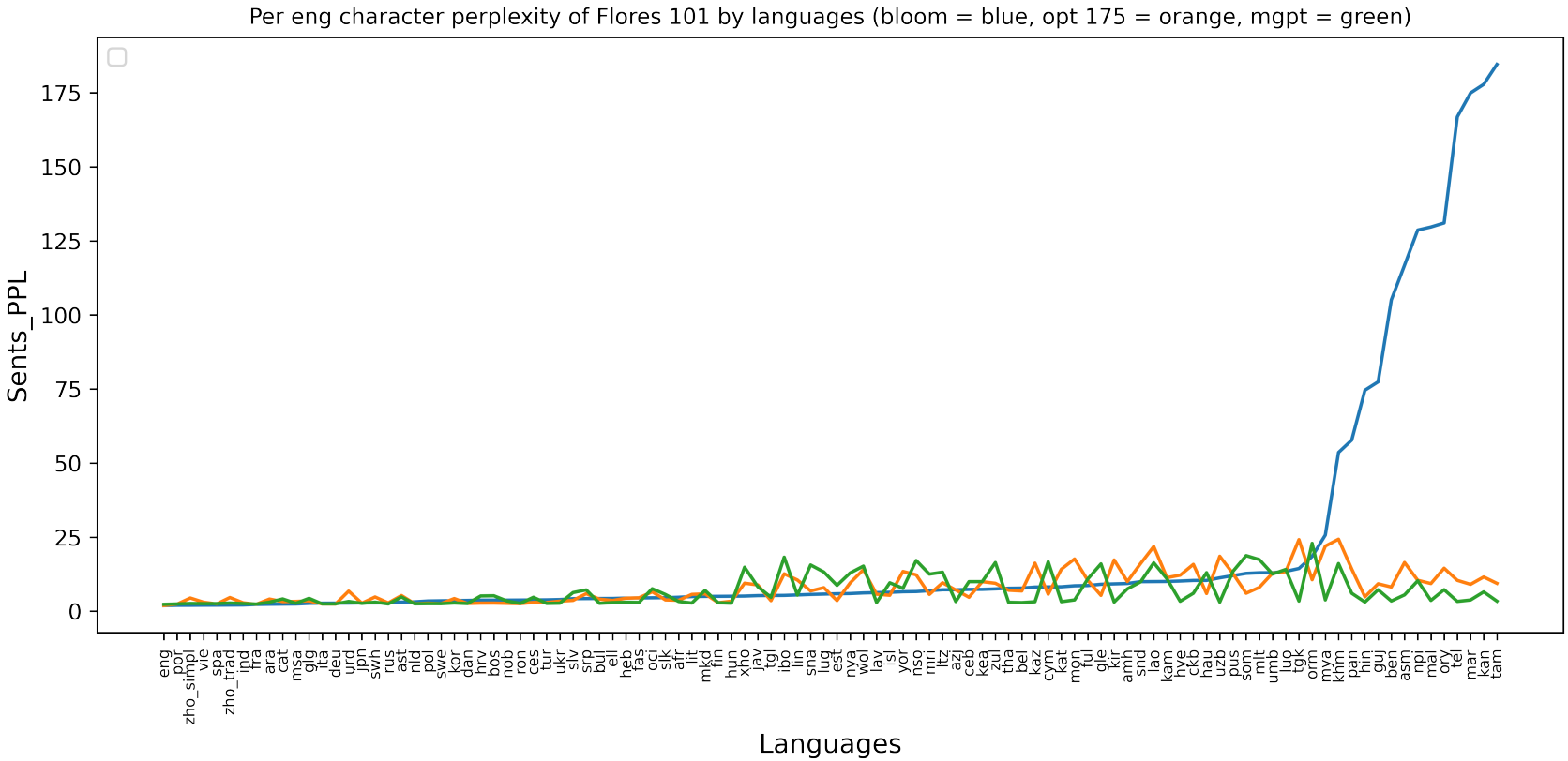


Normalizing per bytes: benefits languages using complex charsets

The Perplexing Perplexity of Multilingual LMs

Questions that you cannot answer

How well is a language modeled in a multilingual LM?

