Low-Resource Natural Language Processing
Word Embeddings

Cristina España-Bonet
DFKI GmbH

Low-Resource NLP:
Multilinguality and Machine Translation
Webinar Series — Session I
8th June 2021

What's this all about?

Webinar Series

What’s this all about?

Who are you going to be listening to?

Cristina España-Bonet was born in Barcelona, Catalonia. She received the B.E. in physics and the M.Sc. in astrophysics and cosmology from the Universitat de Barcelona (Catalonia) in 2002 and 2004 respectively. In 2008, she obtained the M.Sc. in artificial intelligence from the Universitat Politècnica de Catalunya (Catalonia) and the Ph.D. in physics from the Universitat de Barcelona. Since then, she has been working on NLP first at Universitat Politècnica de Catalunya and currently at DFKI and the Universität des Saarlandes (Germany). She is especially interested in interlingual and multilingual approaches and in making available tools and methods for low-resourced languages.

Multilingual Technologies @DFKI

What's this all about?

Low-Resource NLP: Multilinguality and Machine Translation

- 5 sessions, 90 minutes each
- General topic:
  Low-Resource NLP: Multilinguality and Machine Translation
- Special interest:
  The **path** towards low-resource machine translation

What's this all about?

Low-Resource NLP: Multilinguality and Machine Translation

- 5 sessions, 90 minutes each
- General topic:
  Low-Resource NLP: Multilinguality and Machine Translation
- Special interest:
  The **path** towards low-resource machine translation
- Related activity:
  Shared task on multilingual translation at WMT 2021
  (close deadline, evaluation campaign June 29–July 6)

What's this all about?

Low-Resource NLP: Multilinguality and Machine Translation

1. Motivation and Thoughts on LR-NLP
2. Word Embeddings
3. Transformer Models
4. Unsupervised Neural Machine Translation
5. Self-Supervised Neural Machine Translation
6. State-of-the-art: WMT Evaluations

What's this all about?

1. Motivation and Thoughts on LR-NLP
2. Word Embeddings
   - Basics
   - Mono-lingual Embeddings
   - Cross-lingual Embeddings
3. Transformer Models
   - Language Modelling
   - Machine Translation
   - Contextual Embeddings
4. Unsupervised Neural Machine Translation
5. Self-Supervised Neural Machine Translation
6. State-of-the-art: WMT Evaluations

What's this all about?

TODAY! Session I: Low-Resource NLP + Word Embeddings

1. Motivation and Thoughts on LR-NLP
2. Word Embeddings
   - Basics
   - Mono-lingual Embeddings
   - Cross-lingual Embeddings

What's this all about?

TODAY! Session I: Low-Resource NLP + Word Embeddings

1. Motivation and Thoughts on LR-NLP
2. Word Embeddings
   - Basics
   - Mono-lingual Embeddings
   - Cross-lingual Embeddings

But before we start, we'll be 7.5 hours together...

let me know you a bit better!

What's this all about?

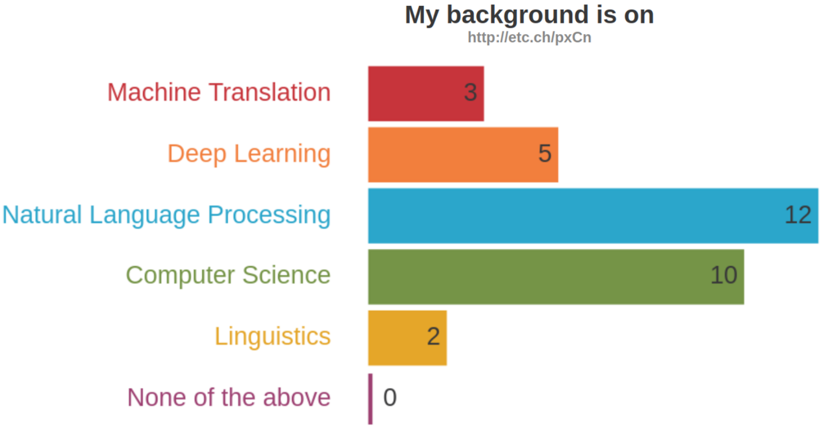
Who am I going to talk to? Let's go interactive! DirectPoll

