Wonders and Mysteries of Multilingual Language Models

François Yvon

ISIR — CNRS and Sorbonne Université

Workshop DataIA - ILLS @ CentraleSupélec

Multilingual Natural Language Processing

Computational Language Documentation

(Bayesian) Probabilistic Models

S. Okabe, P. Godard, L. Besacier

Retrieval Augmented MT

Non Autoregressive MT (LevT), imitation learning

J. Grego, J. Xu, M. Bouthors

Multidomain MT

Adapters, curriculum / sampling policy learning

P.M. Quang, J. Crego

Interacting with bitexts

Alignment & edit-based models

P. Cubaud, J. Crego, A.K Ngo-Ho, J. Xu

MT for Open Science

Integrating Terminology & Phraseology in NMT

S. Abdul-Rauf, R. Bawden, N. Kübler, etc

Multilingual LLMs

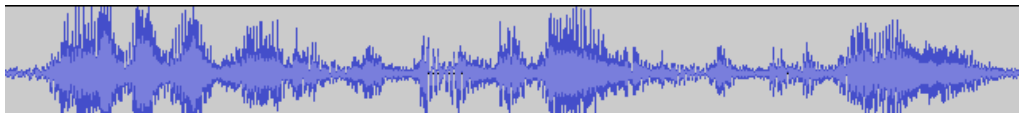
X-lingual Transfer, Probing & Metrics

G. Wisniewski, H. Schütze, A. Imani, M. Sabet, etc.

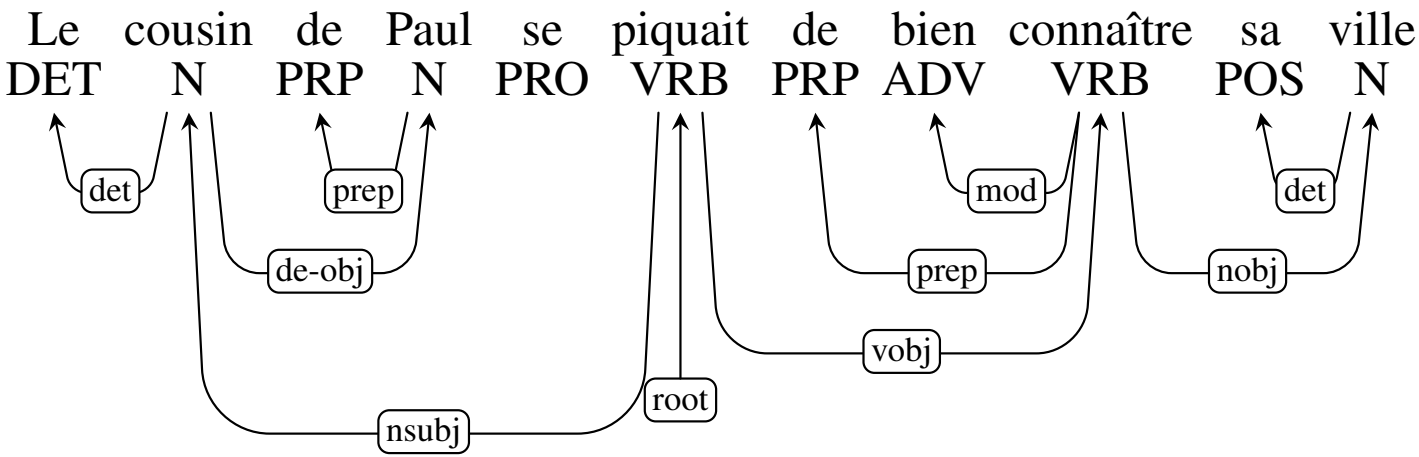
☞ details `fyvo.github.io`

Natural Language Processing : the Data-Based Revolution

Issues in Linguistic Analysis



lə ‖kuzɛ̃‖dø‖ pol ‖ sə ‖ pikɛ ‖ də ‖ bjɛ̃ ‖ ø ːdə ø ːdə ‖ bjɛ̃ ‖ konɛtrə ‖ sa ‖ vil

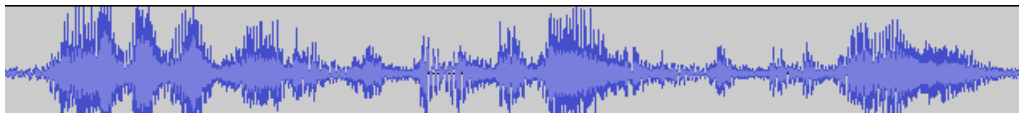


(sp / se-piquer-de-1
: ARG0 (c / cousin)
: ARG1 (k / connaitre-2
: ARG0 (c)
: ARG1 (v /ville))
)

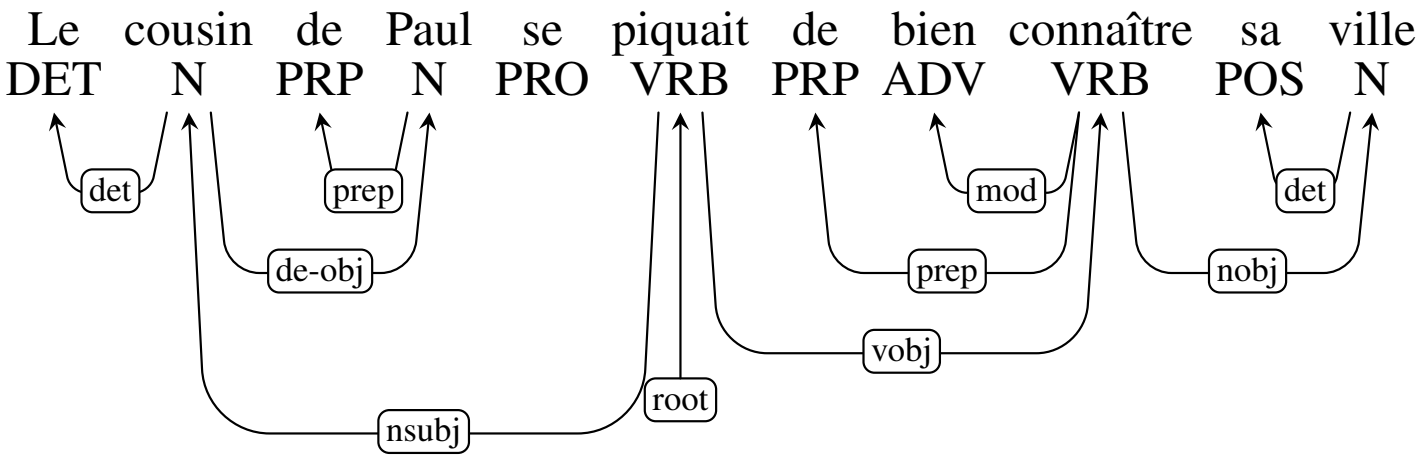
Cousin ?

Natural Language Processing : the Data-Based Revolution

Issues in Linguistic Analysis



lə ‖kuzɛ̃‖dø‖ pol ‖ sə ‖ pikɛ ‖ də ‖ bjɛ̃ ‖ ø ːdə ø ːdə ‖ bjɛ̃ ‖ konɛtrə ‖ sa ‖ vil



(sp / se-piquer-de-1
: ARG0 (c / cousin)
: ARG1 (k / connaitre-2
: ARG0 (c)
: ARG1 (v /ville))
)

Cousin ?