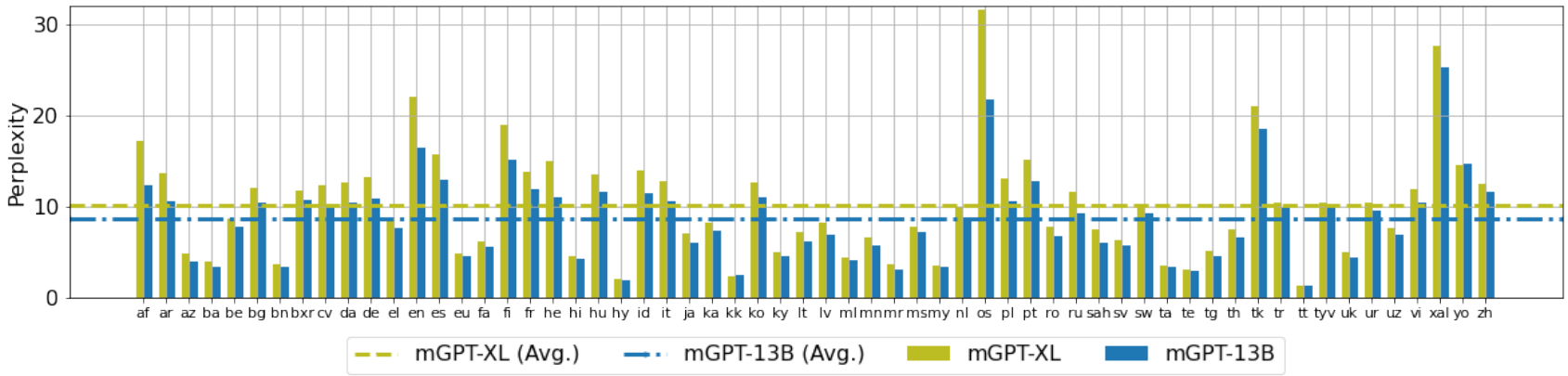


Normalizing per English characters: using less characters helps

The Perplexing Perplexity of Multilingual LMs

Questions that you cannot answer

An evaluation should not be trusted (from mGPT paper)



Strange things happen when test texts are not controlled across languages

Matching Parallel Sentences

Before alignement

In the gayest and happiest spirits she set forward with her father; not always listening , but always agreeing to what he said. They arrived. It is Frank and Miss Fairfax , said Mrs. Weston. I was just going to tell you of our agreeable surprize in seeing him arrive this morning. He stays till to-morrow , and Miss Fairfax has been persuaded to spend the day with us.	Elle partit avec son père, le visage souriant; elle n'écoutait pas toujours, mais elle acquiesçait de confiance. Ils arrivèrent. – C' est Frank et Mlle Fairfax, dit aussitôt Mme Weston. – J'allai justement vous faire part de l'agréable surprise que nous avons eue en le voyant arriver. Il reste jusqu'à demain et Mlle Fairfax a bien voulu, sur notre demande, venir passer la journée .

☞ Jane Austen, Emma https://www.janeausten.org/emma/chapter-54.asp

Matching Parallel Sentences

After alignment

In the gayest and happiest spirits she set forward with her father;	Elle partit avec son père, le visage souriant;
not always listening, but always agreeing to what he said;	elle n' écoutait pas toujours, mais elle acquiesçait de confiance.
They arrived .	Ils arrivèrent .
It is Frank and Miss Fairfax, said Mrs. Weston .	– C'est Frank et Mlle Fairfax, dit aussitôt Mme Weston .
I was just going to tell you of our agreeable surprize in seeing him arrive this morning.	– J'allai justement vous faire part de l'agréable surprise que nous avons eue en le voyant arriver.
He stays till tomorrow, and Miss Fairfax has been persuaded to spend the day with us .	Il reste jusqu'à demain et Mlle Fairfax a bien voulu, sur notre demande , venir passer la journée.

sentence alignment is easy in parallel documents, challenging in comparable corpora

Matching Parallel Sentences

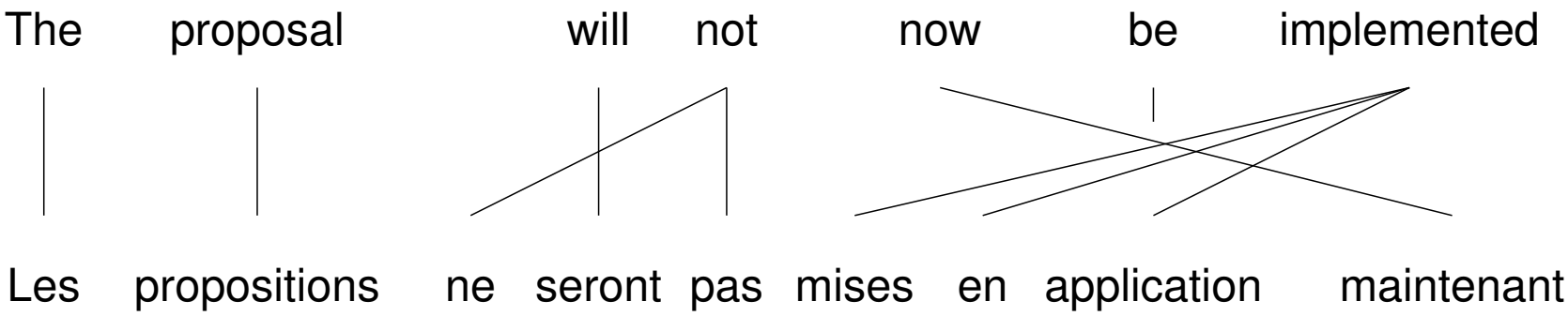
Pretrained multilingual sentence representations are useful