http://etc.ch/pxCn

DirectPoll Link

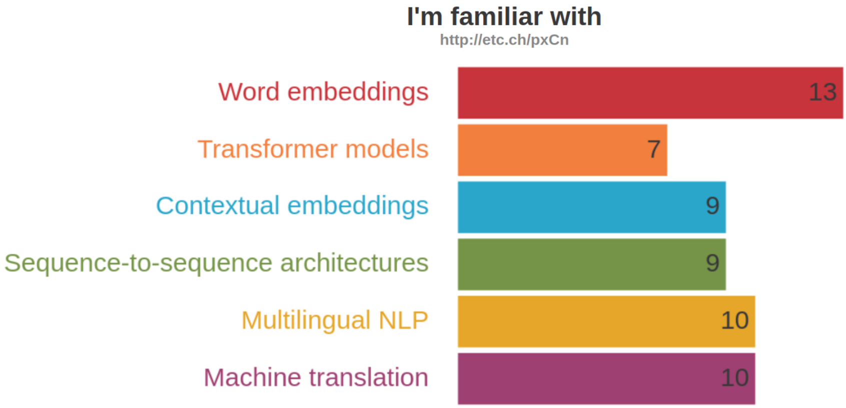
What's this all about?

Let's go interactive! DirectPoll



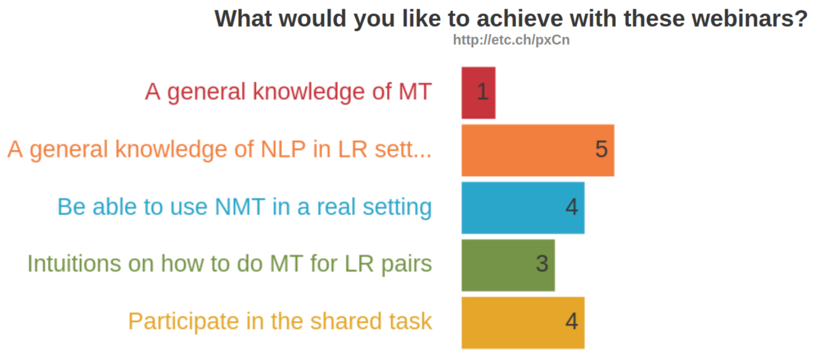
What's this all about?

Let's go interactive! DirectPoll



What's this all about?

Let's go interactive! DirectPoll



Session I

Outline

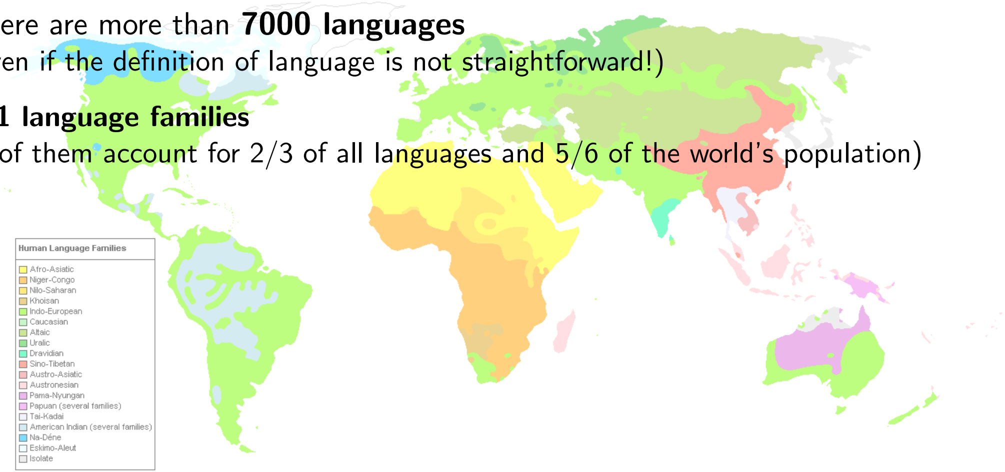
1. Motivation
   - Language Diversity
   - Low-Resource Settings
   - Basic Low-Resource NLP Techniques

2. Word Embeddings
   - Basics
   - Frequency and Prediction-based Embeddings
   - Cross-lingual Embeddings

Motivation

Language Diversity: Some Numbers

- There are more than **7000 languages**
  (even if the definition of language is not straightforward!)
- **141 language families**
  (6 of them account for 2/3 of all languages and 5/6 of the world's population)



