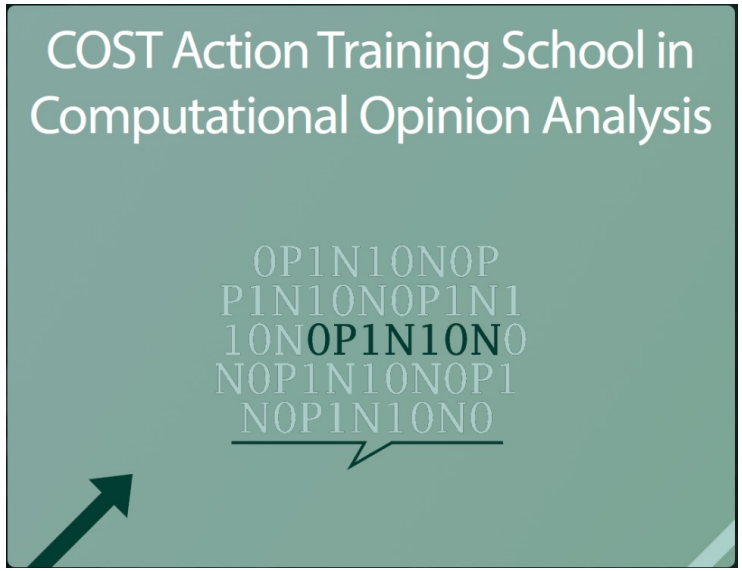
Ollscoil na Gaillimhe
University of Galway

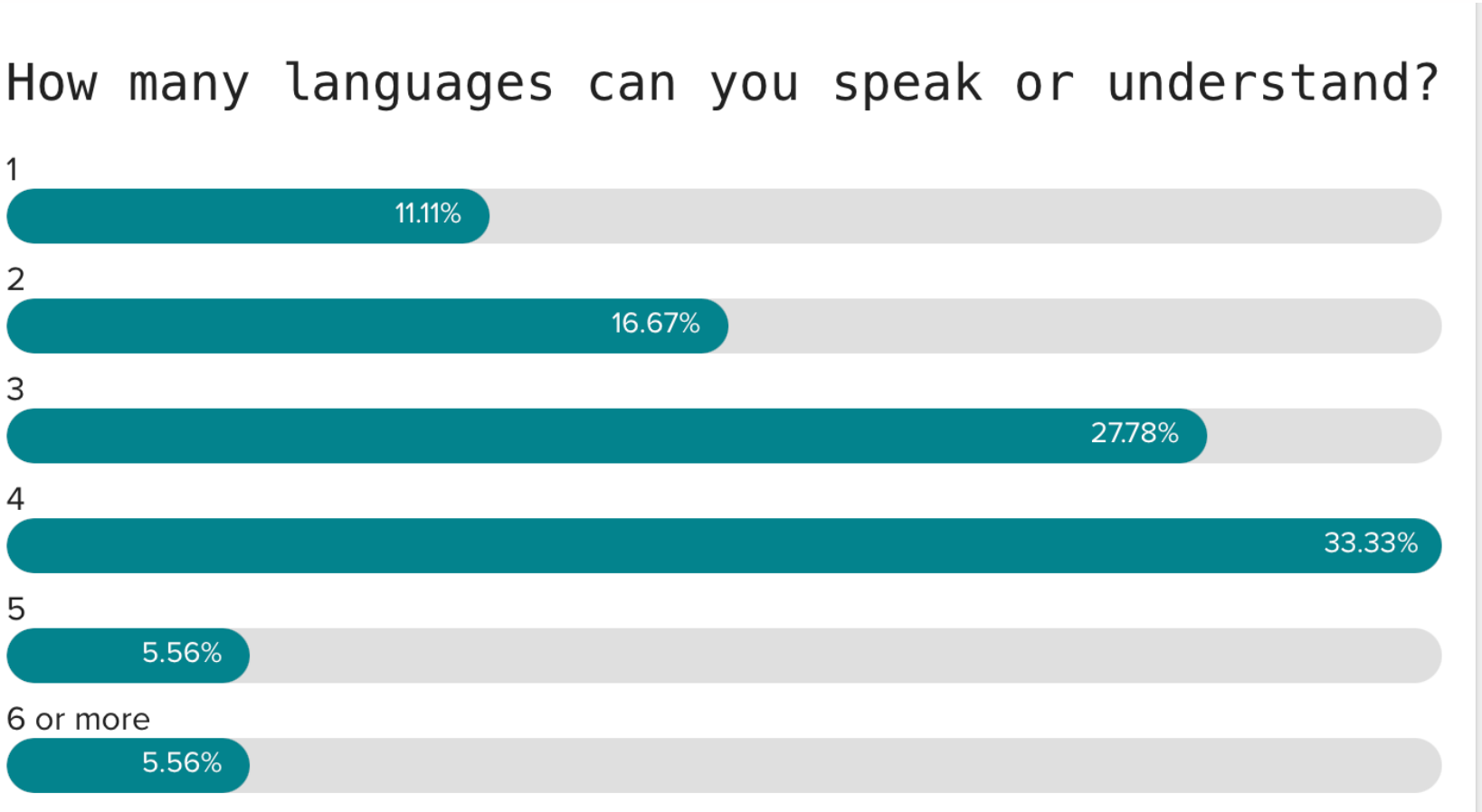
Managing and Processing Multiple Languages