Natural Language Processing : the Data-Based Revolution

Downstream Applications

- **text analysis** : information retrieval and classification, question-answering, spell checking, information extraction, NLU, etc
- **text generation** : machine translation, summarization, dialog, TTS, etc

Natural Language Processing : the Data-Based Revolution

Processing Sequences of Symbols with Hidden Structure

Input	$\boldsymbol{w} = w_1 \ldots w_T \in \Sigma_1^T$
	w_t : *a phoneme, a letter, a morpheme, a word, a sentence*
Output	$\boldsymbol{y} \in \Sigma_2$: *sequence classification*
	y : *filter, topic, polarity, language, author*
Output	$\boldsymbol{y} \in \Sigma_2^T$: *sequence labelling*
	y_t : *PoS, feature, bracketing, etc*
Output	$\boldsymbol{y} \in \Sigma_2^*$: *sequence transduction*
	$\boldsymbol{y}$: *transcription, translation, answer, correction, summary, etc*
Output	$\boldsymbol{y} \in \mathcal{G}$: *sequence parsing*
	$\boldsymbol{y}$: *morphological or syntactic tree, semantic graph, discourse tree*

- Ambiguous or unobserved boundaries
- Ubiquitous ambiguity of basic units w_t
- Structural constraints over possible $\boldsymbol{x}, \boldsymbol{y}$ (unobserved)
- Internal structure within Σ_1 and Σ_2 (unobserved)
- Masive variabilty of data distribution (domain, genre, style, register)

Natural Language Processing : the Data-Based Revolution

The "Unreasonable Effectiveness" of Supervised Learning in NLP (1993-)

There is nothing but tabular data

sent_id = fr_partut-ud-184
text = *La sécurité des transports a trop fait l'actualité ces derniers temps* :

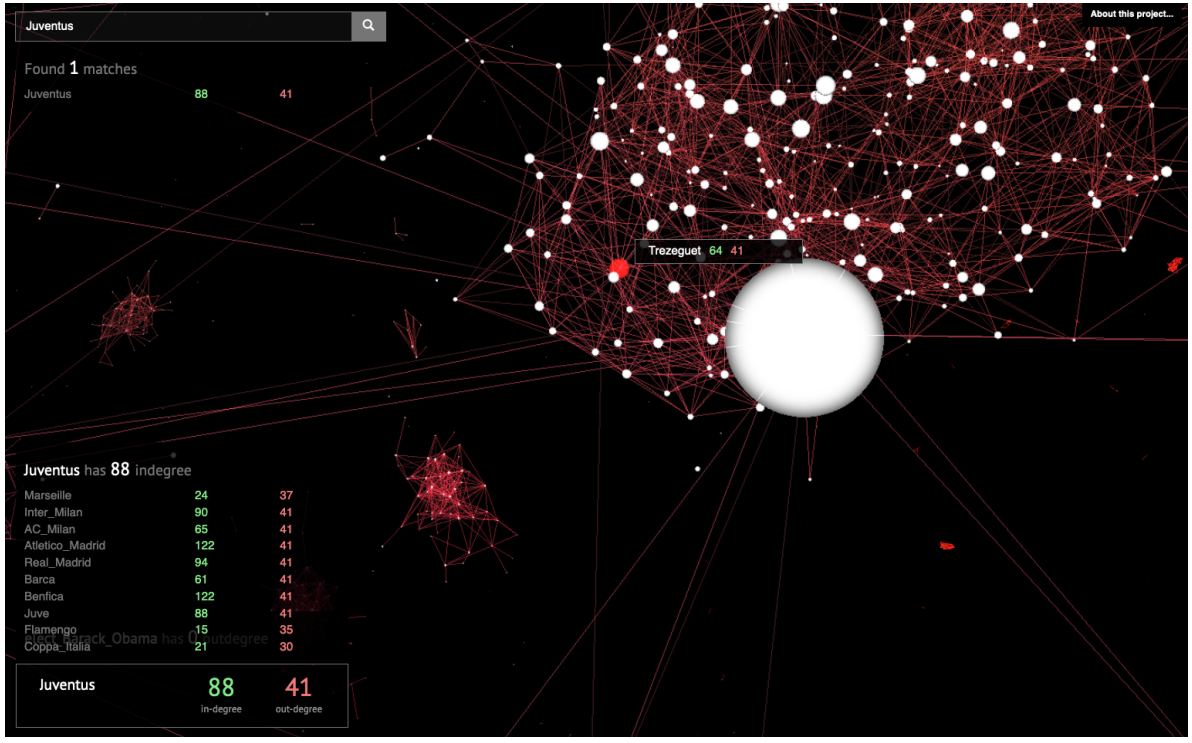
1	La	le	DET	RD	Def=Def|Gdr=Fem|Num=Sing|PronType=Art
2	sécurité	sécurité	NOUN	S	Gdr=Fem|Num=Sing
3-4	des	_	_	_	_
3	de	de	ADP	E	_
4	les	le	DET	RD	Def=Def|Num=Plur|PronType=Art
5	transports	transport	NOUN	S	Gdr=Masc|Num=Plur
6	a	avoir	AUX	VA	Mood=Ind|Num=Sing|Person=3|Tense=Pres
7	trop	trop	ADV	B	
8	fait	faire	VERB	V	Gdr=Masc|Num=Sing|Tense=Past|VbForm=Part
9	l'	le	DET	RD	Def=Def|Num=Sing|PronType=Art
10	actualité	actualité	NOUN	S	Gdr=Fem|Num=Sing
11	ces	ce	DET	DD	Num=Plur|PronType=Dem
12	derniers	dernier	ADJ	NO	Gdr=Masc|Num=Plur|NumType=Ord
13	temps	temps	NOUN	S	Gdr=Masc
14	:	:	PUNCT	FC	_

Picking "low-hanging fruits" with simple learners (classifiers, sequence learning)

Monolingual Pretrained Language Models are Powerful

Modern Lexical Representations : Word Embeddings

Map words w_t in context c onto vector-space representations x as $x = f_{\boldsymbol{\theta}}(w_t; c)$.
$f_{\boldsymbol{\theta}}()$ is a neural net eg. Transformer [Vaswani et al., 2017] with parameters $\boldsymbol{\theta}$.



