Scaling Multilingual Representations beyond 100 Languages

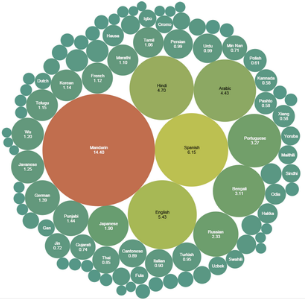
joint work with the NLLB team
Meta AI Research

NAACL - MIA workshop
July 16 2022

Context and Motivation

- 7 151 living languages

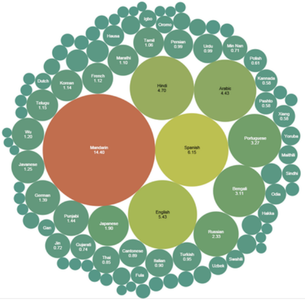
Native speakers



Context and Motivation

- 7 151 living languages
- 40% are endangered

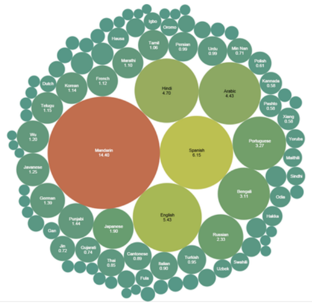
Native speakers



Context and Motivation

- 7 151 living languages
- 40% are endangered
- 23 languages account for half the population
- 200 languages ⇒ 88%
- ≈ 4 000 with developed writing system

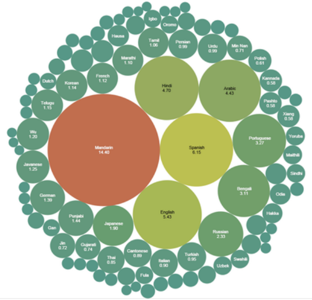
Native speakers



Context and Motivation

- 7 151 living languages
- 40% are endangered
- 23 languages account for half the population
- 200 languages ⇒ 88%
- ≈ 4 000 with developed writing system
- Multilingual approaches: ≈ 130 languages

Native speakers



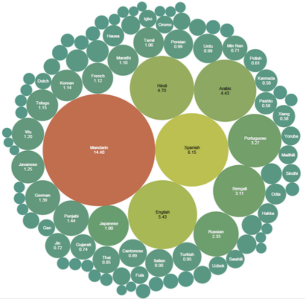
Scaling 100++

META AI

Context and Motivation

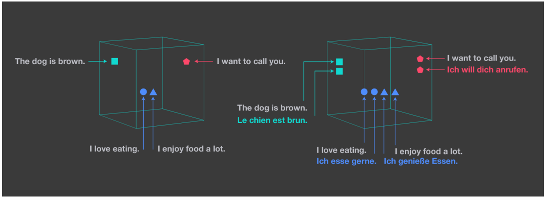
- 7 151 living languages
- 40% are endangered
- 23 languages account for half the population
- 200 languages ⇒ 88%
- ≈ 4 000 with developed writing system
- Multilingual approaches: ≈ 130 languages

Native speakers



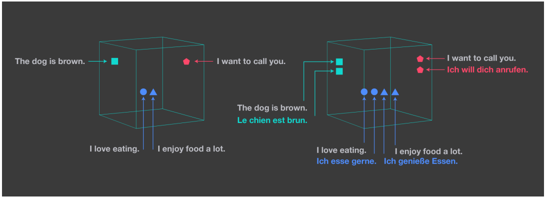
⇒ How can we scale well beyond 100 languages?

Multilingual Sentence Embeddings



- Sentences with similar meaning are close (paraphrases)
- Independently of the language they are written in

Multilingual Sentence Embeddings



- Sentences with similar meaning are close (paraphrases)
- Independently of the language they are written in

Popular approaches

- LASER, *Artexe and Schwenk, arXiv Dec'18, TACL'19*
- mBART, *Liu et al, arXiv'20*
- XLM-R, *Conneau et al, ACL'20*
- LaBSE, *Feng et al, arXiv'20*
- ...