- parallel sentence mining in large multilingual corpora
- towards nearest neighbor machine translation

Better multilingual representations improve multilingual sentence retrieval

Word Alignments and Multilingual Embeddings

Computing word alignments



A vexing and ill-posed problem

Word Alignments and Multilingual Embeddings

Alignments from embeddings [Jalili Sabet et al., 2020]

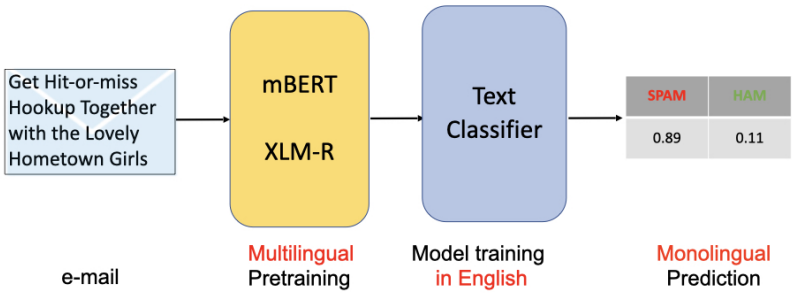
1. evaluate every possible link (e,f) with $\text{sim} \propto \text{mRep}(e)^T \, \text{mRep}(f)^T$
2. compute maximum weight matching in the resulting bipartite graphs

The better the alignment, the better the multilingual space (*)

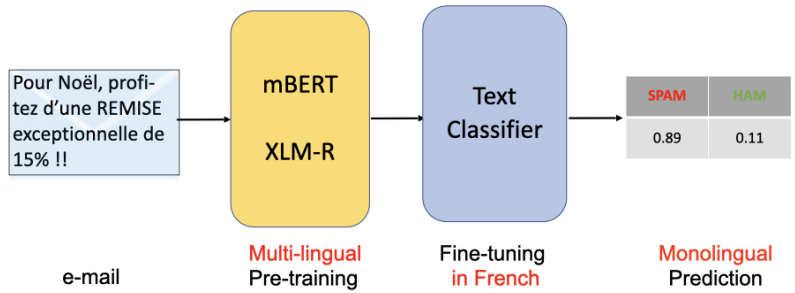
(*) caveat: this requires reference alignments, a rare resource

Evaluating Representations through Cross-Lingual Learning

We adopt the zero-shot cross-lingual transfer setting, where we (1) fine-tune the pre-trained model on English and (2) directly transfer the model to target languages. [Conneau et al., 2020]



English spam filter



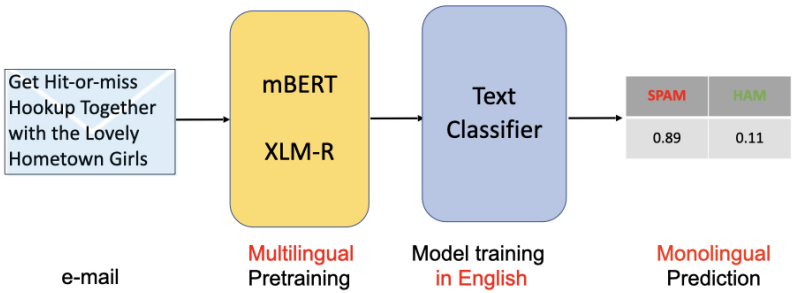
French spam filter

In "zero-shot" mode, no language adaptation is even performed

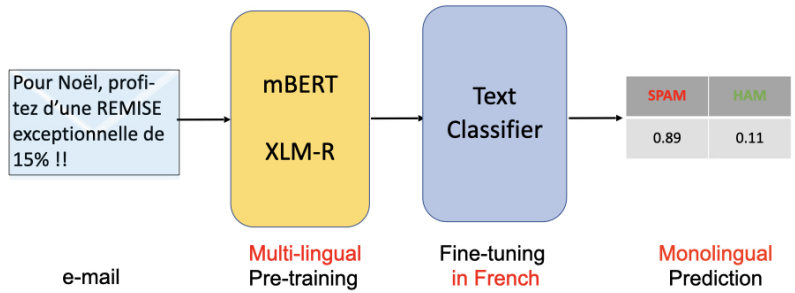
Large benchmarks: X-GLUE [Liang et al., 2020] X-TREME [Hu et al., 2020]