Language Diversity: Some Numbers

- There are more than **7000 languages**
  (even if the definition of language is not straightforward!)
- **141 language families**
  (6 of them account for 2/3 of all languages and 5/6 of the world's population)

Explore:

Ethnologue `https://www.ethnologue.com/`
Glottolog `http://glottolog.org/`
Linguistic Maps `http://linguisticmaps.tumblr.com/`

Motivation

Language Diversity: Related Concepts

language diversity

population density

endangered languages

digital richness

low-resource language

Motivation

Language Diversity: Related Concepts

language diversity

population density

endangered languages

digital richness

low-resource language

low-resource setting vs. low-resource language

Motivation

Endangered Languages, UNESCO Definition (2010)

Vitality	Transmission of the language from one generation to another
Safe	The language is spoken by all generations; intergenerational transmission is uninterrupted.
Vulnerable	Most children speak the language, but it can be restricted to certain areas.
Endangered	Children no longer learn the language as a mother tongue at home.
Severally endangered	The language is spoken by the grandparents; while the generation of parents can understand it, they do not speak it among themselves or with the children.
Critically endangered	The youngest speakers are grandparents and their ancestors, and they speak the language only partially and infrequently.
Extinct	There are no more speakers.

https://www.swisstranslate.ch/en/news/endangered-languages/

Motivation

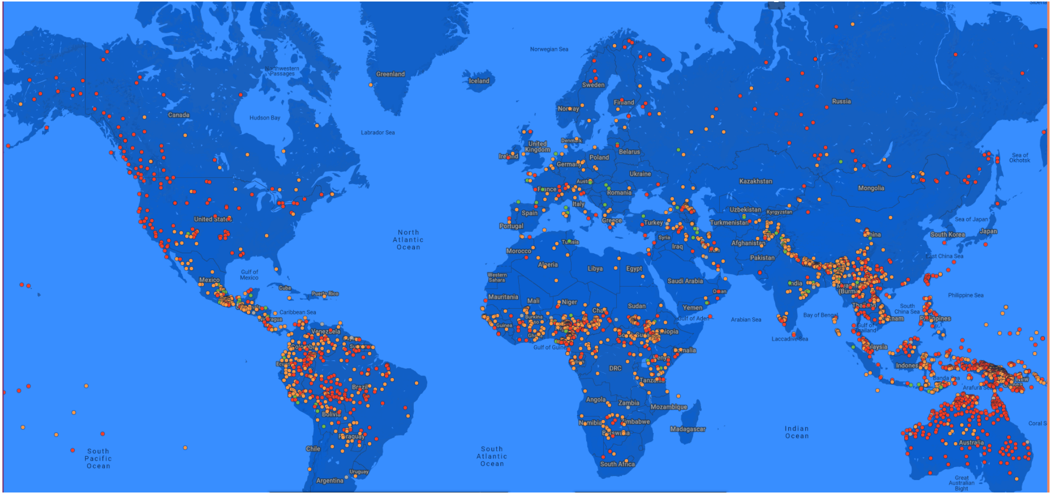
Endangered Languages, UNESCO Statistics (2010)

Status	Africa	America	Asia	Europe	Oceania	Total
Vulnerable	48	260	301	67	34	710
Endangered	72	182	392	133	49	828
Severally endangered	92	181	185	73	80	611
Critically endangered	100	235	174	17	90	616
Extinct (since 1950)	50	95	61	6	21	233
Total	362	953	1 113	296	274	2 998