Massively Multilingual Models

One-for-all approach

- NMT, sentence representations, ...
- Low-resource languages benefit from high-resource ones
  - e.g. Nepali/Hindi or Icelandic/German
- But accounting for the huge size difference is tricky
- Can new low-resource languages be efficiently learned

Massively Multilingual Models

One-for-all approach

- NMT, sentence representations, ...
- Low-resource languages benefit from high-resource ones
  - e.g. Nepali/Hindi or Icelandic/German
- But accounting for the huge size difference is tricky
- Can new low-resource languages be efficiently learned

$\Rightarrow$ *Curse of multilinguality*

Massively Multilingual Models

One-for-all approach

- NMT, sentence representations, ...
- Low-resource languages benefit from high-resource ones
  - e.g. Nepali/Hindi or Icelandic/German
- But accounting for the huge size difference is tricky
- Can new low-resource languages be efficiently learned

$\Rightarrow$ *Curse of multilinguality*

- Do we expect gains combining "unrelated languages"?
  - does Wolof benefit of Indonesian or Italian?
  - does Assamese benefit of Arabic or Albanian?

Massively Multilingual Models

One-for-all approach

- NMT, sentence representations, ...
- Low-resource languages benefit from high-resource ones
  - e.g. Nepali/Hindi or Icelandic/German
- But accounting for the huge size difference is tricky
- Can new low-resource languages be efficiently learned

$\Rightarrow$ *Curse of multilinguality*

- Do we expect gains combining "unrelated languages"?
  - does Wolof benefit of Indonesian or Italian?
  - does Assamese benefit of Arabic or Albanian?
- Some low-resource languages are rather isolated (Quechua, Inuit, ...)

Massively Multilingual Models

Switch to training multiple models

- Train models by groups of similar languages
- Ideally, each group contains a high-resource language

$\Rightarrow$ How can we make sure that these individual models are mutually compatible?

  - e.g. an African and Turkic language

Massively Multilingual Models

- Substantial improved LASER sentence embeddings

Massively Multilingual Models

- Substantial improved LASER sentence embeddings

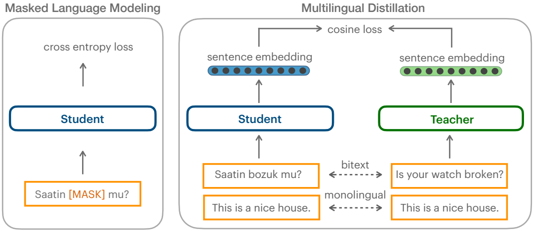
No Language Left Behind (NLLB)

- Single NMT system to translate among 200 languages
- Outperforms previous state-of-the-art by more than 40%

Teacher-Student Training

Idea

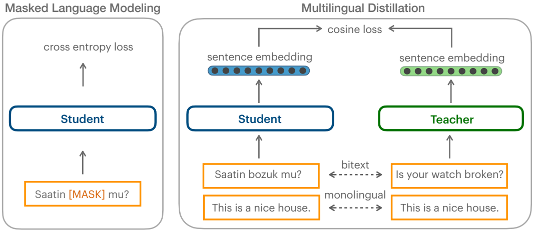
- Do not train new models from scratch (for new languages)
- Extend existing embedding space to more languages



Teacher-Student Training

Idea

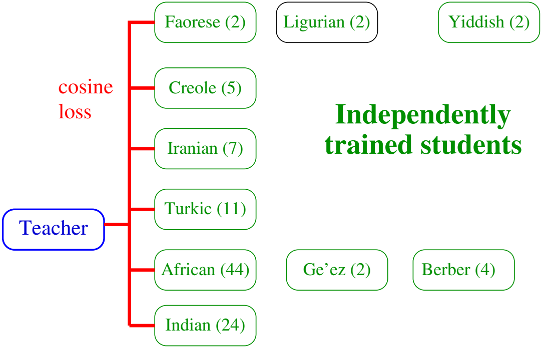
- Do not train new models from scratch (for new languages)
- Extend existing embedding space to more languages



Advantages

- Likely, less resources are needed
- Can be combined with masked LM training

Using Multiple Students



- Multiple students using the same teacher

$\Rightarrow$ The students are mutually compatible

- Each student can be separately optimized (architecture, capacity, vocabulary, convergence, ...)