The structure of word spaces : related words are close

Monolingual Pretrained Language Models are Powerful

Learning θ

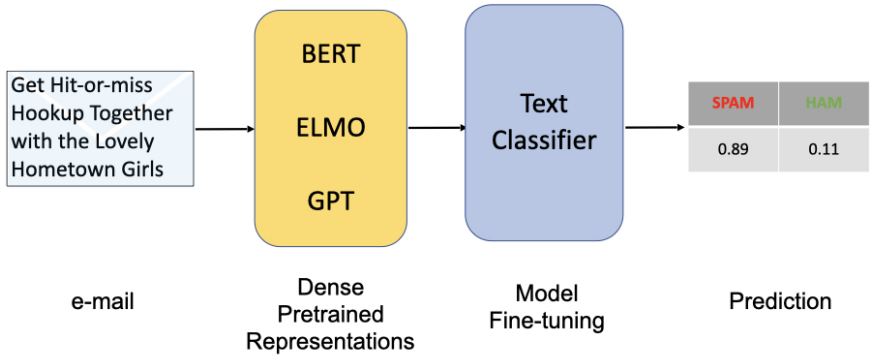
Gigantic collections of texts, simple auxiliary tasks with natural annotations

1. Predict next word given prefix : pure decoder
   Longtemps je me suis couché [mask] ... $\Rightarrow$ train $P_{\theta}(w_t|w_{<t})$
   GPT-2 [Radford et al., 2019], OPT [Zhang et al., 2022], GPT-J [Wang and Komatsuzaki, 2021], etc
2. Predict missing word given bidirectional context : pure encoder
   Longtemps je me suis couché de [mask] bonne heure. $\Rightarrow$ train $P_{\theta}(w_t|w_{-t})$
   BERT [Devlin et al., 2019], Roberta [Liu et al., 2019], CamemBERT [Martin et al., 2020] etc.
3. Denoising sequence to sequence : encoder-decoder
   Longtemts je couché suis de bnone || Longtemps je me suis couché de bonne heure.
   $\Rightarrow$ train $P_{\theta}(w|\tilde{w})$
   BART [Lewis et al., 2020], T5 [Raffel et al., 2020], etc.

Monolingual Pretrained Language Models are Powerful

Using θ

1. Learn + fine-tune task-adapted model $h_{\phi,\theta}(w;c) = h_{\phi}(f_{\theta}(w;c))$



2. With causal LMs, multi-purpose text generation via prompting

GEN Of course. In Chorukor, Monday is ilopagar, Tuesday ilopager, . . .
Wednesday ilopagur, Thursday ilopagir ...

Q&A Answer this : What are the birth date and place of Ludvík Vaculík ? . . .
23 July 1926, in Brumov, Moravia

SA "This Czech writer has written some the most wonderful French novels."
is a positive comment ? . . . [Yes | No]

Monolingual Pretrained Language Models are Powerful

Using θ

1. Learn + fine-tune task-adapted model $h_{\phi,\theta}(w; c) = h_{\phi}(f_{\theta}(w; c))$
2. With causal LMs, multi-purpose text generation via prompting

☺ LM pre-training is mostly unsupervised

☺ Instructions can be learned or fine-tuned

☺ Extremely successful for "high resource" languages

☹ Annotation, corpus and model size matter

Monolingual Pretrained Language Models are Powerful

Using $\boldsymbol{\theta}$

1. Learn + fine-tune task-adapted model $h_{\boldsymbol{\phi},\boldsymbol{\theta}}(w;c) = h_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(w;c))$
2. With causal LMs, multi-purpose text generation via prompting

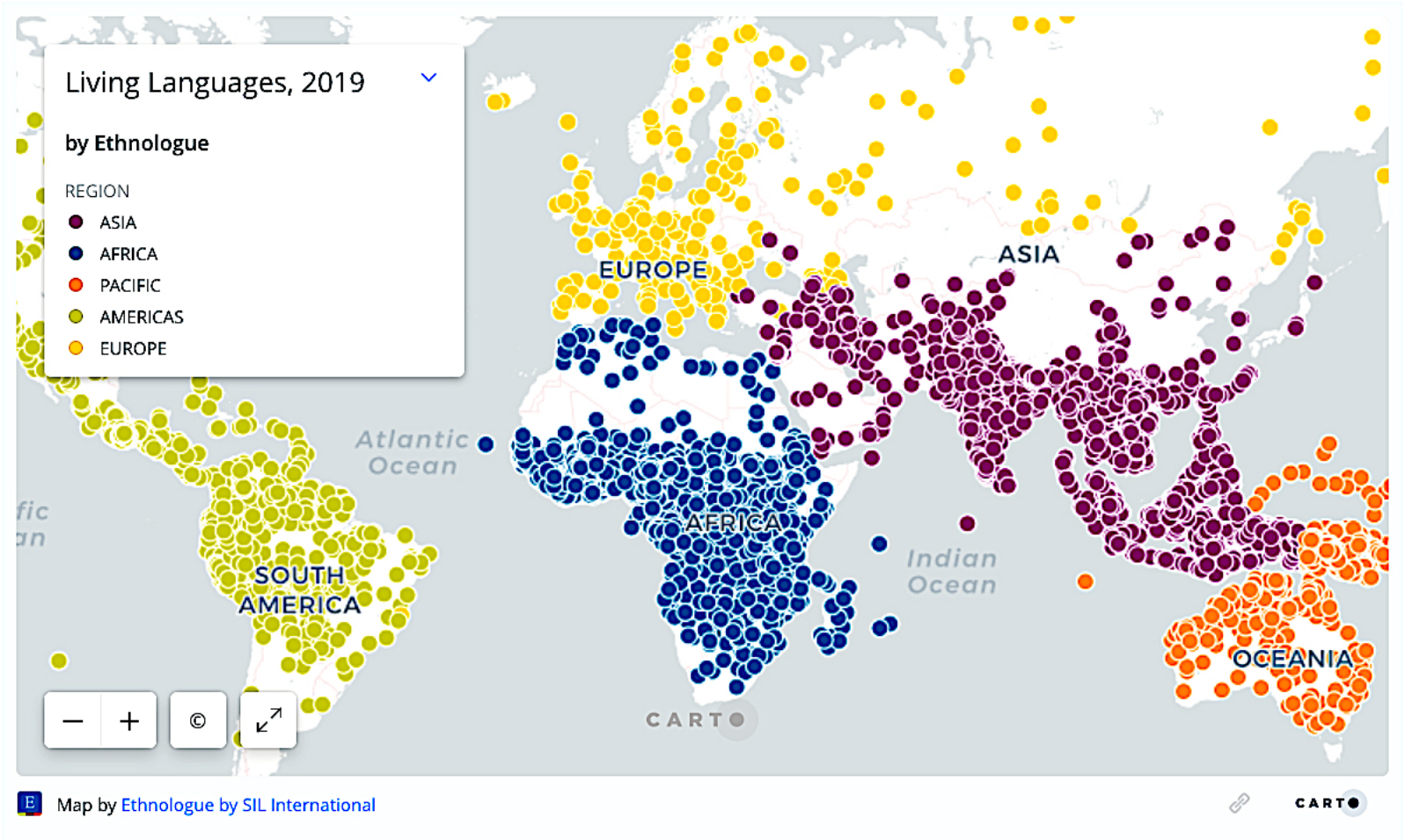
☺ LM pre-training is mostly unsupervised