`https://www.swisstranslate.ch/en/news/endangered-languages/`

Motivation

Endangered Languages as Effect (?) of Diversity



http://www.endangeredlanguages.com

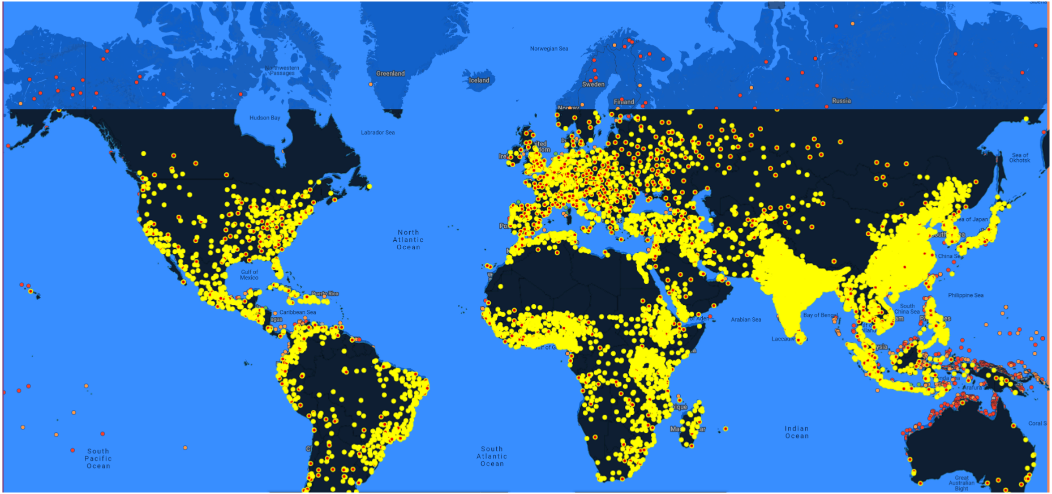
Motivation

Endangered Languages as Effect (?) of Diversity

- The situation is very different in different regions of the world
- Three "hot spots"
  - Central/South America,
  - North Sub-Saharan Africa,
  - South/Southeast Asia and Oceania
- No direct correlation with population density

Motivation

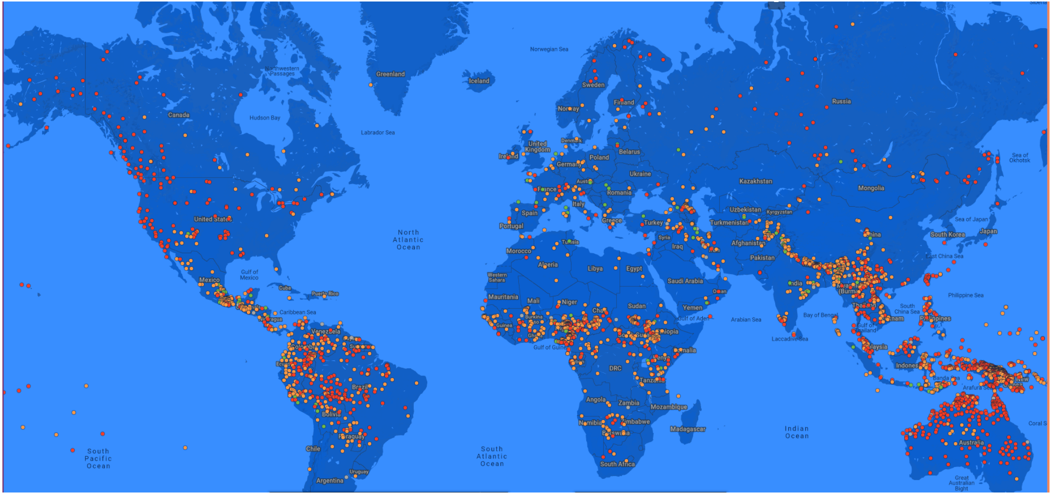
Endangered Languages as Effect (?) of Diversity



https://worldpopulationhistory.org

Motivation

Endangered Languages as Effect (?) of Diversity



http://www.endangeredlanguages.com

Motivation

Endangered Language vs. Low-Resource Language

- No single reason for being endangered language
  - Low population
  - Coexistence with a strong language
  - Politics
  - NEW: Lag behind in digital content

Motivation

Endangered Language vs. Low-Resource Language

- No single reason for being endangered language
  - Low population
  - Coexistence with a strong language
  - Politics
  - NEW: Lag behind in digital content
- Endangered Language $\Rightarrow$ Low-Resource Language (lower levels)

Motivation

Endangered Language vs. Low-Resource Language

- No single reason for being endangered language
  - Low population
  - Coexistence with a strong language
  - Politics
  - NEW: Lag behind in digital content
- Endangered Language $\Rightarrow$ Low-Resource Language (lower levels)
- Low-Resource Language $\not\Rightarrow$ Endangered Language (not necessarily)

Motivation

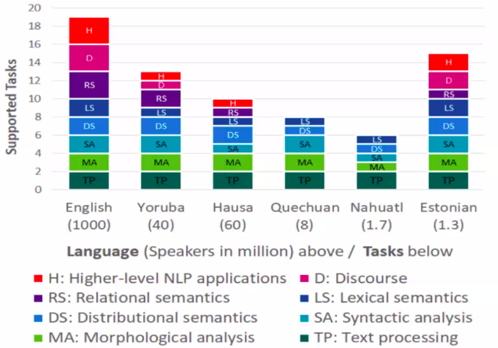
Endangered Language vs. Low-Resource Language

- No single reason for being endangered language
  - Low population
  - Coexistence with a strong language
  - Politics
  - NEW: Lag behind in digital content
- Endangered Language $\Rightarrow$ Low-Resource Language (lower levels)
- Low-Resource Language $\nRightarrow$ Endangered Language (not necessarily)
- No single cause for being low-resource

Motivation

What's the Meaning of Low-Resource?