Comparison with Reimers and Gurevych

Reimers and Gurevych, *Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*, EMNLP'20

	Reimers & Gurevych	**LASER3**
Teacher	SBERT (eng)	LASER (93 langs)
Student	single	multiple
Architecture	same	lang. specific
Initialization	XLM-R	random
Criterion	MSE	cosine
Train. data	xx-eng bitexts only	xx-eng bitexts eng-eng mono. eng-spa bitexts

- Unfortunately, we were not able to make a fair experimental comparison

Evaluation of Multilinguality

Scaling multilingual models

- We may find training data in >1000 languages (e.g. bible)
- But high-quality evaluation data is more limited
  - Tatoeba is very noisy and unbalanced

Evaluation of Multilinguality

Scaling multilingual models

- We may find training data in >1000 languages (e.g. bible)
- But high-quality evaluation data is more limited
  - Tatoeba is very noisy and unbalanced

FLORES

- FLORES-101: ≈1000 sentences in 101 languages
- N-way parallel, sampled from Wikipedia
- NLLB: extension to 204 languages:
  - mostly low-resource languages
  - freely available
- Recently extended to speech (FLEURS-101)

Evaluation of Multilinguality

Bitext mining

- Final goal: improve MT performance
- Costly: train encoder, mine bitexts, train SMT $\rightarrow$ BLEU

Evaluation of Multilinguality

Bitext mining

- Final goal: improve MT performance
- Costly: train encoder, mine bitexts, train SMT $\rightarrow$ BLEU

Proxy: multilingual similarity search `xsim`

- Given a parallel test data (FLORES)
- Search translation with highest margin score

$$\text{score}(x, y) = \frac{cos(x, y)}{\sum_{z \in NN_k(x)} \frac{cos(x,z)}{2k} + \sum_{v \in NN_k(y)} \frac{cos(y,v)}{2k}}$$

- `xsim`: error rate of wrongly matched sentences in FLORES
- Easy to use open-source implementation

Evaluation of LASER3

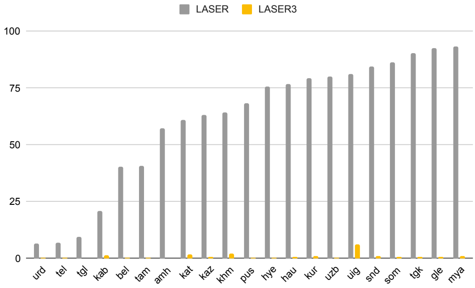
Methology

- Trained LASER3 models for 148 languages
- Transformers perform better than BiLSTM
- Select best model based on `xsim` on FLORES dev
- Mine bitexts against 21.5 billion English sentences
- Train NMT systems
- Compare BLEU on *"human"* versus *"human + mined"*

Evaluation of LASER3

Improving the original LASER

- Originial LASER performed badly on several languages



- Retrained models: avrg `xsim` 61→0.9%
  - Burmese: 93→0.9%, Irish 92→0.8%
  - on-pair with LaBSE

Malayo-Polynesian Languages

Lang.	bitexts	BLEU	xsim %	Monol.	Mined	BLEU
Acehnese	39.2k	0	2.4	2.2M	1.4M	10.3
Buginese	21.8k	0	1.6	0.7M	717k	4.2
Cebuano	1.1M	34.4	0.1	23.6M	8.1M	39.0
Indonesian	11M	-	0.1	-	-	-
Javanese	86k	11.1	0.1	27.2M	8.5M	31.2
Malay	2.3M	34.4	0.0	640M	40.5M	41.4
Pangasinan	327k	15.6	0.7	3.9M	1.9M	18.5
Sundanese	32.3k	1.5	0.6	8.2M	6.1M	28.5
Tagalog	1.3M	40.2	0.1	89M	33M	43.8
Warray	331k	26.5	0.2	26.9M	4.9M	36.5