☺ Instructions can be learned or fine-tuned

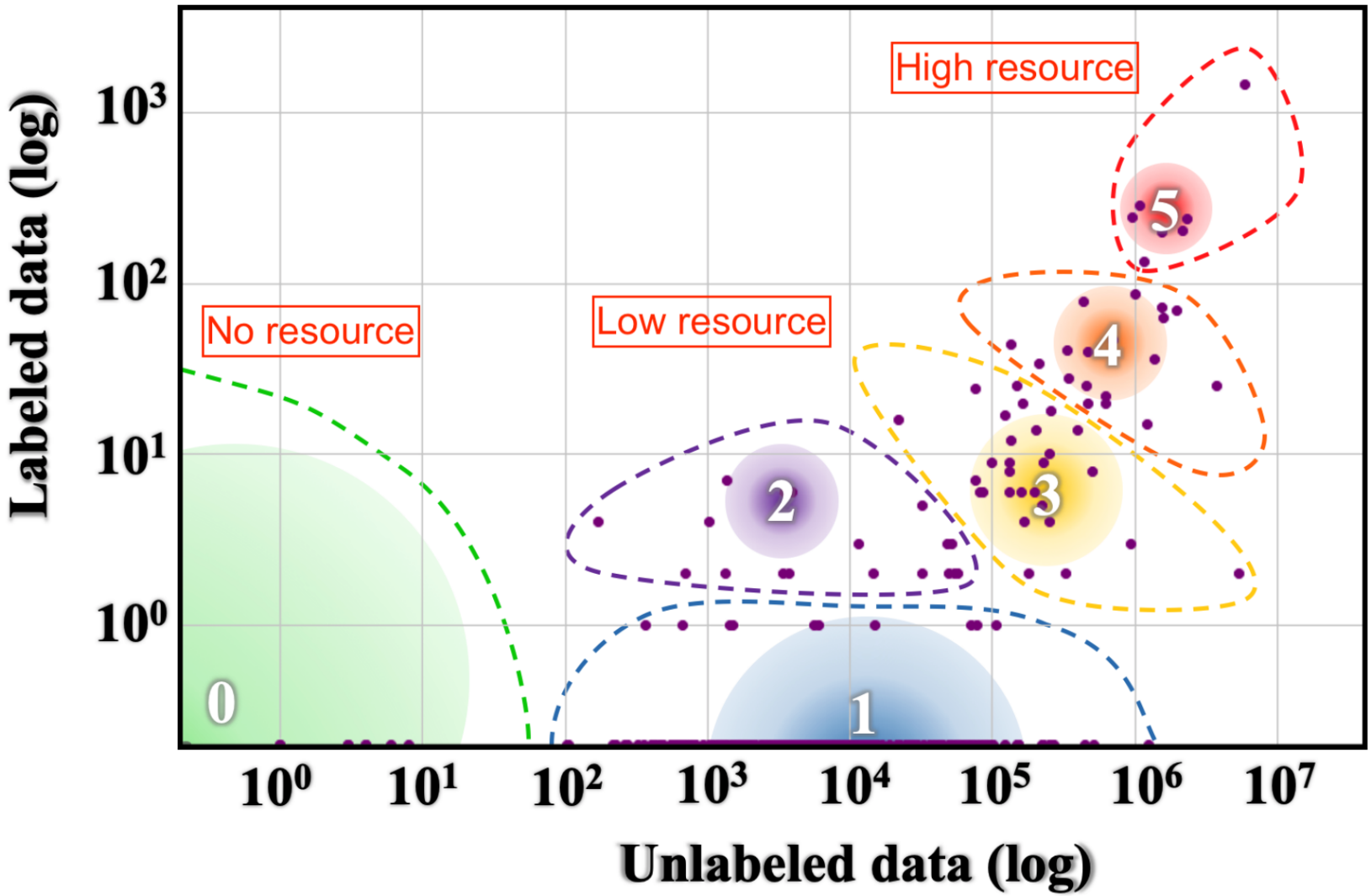
☺ Extremely successful for "high resource" languages

☹ Annotation, corpus and model size matter

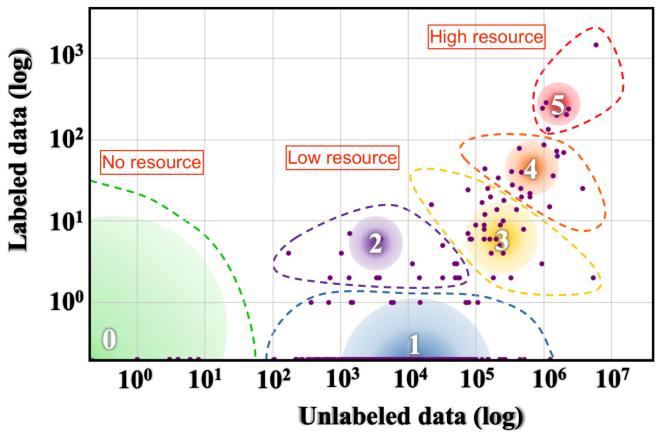
So Many Languages, most of them “Low-Resource”



So Many Languages, most of them “Low-Resource”



So Many Languages, most of them “Low-Resource”



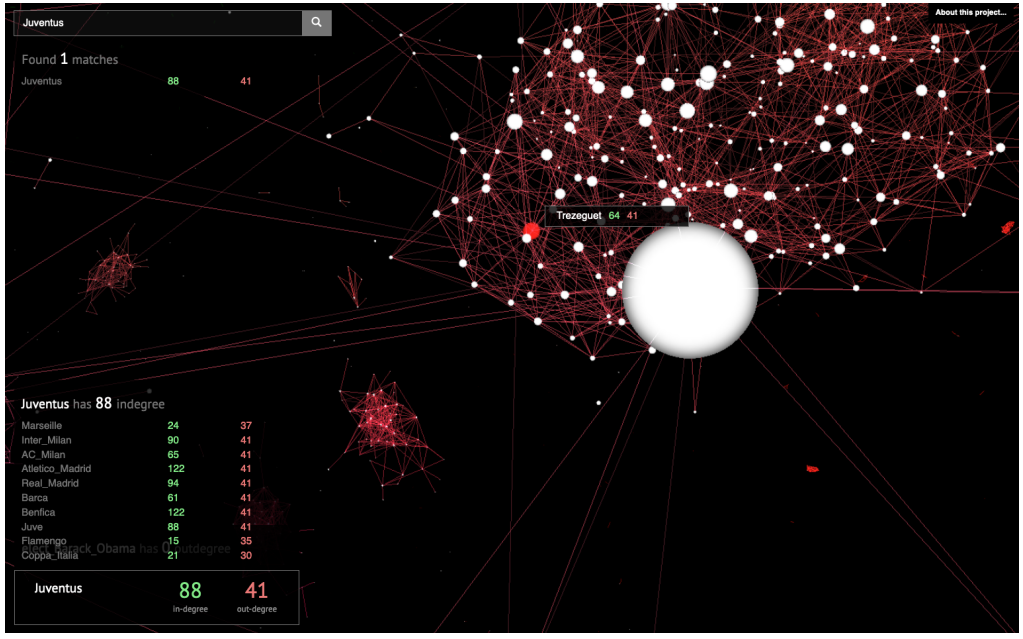
Cls	Example languages	#Lang	#Spkr	% Lang
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07 %
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Data and analysis from [Joshi et al., 2020]

Multilingual Pretrained Language Models - mPLMs

Multilingual Lexical Representations : Word Embeddings

Map words w onto vector-space representations x in context c as $x = f_{\boldsymbol{\theta}}(w; c)$
$f_{\boldsymbol{\theta}}()$ a complex neural net with parameters $\boldsymbol{\theta}$.