There is no universal definition. **Few linguistic resources? Few data?**



(Hedderick et al., 2020)

Motivation

What's the Meaning of Low-Resource?

There is no universal definition. **Few linguistic resources? Few data?**

I prefer to talk about **low-resource setting** because

- Task dependent
  - speech recognition vs. machine translation vs. PoS tagging
- Language (complexity) dependent
  - English vs. Hungarian
- Domain dependent!
  - English text generation: sport vs. corona in March 2020
- Author dependent!

Motivation

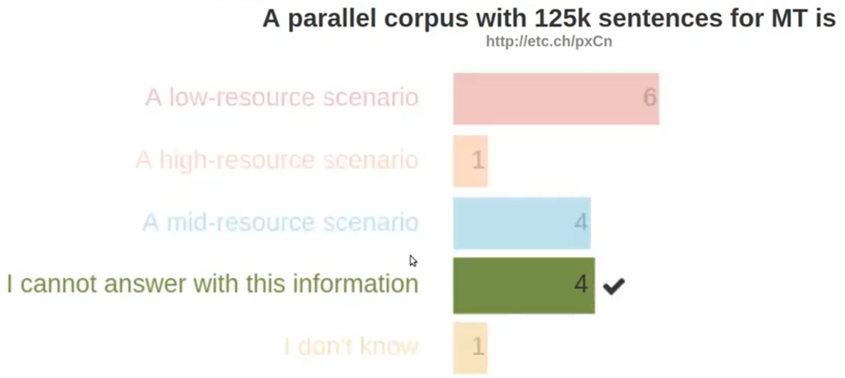
What's a Low-Resource Setting?

Definition. *A low-resource setting is a scenario where standard NLP techniques are not usable (low/null performance).*

Cristina dixit
Don't take it for universal!

Motivation

Example: Low-Resource Machine Translation



Motivation

Example: What is Low-Resource Machine Translation?

AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas (Mager et al. 2021)

Language	**ISO**	**Family**	**Train**	**Dev**	**Test**
Asháninka	cni	Arawak	3883	883	1002
Aymara	aym	Aymaran	6531	996	1003
Bribri	bzd	Chibchan	7508	996	1003
Guarani	gn	Tupi-Guarani	26032	995	1003
Nahuatl	nah	Uto-Aztecan	16145	672	996
Otomí	oto	Oto-Manguean	4889	599	1001
Quechua	quy	Quechuan	125008	996	1003
Rarámuri	tar	Uto-Aztecan	14721	995	1002
Shipibo-Konibo	shp	Panoan	14592	996	1002
Wixarika	hch	Uto-Aztecan	8966	994	1003

Motivation

Example: What is Low-Resource Machine Translation?

AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas (Bollmann et al. 2021)

Set	System	Track	Languages									
			AYM	BZD	CNI	GN	HCH	NAH	OTO	QUY	SHP	TAR
DEV	CoAStaL-1: Phrase-based	1	2.57	3.83	2.79	2.59	6.81	2.33	1.44	1.73	3.70	1.26
	CoAStaL-2: Random	2	0.02	0.03	0.04	0.02	1.14	0.02	0.02	0.02	0.06	0.02
TEST	Helsinki-2 (best)	1	2.80	5.18	6.09	8.92	15.67	3.25	5.59	5.38	10.49	3.56
	CoAStaL-1: Phrase-based	1	1.11	3.60	3.02	2.20	8.80	2.06	2.72	1.63	3.90	1.05
	+ extra data	1	1.07	–	–	2.24	–	2.06	–	1.24	–	–
	CoAStaL-2: Random	2	0.05	0.06	0.03	0.03	2.07	0.03	0.03	0.02	0.04	0.06
	Baseline	2	0.01	0.01	0.01	0.12	2.20	0.01	0.00	0.05	0.01	0.00