Malayo-Polynesian Languages

Lang.	bitexts	BLEU	xsim %	Monol.	Mined	BLEU
Acehnese	39.2k	0	2.4	2.2M	1.4M	10.3
Buginese	21.8k	0	1.6	0.7M	717k	4.2
Cebuano	1.1M	34.4	0.1	23.6M	8.1M	39.0
Indonesian	11M	-	0.1	-	-	-
Javanese	86k	11.1	0.1	27.2M	8.5M	31.2
Malay	2.3M	34.4	0.0	640M	40.5M	41.4
Pangasinan	327k	15.6	0.7	3.9M	1.9M	18.5
Sundanese	32.3k	1.5	0.6	8.2M	6.1M	28.5
Tagalog	1.3M	40.2	0.1	89M	33M	43.8
Warray	331k	26.5	0.2	26.9M	4.9M	36.5

- Very low `xsim` error rates for most languages despite <100k bitexts for some languages

⇒ Training a language specific encoder seems to be beneficial

Malayo-Polynesian Languages

Lang.	bitexts	BLEU	xsim %	Monol.	Mined	BLEU
Acehnese	39.2k	0	2.4	2.2M	1.4M	10.3
Buginese	21.8k	0	1.6	0.7M	717k	4.2
Cebuano	1.1M	34.4	0.1	23.6M	8.1M	39.0
Indonesian	11M	-	0.1	-	-	-
Javanese	86k	11.1	0.1	27.2M	8.5M	31.2
Malay	2.3M	34.4	0.0	640M	40.5M	41.4
Pangasinan	327k	15.6	0.7	3.9M	1.9M	18.5
Sundanese	32.3k	1.5	0.6	8.2M	6.1M	28.5
Tagalog	1.3M	40.2	0.1	89M	33M	43.8
Warray	331k	26.5	0.2	26.9M	4.9M	36.5

- Large amounts of monolingual data

$\Rightarrow$ Optimal conditions for mining

Malayo-Polynesian Languages

Lang.	bitexts	BLEU	xsim %	Monol.	Mined	BLEU
Acehnese	39.2k	0	2.4	2.2M	1.4M	10.3
Buginese	21.8k	0	1.6	0.7M	717k	4.2
Cebuano	1.1M	34.4	0.1	23.6M	8.1M	39.0
Indonesian	11M	-	0.1	-	-	-
Javanese	86k	11.1	0.1	27.2M	8.5M	31.2
Malay	2.3M	34.4	0.0	640M	40.5M	41.4
Pangasinan	327k	15.6	0.7	3.9M	1.9M	18.5
Sundanese	32.3k	1.5	0.6	8.2M	6.1M	28.5
Tagalog	1.3M	40.2	0.1	89M	33M	43.8
Warray	331k	26.5	0.2	26.9M	4.9M	36.5

- BLEU gain >20: Javanese and Sundanese

Malayo-Polynesian Languages

Lang.	bitexts	BLEU	xsim %	Monol.	Mined	BLEU
Acehnese	39.2k	0	2.4	2.2M	1.4M	10.3
Buginese	21.8k	0	1.6	0.7M	717k	4.2
Cebuano	1.1M	34.4	0.1	23.6M	8.1M	39.0
Indonesian	11M	-	0.1	-	-	-
Javanese	86k	11.1	0.1	27.2M	8.5M	31.2
Malay	2.3M	34.4	0.0	640M	40.5M	41.4
Pangasinan	327k	15.6	0.7	3.9M	1.9M	18.5
Sundanese	32.3k	1.5	0.6	8.2M	6.1M	28.5
Tagalog	1.3M	40.2	0.1	89M	33M	43.8
Warray	331k	26.5	0.2	26.9M	4.9M	36.5

- BLEU gain >20: Javanese and Sundanese
- High resource languages also improve

European Minority Languages

Lang.	fao	fur	lij	lim	lmo	ltz	srd	szl	vec	ydd
Addtl. Lang	deu	ita	ita	nld	ita	deu	ita	pol	ita	deu
Bitexts [k]	6.6	6.3	2.2	5.4	1.3	9.8	1.4	6.4	1.2	6.2
BLEU	0	0	0	0	0	0	0	0	0	0
`xsim` **[%]**	2.57	0.1	0.2	16.1	1.09	0.59	0.1	0.69	2.77	0.1
Monolingual	1.2M	737k	106k	15M	61M	123M	515k	2.5M	12M	12M
Mined	1.6M	532k	631k	2.0M	4.1M	5.5M	723k	1.0M	2.5M	3.3M
BLEU	10.6	23.5	13.4	5.5	20.7	37.0	20.9	18.9	17.8	30.1