Neighbours should be semantically similar **cross-linguistically**

Multilingual Pretrained Language Models - mPLMs

Multilingual representations of texts

_Tous _les _être s _humain s _na issent _libre s _et _ég aux _en _digni té _et _en _droits . _Ils _sont _do u és _de _raison _et _de _conscience _et _doivent _agir _les _uns _en vers _les _autres _dans _un _ esprit _de _frater n ité .
_Všichni _lidé _rod í _se _svobod ní _a _sobě _rov ní _co _do _d ů stoj nosti _a _práv . _Jsou _na dán i _rozum em _a _s vědomí m _a _mají _spolu _jedna t _v _du chu _brat r ství .
_Tutti _gli _esse ri _umani _na scono _liberi _ed _e gu ali _in _digni tà _e _diritti . _Es si _sono _do tati _di _ragione _e _di _coscienza _e _devono _agir e _gli _uni _verso _gli _altri _in _spirito _di _fra tella nza .

Low-level, language independent segmentations

Multilingual Pretrained Language Models - mPLMs

Multilingual representations of texts

_Tous _les _être s _humain s _na issent _libre s _et _ég aux _en _digni té _et _en _droits . _Ils _sont _do u és _de _raison _et _de _conscience _et _doivent _agir _les _uns _en vers _les _autres _dans _un _ esprit _de _frater n ité .
_Všichni _lidé _rod í _se _svobod ní _a _sobě _rov ní _co _do _d ů stoj nosti _a _práv . _Jsou _na dán i _rozum em _a _s vědomí m _a _mají _spolu _jedna t _v _du chu _brat r ství .
_Tutti _gli _esse ri _umani _na scono _liberi _ed _e gu ali _in _digni tà _e _diritti . _Es si _sono _do tati _di _ragione _e _di _coscienza _e _devono _agir e _gli _uni _verso _gli _altri _in _spirito _di _fra tella nza .

Low-level, language independent segmentations

Multilingual Pretrained Language Models - mPLMs

Learning multilingual θ

Multilingual Mixtures of corpora, simple auxiliary tasks with natural annotations :

1. Predict next word given prefix : pure decoder.
   mGPT [Shliazhko et al., 2022], XGLM [Lin et al., 2021], etc
2. Predict missing word given full context : pure encoder.
   mBERT [Devlin et al., 2019], XLM-R [Conneau et al., 2020]
3. Denoising sequence to sequence : encoder-decoder
   mBART50 [Liu et al., 2020b], mT5 [Xue et al., 2020], etc

+ Complementary objectives to bridge between languages :
parallel corpora, bilingual dictionaries, synthetic code-switched data [Conneau et al., 2020, Chi et al., 2021], etc.

Multilingual Pretrained Language Models - mPLMs

Using multilingual θ

1. Learn / fine-tune task-adapted model $h_{\phi,\theta}(w;c) = h_{\phi}(f_{\theta}(w;c))$ on L_1, perform zero-shot X-lingual transfer to L_2
2. Multilingual text generation with prompting

Translate into English

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

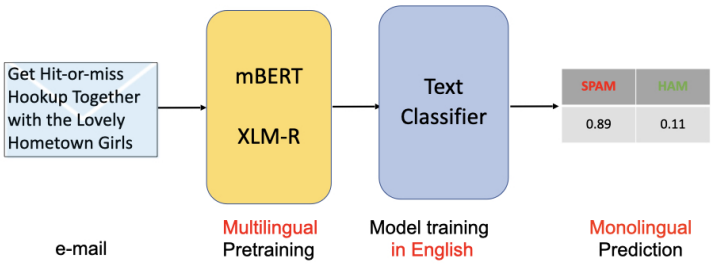
D'ici la fin de l'année, nous aurons sept nouveaux pharmaciens.

☺ Hardly more difficult than monolingual

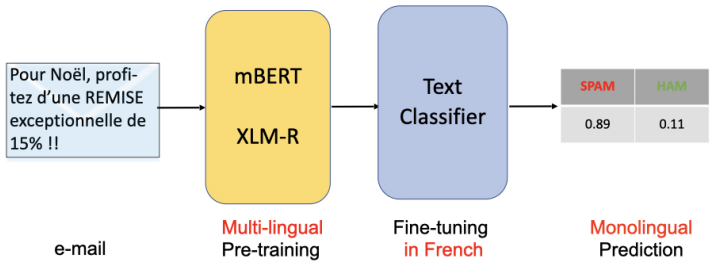
☺ Bring low resource languages up to steam

The effectiveness of multilingual PLMs

One model handling multiple languages



English spam filter

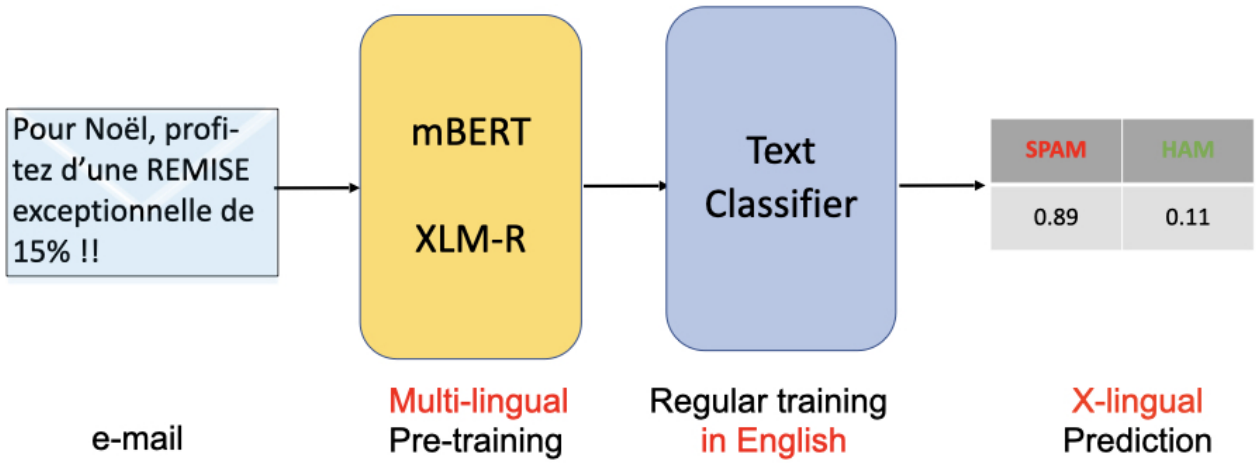


French spam filter

Effective for many languages / tasks [Pires et al., 2019, Wu and Dredze, 2020]