Main Approaches, Keywords

transfer learning

Basic Low-Resource NLP Techniques

Main Approaches, Keywords

transfer learning

few-shot learning

Basic Low-Resource NLP Techniques

Main Approaches, Keywords

transfer learning

few-shot learning

pretraining

Basic Low-Resource NLP Techniques

Main Approaches, Keywords

transfer learning

multi-task learning

few-shot learning

pretraining

Basic Low-Resource NLP Techniques

Main Approaches, Keywords

transfer learning

multi-task learning

few-shot learning

pretraining

data augmentation

Basic Low-Resource NLP Techniques

Main Approaches, Keywords

transfer learning

multi-task learning

few-shot learning

semi-supervised training

pretraining

data augmentation

Basic Low-Resource NLP Techniques

Main Approaches, Keywords

transfer learning

multi-task learning

few-shot learning

zero-shot learning

semi-supervised training

pretraining

data augmentation

Basic Low-Resource NLP Techniques

Main Approaches, Keywords

transfer learning

multi-task learning

few-shot learning

zero-shot learning

semi-supervised training

weak supervision

pretraining

data augmentation

Basic Low-Resource NLP Techniques

Main Approaches, Keywords

transfer learning

multi-task learning

few-shot learning

zero-shot learning

semi-supervised training

weak supervision

distillation

pretraining

data augmentation

Basic Low-Resource NLP Techniques

Main Approaches, Keywords

transfer learning

multi-task learning

few-shot learning

zero-shot learning

semi-supervised training

domain adaptation

weak supervision

distillation

pretraining

data augmentation

Basic Low-Resource NLP Techniques

Main Approaches, Keywords

transfer learning

multi-task learning

few-shot learning

zero-shot learning

semi-supervised training

domain adaptation

weak supervision

distillation

pretraining

data augmentation

Basic Low-Resource NLP Techniques

Main Approaches

1. Data enrichment
   - Data collection
   - Data augmentation
2. General machine learning
   - Unsupervised learning
   - Weak supervision
   - Transfer learning
3. Multilinguality and/or multimodality
4. Specialised architectures

Main Approaches: Data Augmentation Examples

Data augmentation is good in general (acts as a regularisation for NN)

- **Some methods:**
  - Oversampling,
  - transformations of existing instances,
  - create new instances...

- **Some examples:**
  - Backtranslation or noise addition for MT,
  - transformation of images (geometry, color...) for vision,
  - change the sample rate of waveforms or modify the spectrogram for ASR...

Basic Low-Resource NLP Techniques

Main Approaches: General Machine Learning

Weak Supervision – Related to data augmentation

Usage of noisy instances as signals to label large amounts of training data
⇒ supervised learning

- Useful for NER, PoS tagging, etc.
- But also MT-like (backtranslation)

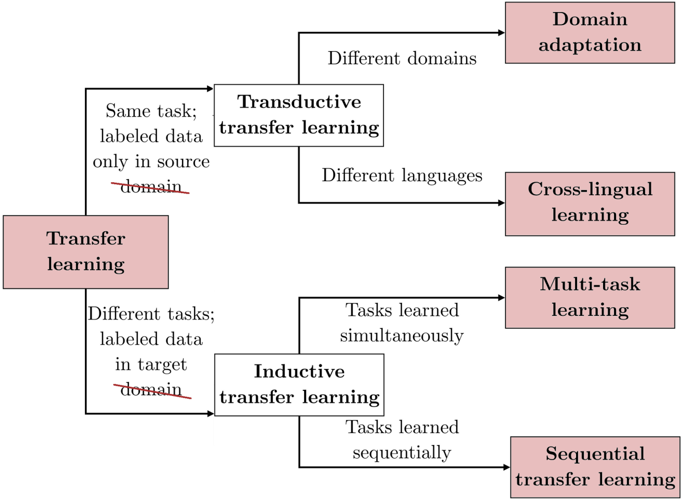
Transfer learning

Use a pretrained model for a different but related task as starting point.

- Useful for domain adaptation, task adaptation, etc.
- But also for few, zero-shot languages

Basic Low-Resource NLP Techniques

Main Approaches: More on Transfer Learning



Sebastian Ruder's PhD thesis

Basic Low-Resource NLP Techniques

Example in Machine Translation: Yorùbá–English (Adelani et al., 2021)

How much data do we have?