Evaluating Representations through Cross-Lingual Learning

We adopt the zero-shot cross-lingual transfer setting, where we (1) fine-tune the pre-trained model on English and (2) directly transfer the model to target languages. [Conneau et al., 2020]



English spam filter



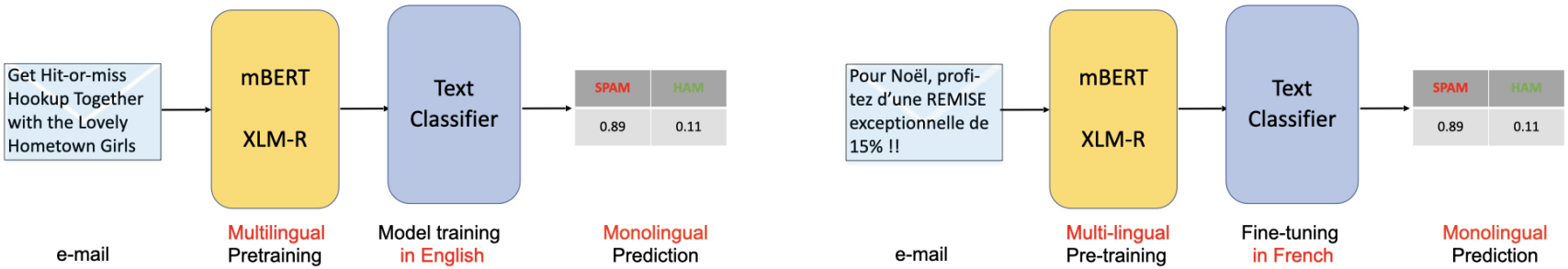
French spam filter

In "zero-shot" mode, no language adaptation is even performed

Large benchmarks: X-GLUE [Liang et al., 2020] X-TREME [Hu et al., 2020]

Evaluating Representations through Cross-Lingual Learning

We adopt the zero-shot cross-lingual transfer setting, where we (1) fine-tune the pre-trained model on English and (2) directly transfer the model to target languages. [Conneau et al., 2020]



English spam filter

French spam filter

In "zero-shot" mode, no language adaptation is even performed

Large benchmarks: X-GLUE [Liang et al., 2020] X-TREME [Hu et al., 2020]

Evaluating mLMs with Machine Translation

Instructions for Machine Translation

Translate into French "By the end of the year, we will have seven new pharmacists.":

D'ici la fin de l'année, nous aurons sept nouveaux pharmaciens.

Model	En-Fr	Fr-En
GPT-2 [1,5b]	5	11.5
GPT-3 [175b]	21.2	25.2
PALM [540B]	38.5	41.1
SOTA	45.6	45.4

Test representations (in source) and generation (in targert), closing the gap with well trained bilingual MT

Lessons and Challenges in Evaluating Multilingual LMs

mLMs are powerful tools and help improve SOTA for many tasks

Open questions

- the apories of counting the languages of a mLM
- how to achieve fairness in mLMs design ?
- the reason why transfert is positive or negative
- mLMs: a collection of monolingual models ?
- the generating code-switched languages

Bibliography I

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Jennifer C. Lai, and Robert L. Mercer. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40, 1992. URL https://aclanthology.org/J92-1002.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Bibliography II

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.280. URL https://aclanthology.org/2021.naacl-main.280.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL https://aclanthology.org/2022.tacl-1.30.

Bibliography III

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hu20b.html.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.147. URL https://www.aclweb.org/anthology/2020.findings-emnlp.147.

Bibliography IV

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2022. URL https://arxiv.org/abs/2211.09110.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*, 2020.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Bibliography V

Clara Meister and Ryan Cotterell. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.414. URL https://aclanthology.org/2021.acl-long.414.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1491. URL https://aclanthology.org/P19-1491.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.3. URL https://aclanthology.org/2021.emnlp-main.3.

Bibliography VI

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*, 2019.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL https://aclanthology.org/2020.tacl-1.25.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2019.