Jamal Nasir



University
*of*Galway.ie

How many languages can you speak or understand?


How many languages can you speak or understand?
1
11.11%
2
16.67%
3
27.78%
4
33.33%
5
5.56%
6 or more
5.56%


Ollscoil na Gaillimhe
University of Galway
Managing and Processing Multiple Languages: Data or Technology
University ofGalway.ie

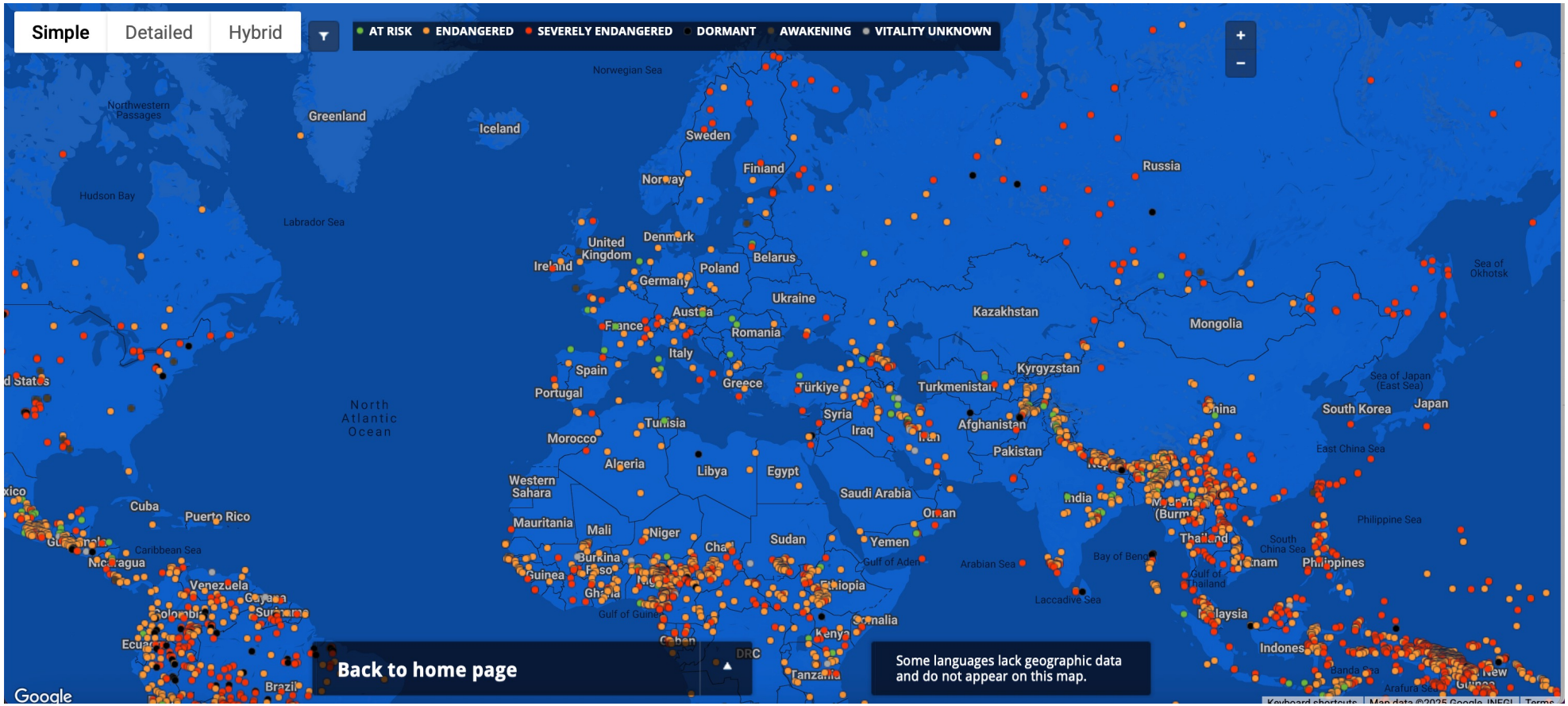
How many Languages are there in the world?