Domain	**Train. Set**	**Dev. Set**	**Test Set**
Standard (religious) corpora			
Bible	30,760	–	–
JW300	459,871	–	–

Is it enough?

Basic Low-Resource NLP Techniques

Example in Machine Translation: Yorùbá–English (Adelani et al., 2021)

How much data do we have?

Domain	**Train. Set**	**Dev. Set**	**Test Set**
Standard (religious) corpora			
Bible	30,760	–	–
JW300	459,871	–	–

Is it enough? What's translation quality (BLEU) on out-of-domain?

Transformer trained with	*en***2***yo*	*yo***2***en*
Bible	2.2±0.1	1.4±0.1
JW300	7.5±0.2	9.6±0.3
JW300+Bible	8.1±0.2	10.8±0.3

Basic Low-Resource NLP Techniques

Example in Machine Translation: Yorùbá–English (Adelani et al., 2021)

1. Data enrichment: data collection (number of sentences in MENYO-20k)

Domain	**Train. Set**	**Dev. Set**	**Test Set**
Standard (religious) corpora			
Bible	30,760	–	–
JW300	459,871	–	–
MENYO-20k			
News	4,995	1,391	3,102
TED Talks	507	438	2,000
Book	-	1,006	1,008
IT	356	312	273
Proverbs	2,200	250	250
Others	2,012	250	250
TOTAL	500,701	3,397	6,633

Basic Low-Resource NLP Techniques

Example in Machine Translation: Yorùbá–English *(Adelani et al., 2021)*

2 Transfer learning for domain adaptation (with MENYO-20k 1)

3 Backtranslation (with the best system)

4 Weak supervision (supervised training with additional data 3)

2 Transfer learning (with MENYO-20k 1) from pretrained models

1 Multilinguality (baseline 0)

Basic Low-Resource NLP Techniques

Example in Machine Translation: Yorùbá–English (Adelani et al., 2021)

Model (tested on MENYO-20k)	*en*2*yo*	*yo*2*en*
JW300+Bible **baseline**	8.1±0.2	10.8±0.3
+Transfer learning **domain adaptation**	12.3±0.3	13.2±0.3
JW300+Bible+MENYO-20k **domain adaptation**	10.9±0.3	14.0±0.3
+Transfer learning **domain adaptation**	**12.4±0.3**	14.6±0.3
+ Backtranslation **data augmentation**	12.0±0.3	**18.2±0.4**

Basic Low-Resource NLP Techniques

Example in Machine Translation: Yorùbá–English (Adelani et al., 2021)

Model (tested on Menyo-20k)	*en***2***yo*	*yo***2***en*
JW300+Bible **baseline**	8.1±0.2	10.8±0.3
+Transfer learning **domain adaptation**	12.3±0.3	13.2±0.3
JW300+Bible+Menyo-20k **domain adaptation**	10.9±0.3	14.0±0.3
+Transfer learning **domain adaptation**	**12.4±0.3**	14.6±0.3
+ Backtranslation **data augmentation**	12.0±0.3	**18.2±0.4**
mT5-base+Transfer learning **pretraining** **task adaptation**	11.5±0.3	16.3±0.4