- Pairing low-resource with similar high-resource language is very effective
- BLEU > 20: Faroese, Lombard and Sardinian
- BLEU > 30: Luxemburgish and Yiddish

Creole Languages

Lang.	hat	kea	pap	sag	tpi
Addtl. Lang	fra	por	spa por	lin	eng
Bitexts	334	6	5	282	458
BLEU	20.2	0	0	4.8	14.7
`xsim` **[%]**	1.19	1.19	0.1	8.6	0.2
Monolingual	14M	227k	28M	645k	1.7M
Mined	8.0M	656k	7.3M	1.9M	1.2M
BLEU	29.2	4.9	40.9	5.3	16.1

- Papiemento: mono=28M $\rightarrow$ BLEU=40.9
- Tok Pisin: mono=1.7M $\rightarrow$ BLEU=16.1
- Kabuverdianu: mono<300k $\rightarrow$ BLEU=4.9

Creole Languages

Lang.	hat	kea	pap	sag	tpi
Addtl. Lang	fra	por	spa por	lin	eng
Bitexts	334	6	5	282	458
BLEU	20.2	0	0	4.8	14.7
`xsim` **[%]**	1.19	1.19	0.1	8.6	0.2
Monolingual	14M	227k	28M	645k	1.7M
Mined	8.0M	656k	7.3M	1.9M	1.2M
BLEU	29.2	4.9	40.9	5.3	16.1

- Papiemento: mono=28M $\rightarrow$ BLEU=40.9
- Tok Pisin: mono=1.7M $\rightarrow$ BLEU=16.1
- Kabuverdianu: mono$<$300k $\rightarrow$ BLEU=4.9

$\Rightarrow$ The amount of monolingual data is crucial

Berber Languages (14M speakers)

Lang. **Script**	Kabyle Latin	Tifinagh Latin	Tifinagh Tifinagh	Tamazight Tifinagh
bitexts	72k	10.2k	4k	6.2k
BLEU	1.2	0	0	0
`xsim` [%]	0.99	24.11	35.57	3.66
Monolingual	3.4M	23k	5k	59k
Mined	3.1M	240k	-	111k
BLEU	6.2	1.2	-	3.8

- Extremely limited resources, except Kabyle
- Kabyle: some mined bitexts and BLEU>6
- Tamazight: very modest BLEU score of ≈ 4
- Tifinagh: insufficient monolingual data

Berber Languages (14M speakers)

Lang. **Script**	Kabyle Latin	Tifinagh Latin	Tifinagh Tifinagh	Tamazight Tifinagh
bitexts	72k	10.2k	4k	6.2k
BLEU	1.2	0	0	0
`xsim` [%]	0.99	24.11	35.57	3.66
Monolingual	3.4M	23k	5k	59k
Mined	3.1M	240k	-	111k
BLEU	6.2	1.2	-	3.8

- Extremely limited resources, except Kabyle
- Kabyle: some mined bitexts and BLEU>6
- Tamazight: very modest BLEU score of ≈ 4
- Tifinagh: insufficient monolingual data

$\Rightarrow$ Typical examples of very low-resource languages for which it is very hard to collect written material

African Languages

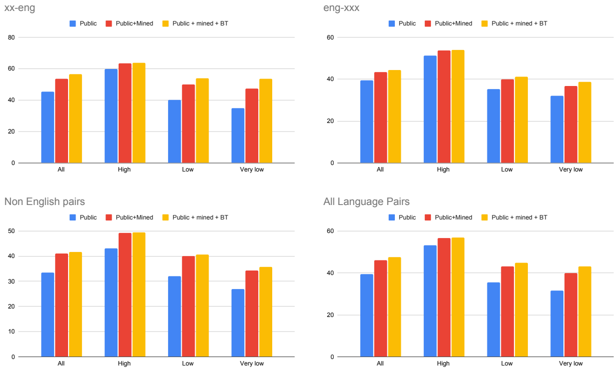
- 1.2 billion people, estimated 2000 languages
- Existing systems support only few African languages
  - LaBSE: 14 (+4)
  - Google translate: 22
- We trained encoders for 55 languages, 48 are low resource
- Specific encoder for languages with Ge'ez script: Amharic and Tigrinya
- Average over 44 languages: BLEU 11.0 $\rightarrow$ 14.8 with mined data