7,159 languages are in use today. ☐

That number is constantly in flux, because we're learning more about the world's languages every day. And beyond that, the languages *themselves* are in flux. They're living and dynamic, used by communities whose lives are shaped by our rapidly changing world. This is a fragile time: Roughly 44% of all languages are now endangered, often with fewer than 1,000 users remaining. Meanwhile, the world's 20 largest languages are the native tongue of more than 3.7 billion people total. That's just 0.3% of the world's languages accounting for nearly half of the world's population!

https://www.ethnologue.com/insights/how-many-languages/

Endangered Languages



https://endangeredlanguages.com

Why Multilingual NLP?

Speaking **more languages** means communicating with **more people**…
…and reaching **more users and customers**…

Why Multilingual NLP?

- Decreasing the digital language divide

Why Multilingual NLP?

- Dealing with inequality of information (access)
- Mitigating cross-cultural biases
- Deploying language technology for underrepresented languages, dialects, minorities; societal impact
- Understanding cross-linguistic differences

"95% of all languages in use today will never gain traction online" (Andras Kornai)

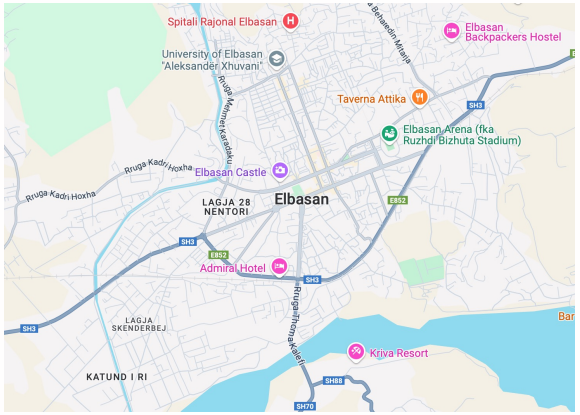
Why Multilingual NLP?

- Inequality of information and representation can also affect how we understand places, events, processes...

We're in Elbasan searching for...

... restorantet (Al)

...Restaurant (En)



...éttermek (HU)

How to Cope Multilingual?

Better Data:

- ❖ – every piece of relevant data can help - be resourceful!
- ❖ – make data if necessary - be connected!