The effectiveness of multilingual PLMs

Knowledge transfer between languages

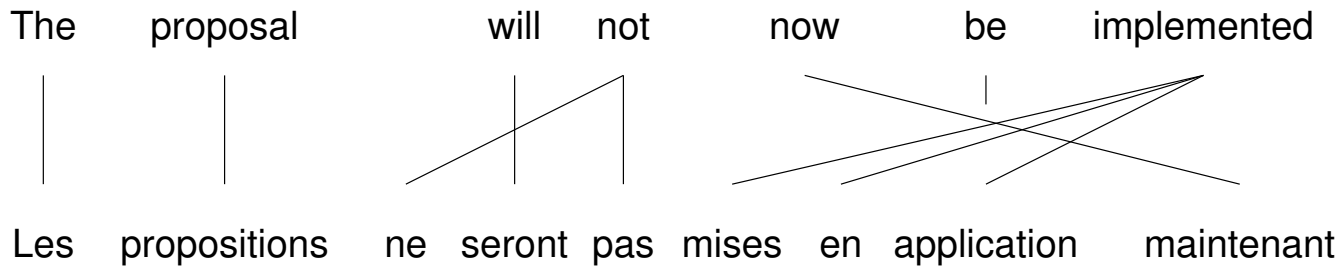


Multilingual zero-shot spam filter

- ☺ No foreign training data required
- ☺ Effective for many languages and tasks : X-GLUE [Liang et al., 2020]
  X-TREME+ [Hu et al., 2020, Ruder et al., 2023]

The effectiveness of multilingual PLMs

Multilingual Embeddings Align Lexical Representations



- Word Alignment = Bipartite Graph
- Multilingual embeddings + graph algorithms = unsupervised alignments
- Aligns code-switching and multilingual texts
- Also : unsupervised sentence alignments

	ENG-CES		ENG-DEU		ENG-FAS	
Method	F_1	AER	F_1	AER	F_1	AER
IBM2	.76	.25	.71	.29	.57	.43
IBM4	.75	.26	.77	.23	.51	.49
eflomal	.85	.15	.77	.23	.61	.39
fastText	.70	.30	.60	.40	.50	.50
mBERT[8]	**.87**	**.13**	**.79**	**.21**	.67	.33
XLM-R[8]	**.87**	**.13**	**.79**	**.21**	**.70**	**.30**

Word alignment, Argmax method

Jalili Sabet et al. [2020]

The effectiveness of multilingual PLMs

Unsupervised Alignments in Large Scale Evaluation

Glot-500 [ImaniGooghari et al., 2023] : a multilingual (XML-type) model for 500 languages

- New challenges
  - data selection and filtering
  - language identification at scale
- New evaluation methods
  - Round trip alignment
  - Sentence retrieval

Language-Script	XLM-R-B	XLM-R-L	Glot500-m	Language-Script	XLM-R-B	XLM-R-L	Glot500-m	Language-Script	XLM-R-B	XLM-R-L	Glot500-m
efi_Latn	2.55	3.25	**6.23**	mfe_Latn	3.61	4.19	**6.26**	toi_Latn	3.19	4.10	**4.31**
ell_Grek	2.79	3.38	**4.77**	mgh_Latn	2.78	3.28	**3.48**	toj_Latn	1.43	1.84	**2.25**
eng_Latn	4.02	4.49	**6.39**	mgr_Latn	3.32	4.06	**6.39**	ton_Latn	2.01	2.64	**3.63**
enm_Latn	3.77	4.60	**7.19**	mhr_Cyrl	2.75	3.28	**5.32**	top_Latn	1.56	2.16	**2.19**
epo_Latn	4.01	4.83	**5.88**	min_Latn	2.62	3.05	**3.78**	tpi_Latn	2.44	2.71	**5.96**
est_Latn	4.34	5.24	**8.21**	miq_Latn	2.23	3.13	**4.12**	tpm_Latn	2.79	3.39	**4.67**
eus_Latn	3.12	3.80	**4.19**	mkd_Cyrl	3.99	4.54	**7.37**	tsn_Latn	2.82	3.12	**4.63**
ewe_Latn	2.22	2.67	**4.74**	mlg_Latn	3.34	3.81	**6.33**	tso_Latn	2.40	3.05	**5.00**
fao_Latn	3.85	4.62	**5.75**	mlt_Latn	2.94	3.57	**4.87**	tsz_Latn	2.68	3.14	**4.20**
fas_Arab	4.54	4.48	**7.00**	mos_Latn	2.71	3.24	**4.25**	tuc_Latn	1.43	1.83	**2.36**
fij_Latn	2.81	3.17	**4.94**	mps_Latn	1.50	1.65	**3.05**	tui_Latn	2.47	2.83	**4.53**
fil_Latn	3.26	3.92	**4.80**	mri_Latn	2.81	3.44	**5.49**	tuk_Cyrl	2.74	3.68	**4.33**
fin_Latn	4.06	5.19	**6.03**	mrw_Latn	2.69	3.24	**4.58**	tuk_Latn	2.43	3.23	**4.74**
fon_Latn	1.63	1.89	**3.70**	msa_Latn	3.17	3.50	**5.38**	tum_Latn	3.41	4.13	**6.15**
fra_Latn	3.19	3.97	**5.08**	mwm_Latn	1.74	1.99	**3.20**	tur_Latn	5.18	4.86	**7.45**
fry_Latn	3.36	3.99	**4.52**	mxv_Latn	1.75	2.11	**2.31**	twi_Latn	3.05	4.06	**6.70**
gaa_Latn	2.74	3.26	**6.01**	mya_Mymr	1.54	1.53	**2.46**	tyv_Cyrl	2.31	2.83	**3.33**
gil_Latn	2.76	3.20	**4.50**	myv_Cyrl	2.90	3.42	**4.46**	tzh_Latn	2.16	2.50	**3.08**
giz_Latn	3.00	3.43	**5.40**	mzh_Latn	2.62	3.02	**4.10**	tzo_Latn	2.01	2.29	**2.77**
gkn_Latn	1.93	2.07	**3.31**	nan_Latn	1.99	2.51	**2.56**	udm_Cyrl	2.90	3.48	**4.72**
gkp_Latn	1.88	2.25	**3.40**	naq_Latn	2.42	3.15	**4.41**	uig_Arab	2.58	3.11	**3.61**
gla_Latn	2.90	3.48	**3.61**	nav_Latn	1.75	2.10	**2.71**	uig_Latn	2.26	2.76	**3.79**
gle_Latn	3.52	4.24	**4.49**	nbl_Latn	3.09	3.87	**4.85**	ukr_Cyrl	5.71	5.96	**7.47**
glv_Latn	2.76	3.38	**4.45**	nch_Latn	2.18	2.74	**3.32**	urd_Arab	1.88	2.88	**3.96**
gom_Latn	3.05	3.59	**4.40**	ncj_Latn	2.64	3.40	**3.69**	urd_Latn	2.29	2.97	**3.03**
gor_Latn	2.26	2.73	**3.71**	ndc_Latn	3.32	3.85	**6.67**	uzb_Cyrl	2.73	3.26	**7.24**
grc_Grek	1.11	2.00	**2.93**	nde_Latn	4.00	4.60	**6.05**	uzb_Latn	3.32	3.98	**5.91**
guc_Latn	1.46	1.80	**2.23**	ndo_Latn	3.21	3.85	**5.61**	uzn_Cyrl	2.61	3.06	**5.86**
gug_Latn	2.60	3.23	**4.70**	nds_Latn	2.98	3.69	**4.70**	ven_Latn	2.96	3.64	**5.34**
guj_Gujr	3.18	4.15	**4.38**	nep_Deva	3.02	2.97	**6.31**	vie_Latn	3.99	4.48	**6.69**
gur_Latn	2.14	2.59	**3.22**	ngu_Latn	1.86	2.34	**3.39**	wal_Latn	2.87	3.65	**4.24**
guw_Latn	2.18	2.54	**4.56**	nia_Latn	2.75	**3.47**	3.24	war_Latn	3.04	3.74	**5.43**
gya_Latn	1.94	2.25	**4.63**	nld_Latn	2.81	3.63	**4.90**	wbm_Latn	2.44	2.86	**6.53**
gym_Latn	1.44	1.78	**2.63**	nmf_Latn	3.30	4.27	**5.05**	wol_Latn	3.47	4.48	**6.10**
hat_Latn	3.21	3.64	**6.39**	nnb_Latn	2.46	3.14	**4.08**	xav_Latn	0.87	1.03	**1.12**
hau_Latn	3.69	4.24	**6.31**	nno_Latn	3.90	4.61	**7.41**	xho_Latn	3.61	4.27	**5.90**
haw_Latn	2.25	2.63	**3.55**	nob_Latn	3.88	4.81	**5.83**	yan_Latn	2.95	3.35	**5.59**
heb_Hebr	1.85	2.41	**3.92**	nor_Latn	3.31	4.14	**5.82**	yao_Latn	2.01	2.66	**3.87**
hif_Latn	2.90	3.43	**3.60**	npi_Deva	3.29	3.30	**5.93**	yap_Latn	2.86	3.41	**3.45**
hil_Latn	2.92	3.48	**4.88**	nse_Latn	3.29	4.06	**5.74**	yom_Latn	3.25	4.00	**5.17**
hin_Deva	3.39	3.80	**5.13**	nso_Latn	3.06	3.92	**5.51**	yor_Latn	2.24	2.68	**3.88**
hin_Latn	2.94	3.20	**4.77**	nya_Latn	2.76	**3.19**	**5.96**	yua_Latn	2.04	2.26	**2.86**
hmo_Latn	2.43	2.70	**6.12**	nyn_Latn	2.77	3.50	**5.59**	yue_Hani	2.37	**3.19**	2.95
hne_Deva	2.48	2.53	**4.95**	nyy_Latn	2.21	2.74	**2.95**	zai_Latn	3.22	3.76	**5.21**
hnj_Latn	2.14	2.53	**4.28**	nzi_Latn	2.09	2.70	**4.20**	zho_Hani	2.77	4.38	**5.03**
hra_Latn	3.32	3.86	**5.19**	ori_Orya	2.73	2.77	**3.92**	zlm_Latn	4.39	5.15	**7.54**
hrv_Latn	4.14	5.24	**7.02**	ory_Orya	3.27	3.20	**4.39**	zom_Latn	3.65	4.45	**5.36**
hui_Latn	1.84	2.10	**3.47**	oss_Cyrl	2.20	2.52	**5.85**	zsm_Latn	4.49	5.07	**8.83**
hun_Latn	4.54	4.10	**5.62**	ote_Latn	1.89	2.23	**2.66**	zul_Latn	3.67	4.39	**5.44**
hus_Latn	1.70	2.00	**2.42**	pag_Latn	2.93	3.44	**4.56**				