Challenges

- It seems very difficult to crawl textual resources for several languages

Massively Multilingual NMT

Impact of mined bitexts (chrF++)



- Substantial gains in chrF++ when adding mined data
  - very low-resource xx/eng: +12.5 chrF++
  - very low-resource eng/xx: +4.7 chrF++

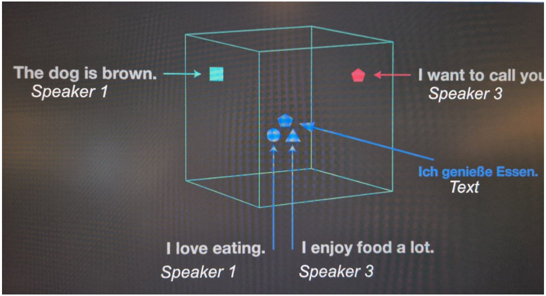
⇒ Mined data is crucial for very low-resource languages

Going Multimodal

What about other modalities?

- Many languages are rather spoken than written

⇒ multilingual and -modal representation



Going Multimodal

Speech LASER

- Apply teacher-student approach to speech

$\Rightarrow$ Fit fixed-size **speech** representation to LASER
  - train with transcriptions, translations or both

- NeurIPS'21 paper:
  P.-A. Duquenne, H. Gong, H. Schwenk, *Multimodal and Multilingual Embeddings for Large-Scale Speech Mining*
- Recent similar works: Data2vec, mSLAM

Large-Scale Speech Mining

Speech-to-text

- SpeechLASER compatible with LASER2 encoder

$\Rightarrow$ We can mine speech against all 200 NLLB languages !

- Mining in Librivox audio books
  - ≈20 000h of audio-text alignments
  - Data substantially boosts S2T translation

Large-Scale Speech Mining

Speech-to-text

- SpeechLASER compatible with LASER2 encoder

$\Rightarrow$ We can mine speech against all 200 NLLB languages !

- Mining in Librivox audio books
  - ≈20 000h of audio-text alignments
  - Data substantially boosts S2T translation

Speech-to-speech mining

- Mine directly speech against speech
- No need to transcribe or translate
- Librivox: 1433h of mined S2S in eng, deu, fra and spa
- Enabled improved S2S translation:
  - A. Lee et al., *Textless Speech-to-Speech Translation on Real Data*, NAACL'22

Open-Source Activities

NLLB: main entry point

- `https://github.com/facebookresearch/fairseq/tree/nllb`
- LID, NMT models
- scripts to reproduce data
- `LASER3 teacher-student training`
- `stopes`: data processing and large-scale mining

Open-Source Activities

LASER github

- `https://github.com/facebookresearch/LASER`
- All LASER3 models
- Mined Bitexts
  - 24 African languages: link WMT'22 workshop
  - remaining languages: soon to come
- 1433h of mined speech-to-speech data in LibriVox

Conclusion

Scaling LASER

- Moved away from the popular one-for-all approach
  - train multiple mutually language specific models
  - alternative to adapters?
- Teacher-student with multiple mutually compatible encoders seems to be very efficient
- Mined more than 1 billion new bitexts
- Enabled scaling NMT to 200 languages and boosted performance
- First successful speech-to-speech mining
- Can we use LASER3 embeddings for other multilingual tasks?

Conclusion

Challenges

- It is very hard to find textual resources for low-resource languages
- Does it make sense to scale translation to thousands of languages?

Conclusion

Challenges

- It is very hard to find textual resources for low-resource languages
- Does it make sense to scale translation to thousands of languages?
- Yes, but I believe that we should switch to the speech modality


Thank You
Asante
Kiitos
Welalin
Terma Kasih
Salamat
Grazie
Dankon
Multumesc
Matondo
Nirringrazzjak
Merci
Spasibo
Dank Je
Ua Tsaug Rau Koj
Mochchakkeram
Obrigado
Arigato
Kia Ora
Raibh Maith Agat