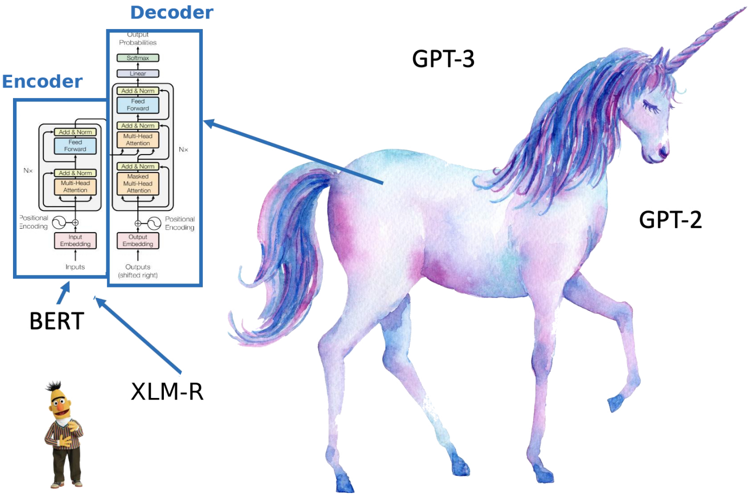
Basic Low-Resource NLP Techniques

Example in Machine Translation: Yorùbá–English *(Adelani et al., 2021)*

Model (tested on MENYO-20k)	*en*2*yo*	*yo*2*en*
JW300+Bible **baseline**	8.1±0.2	10.8±0.3
+Transfer learning **domain adaptation**	12.3±0.3	13.2±0.3
JW300+Bible+MENYO-20k **domain adaptation**	10.9±0.3	14.0±0.3
+Transfer learning **domain adaptation**	**12.4±0.3**	14.6±0.3
+ Backtranslation **data augmentation**	12.0±0.3	**18.2±0.4**
mT5-base+Transfer learning **pretraining** **task adaptation**	11.5±0.3	16.3±0.4
Google GMNMT **multilingual**	3.7±0.2	**22.4±0.5**
Facebook M2M-100 **multilingual**	3.3±0.2	4.6±0.3
OPUS-MT **bilingual**	–	5.9±0.2

Basic Low-Resource NLP Techniques

Short Digression: Parameters in Pretrained Models



(Adapted from https://www.programmersought.com/article/24793362644/)

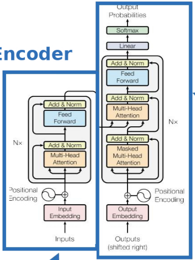
Basic Low-Resource NLP Techniques

Short Digression: Parameters in Pretrained Models

Number of layers ⤳

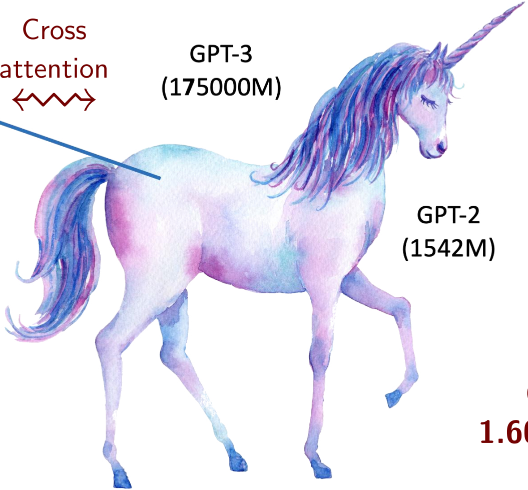
Encoder

Decoder



Cross attention ⟿

GPT-3 (175000M)



GPT-2 (1542M)

BERT (340M)

XLM-R (550M)

Google 1.600.000M!

(Adapted from https://www.programmersought.com/article/24793362644/)

Short Digression: Parameters in Pretrained Models, Energy Costs

You are viewing Rada Mihalcea's screen View Options

Energy & Cost Considerations

(Strubell et al., 2020)

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$T2T_{base}$	P100x8	1415.78	12	27	26	\$41–\$140
$T2T_{big}$	P100x8	1515.43	84	201	192	\$289–\$981
ELMo	P100x3	517.66	336	275	262	\$433–\$1472
$BERT_{base}$	V100x64	12,041.51	79	1507	1438	\$3751–\$12,571
$BERT_{base}$	TPUv2x16	—	96	—	—	\$2074–\$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	\$942,973–\$3,201,722
NAS	TPUv2x1	—	32,623	—	—	\$44,055–\$146,848
GPT-2	TPUv3x32	—	168	—	—	\$12,902–\$43,008

Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

Basic Low-Resource NLP Techniques

Some (Final?) Thoughts

- Low-resource NLP is today a **hard problem**
  - No data in a data world
- This introduction is **not exhaustive** at all
  - I've given keywords, now we'll see the basics and end with SotA
- There is no **universal solution**
  - Discussion on the previous thoughts in the last session

Basic Low-Resource NLP Techniques

Some (Final?) Thoughts II

What to do next?

- A mini-break and learn the very very basics of word embeddings :-)
- Already an expert? Join us:
  - Shared Task: Multilingual Low-Resource Translation for Indo-European Languages (WMT2021@EMNLP2021)
  - First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL2021@RANLP2021)

More to come!!

Thanks! And...

wait!

Low-Resource Natural Language Processing
Word Embeddings

Cristina España-Bonet
DFKI GmbH

Low-Resource NLP:
Multilinguality and Machine Translation
Webinar Series — Session I
8th June 2021