Table 22: Accuracy of XLM-R-B, XLM-R-L, and Glot500-m on Round Trip Alignment (Part II).

Round Trip alignments, Part II

From 100^2 to 500^2 unsupervised word alignments

The effectiveness of multilingual PLMs

Multilingual Text Generation

If the French says : Ça n'arrête pas, nous sommes bien placés pour le savoir...

then the English should say : It doesn't stop, we know that well...

Model	En-Fr	Fr-En
GPT-2 [1,5b] [Radford et al., 2019]	5	11.5
GPT-3 [175b] [Brown et al., 2020]	21.2	25.2
PALM [540B] [Chowdhery et al., 2022]	38.5	41.1
BLOOM [176B] [BigScience et al., 2022]	??	??
SOTA [Liu et al., 2020a]	44.1	-

Closing the gap with well trained bilingual MT, really ?

The effectiveness of multilingual PLMs

Understanding MT performance of BLOOM

- Methods :
  - Multiple prompts, in multiple languages
  - 3 datasets and tasks
  - Dozens of language pairs (H-H, L-H, L-L)
  - Multiple metrics
- Main takeaways
  - a call for clarity when reporting results
  - the failures of 0-shot learning, “language hallucinations”
  - "Free-rider effects” and language similarities
  - poor MT quality for LR languages

Src ↓	Trg →	ar	en	es	fr	zh
ar	BLOOM	–	40.3	23.3	33.1	17.7
	M2M	–	25.5	16.7	25.7	13.1
en	BLOOM	28.2	–	29.4	45.0	26.7
	M2M	17.9	–	25.6	42.0	19.3
es	BLOOM	18.8	32.7	–	24.8	20.9
	M2M	12.1	25.1	–	29.3	14.9
fr	BLOOM	23.4	45.6	27.5	–	23.2
	M2M	15.4	37.2	25.6	–	17.6
zh	BLOOM	15.0	30.5	20.5	26.0	–
	M2M	11.6	20.9	16.9	24.3	–

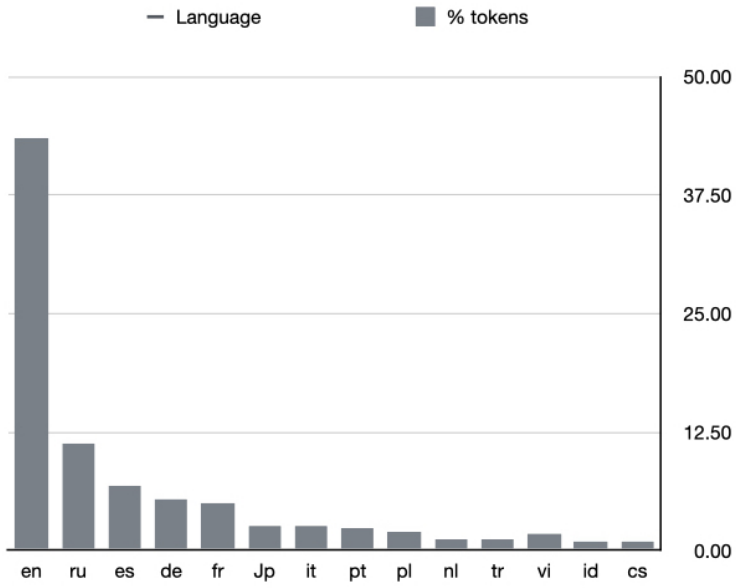
(a) High-resource language pairs.

Src ↓	Trg →	en	fr	hi	id	vi
en	BLOOM	–	45.0	27.2	39.0	28.5
	M2M	–	42.0	28.1	37.3	35.1
fr	BLOOM	45.6	–	18.5	31.4	32.8
	M2M	37.2	–	22.9	29.1	30.3
hi	BLOOM	35.1	27.6	–	–	–
	M2M	27.9	25.9	–	–	–
id	BLOOM	43.2	30.4	–	–	–
	M2M	33.7	30.8	–	–	–
vi	BLOOM	38.7	26.8	–	–	–
	M2M	29.5	25.8	–	–	–

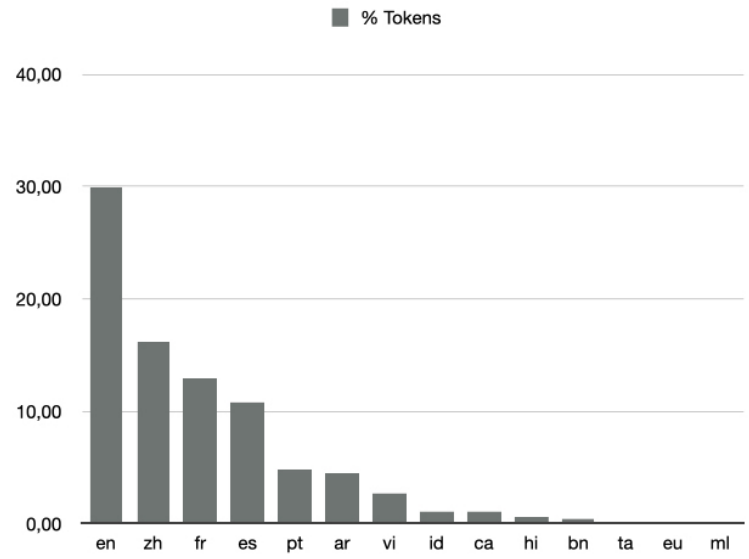
(b) High→mid-resource language pairs.

The effectiveness of multilingual PLMs

Caveats



Languages in mT5
(104 languages)



Languages in Bloom
(59 languages)

- ☹ English centric
- ☹ Poorly documented, hard to reproduce – what is a language anyway ?

What we would like to know

Why can this work ?

- Lexical overlap not necessary

  L1 : Longtemps je me suis couché de bonne heure .
  L2 : Mpohufnqt kf nf tvjt dpvdiê ef cpoof ifvsf !

- Transfer breaks with reversal or words shuffle

  L1 :Longtemps je me suis couché de bonne heure .
  L2 :. heure bonne de douché suis me je longtemps

- Vocabulary alignment helps
- Language typological similarity helps

Analyses in [K et al., 2020, Dufter and Schütze, 2020, Deshpande et al., 2022, ImaniGooghari et al., 2023], etc

What we would like to know

Why can this work ?

- Lexical overlap not necessary

  L1 : Longtemps je me suis couché de bonne heure .
  L2 : Mpohufnqt kf nf tvjt dpvdiê ef cpoof ifvsf !

- Transfer breaks with reversal or words shuffle

  L1 :Longtemps je me suis couché de bonne heure .
  L2 :. heure bonne de douché suis me je longtemps

- Vocabulary alignment helps
- Language typological similarity helps

Analyses in [K et al., 2020, Dufter and Schütze, 2020, Deshpande et al., 2022, ImaniGooghari et al., 2023], etc

? Impact of language distributions
? Impact of model size
? Impact of number of languages - curse of multilinguality
? Which linguistic properties help / break X-lingual transfer
? Metrics for “coverage”

What we would like to know

Can mPLMs handle truely multilingual texts ?

? Generate code-switched output

> Et le premier ministre nous répond que *a farmer is a farmer a Canadian is a Canadian* d'un bout à l'autre du Canada.
> Autrement dit *they are getting out of the closet* parce que cela leur donne le droit d avoir deux enfants.

(Exemples from [Carpuat, 2014])

? Generate text with multilingual prompts

Přeložit do angličtiny : Ça n'arrête pas, nous sommes bien placés pour le savoir... ; :

It doesn't stop, we know that well...

? Answer questions X-linguistically

? Generate summaries from multilingual sources

Take Aways

Large mPLMs serve practical purposes

They learn linguistic features

They display amazing emerging properties

Multilingual models also useful in MT, parsing, etc

They improve technological support for many languages [Ruder et al., 2023]

They remain poorly understood

Bibliography I

Workshop BigScience, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor,

Bibliography II

Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly

Bibliography III

Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller,

Bibliography IV

Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. BLOOM : A 176b-parameter open-access multilingual language model. CoRR, abs/2211.05100, 2022. doi : 10.48550/arXiv.2211.05100. URL https://doi.org/10.48550/arXiv.2211.05100.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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, 2020. URL https://arxiv.org/abs/2005.14165.

Marine Carpuat. Mixed language and code-switching in the Canadian Hansard. In Proceedings of the First Workshop on Computational Approaches to Code Switching, pages 107–115, Doha, Qatar, October 2014. Association for Computational Linguistics. doi : 10.3115/v1/W14-3913. URL https://aclanthology.org/W14-3913.

Bibliography V

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM : scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Bibliography VI

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. URL https://www.aclweb.org/anthology/2020.acl-main.747.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. When is BERT multilingual ? isolating crucial ingredients for cross-lingual transfer. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, pages 3610–3623, Seattle, United States, July 2022. Association for Computational Linguistics. doi : 10.18653/v1/2022.naacl-main.264. URL https://aclanthology.org/2022.naacl-main.264.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi : 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Bibliography VII

Philipp Dufter and Hinrich Schütze. Identifying elements essential for BERT's multilinguality. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4423–4437, Online, November 2020. Association for Computational Linguistics. doi : 10.18653/v1/2020.emnlp-main.358. URL https://aclanthology.org/2020.emnlp-main.358.

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.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André F. T. Martins, François Yvon, and Hinrich Schütze. Glot500 : Scaling Multilingual Corpora and Language Models to 500 Languages, 2023.

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 VIII

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282–6293, 2020.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert : An empirical study. In Proc. International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=HJeT3yrtDr.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi : 10.18653/v1/2020.acl-main.703. URL https://www.aclweb.org/anthology/2020.acl-main.703.

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.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual language models, 2021. URL https://arxiv.org/abs/2112.10668.

Bibliography IX

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. Very deep transformers for neural machine translation, 2020a. URL https://arxiv.org/abs/2008.07772.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa : A robustly optmized bert pretraining approach. arXiv preprint arXiv :1907.11692, 2019.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. arXiv preprint arXiv :2001.08210, 2020b.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT : a tasty French language model. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7203–7219, Online, July 2020. Association for Computational Linguistics. doi : 10.18653/v1/2020.acl-main.645. URL https://aclanthology.org/2020.acl-main.645.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT ? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi : 10.18653/v1/P19-1493. URL https://www.aclweb.org/anthology/P19-1493.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Bibliography X

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140) :1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A. Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David I. Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. Xtreme-up : A user-centric scarce-data benchmark for under-represented languages, 2023.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. mGPT : few-shot learners go multilingual, 2022. URL https://arxiv.org/abs/2204.07580.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Ben Wang and Aran Komatsuzaki. GPT-J-6B : A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, 2021.

Bibliography XI

Shijie Wu and Mark Dredze. Do explicit alignments robustly improve multilingual encoders ? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4471–4482, Online, November 2020. Association for Computational Linguistics. doi : 10.18653/v1/2020.emnlp-main.362. URL https://www.aclweb.org/anthology/2020.emnlp-main.362.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5 : A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv :2010.11934, 2020.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT : Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.