Better Models or Algorithms:

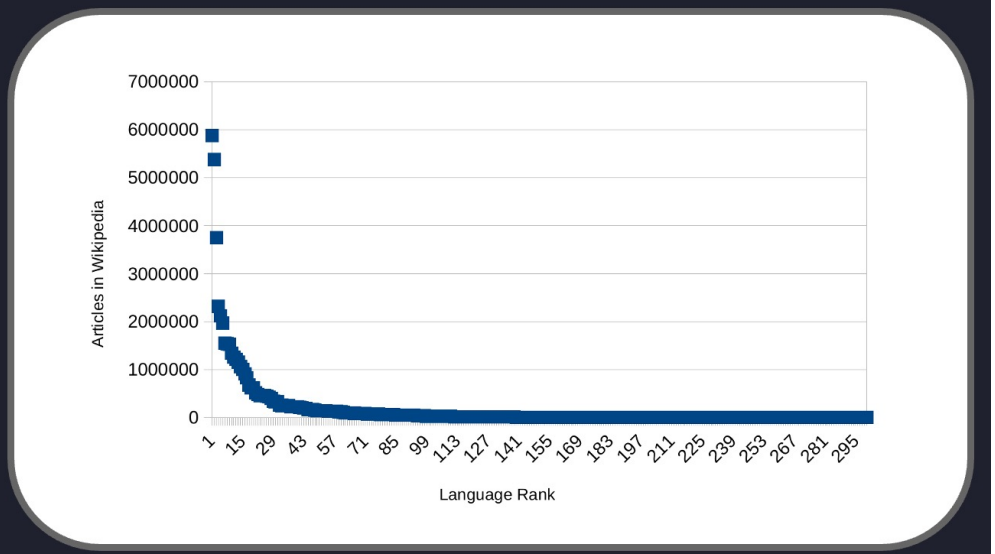
- ❖ – sophisticated modelling/training methods - know NLP/ML!
- ❖ – linguistically informed methods - know linguistics!

Better Deployment:

- ❖ – different situations require different solutions - be aware!


Ollscoil na Gaillimhe
University of Galway
Data
University
ofGalway.ie

Issues: The Long Tail of Data



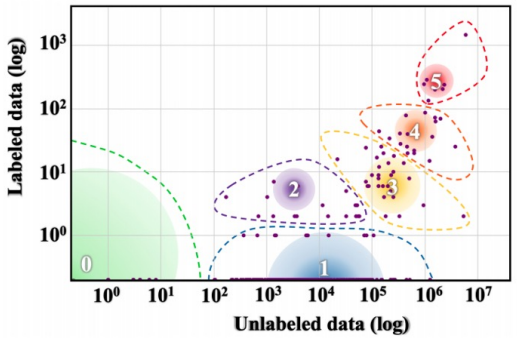
Even getting “raw” unannotated data is problematic for many languages...

Natural Language Processing for Multilingual Task-Oriented Dialogue

Tutorial Abstract

Evgeniia Razumovskaia[1], Goran Glavaš[2], Olga Majewska[1]
Edoardo Maria Ponti[4,5], Ivan Vulić[1,6]
[1] Language Technology Lab, University of Cambridge
[2] Center for Artificial Intelligence and Data Science, University of Würzburg
[4]Mila – Quebec Artificial Intelligence Institute [5]McGill University [6]PolyAI Limited
{er563,om304,iv250}@cam.ac.uk goran@informatik.uni-mannheim.de
edoardo-maria.ponti@mila.quebec

Are all languages created equal?



Most languages are "Left-Behinds"
[Joshi et al., ACL-20; Blasi et al., ACL-22]

Is creating equitable language technology across different languages then even possible?

Can we at least try to 'approximate' equality?

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
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%

Natural Language Processing for Multilingual Task-Oriented Dialogue
Tutorial Abstract

Evgeniia Razumovskaia[1], **Goran Glavaš**[2], **Olga Majewska**[1]
Edoardo Maria Ponti[4,5], **Ivan Vulić**[1,6]
[1] Language Technology Lab, University of Cambridge
[2] Center for Artificial Intelligence and Data Science, University of Würzburg
[4]Mila – Quebec Artificial Intelligence Institute [5]McGill University [6]PolyAI Limited
{er563,om304,iv250}@cam.ac.uk goran@informatik.uni-mannheim.de
edoardo-maria.ponti@mila.quebec

What we can do in Multilingual Setup?

- Many NLP tasks share common knowledge about language (e.g. linguistic representations, structural similarities)
- Languages share common structure (on the lexical, syntactic, and semantic level)
- Annotated data is rare, make use of as much supervision as available

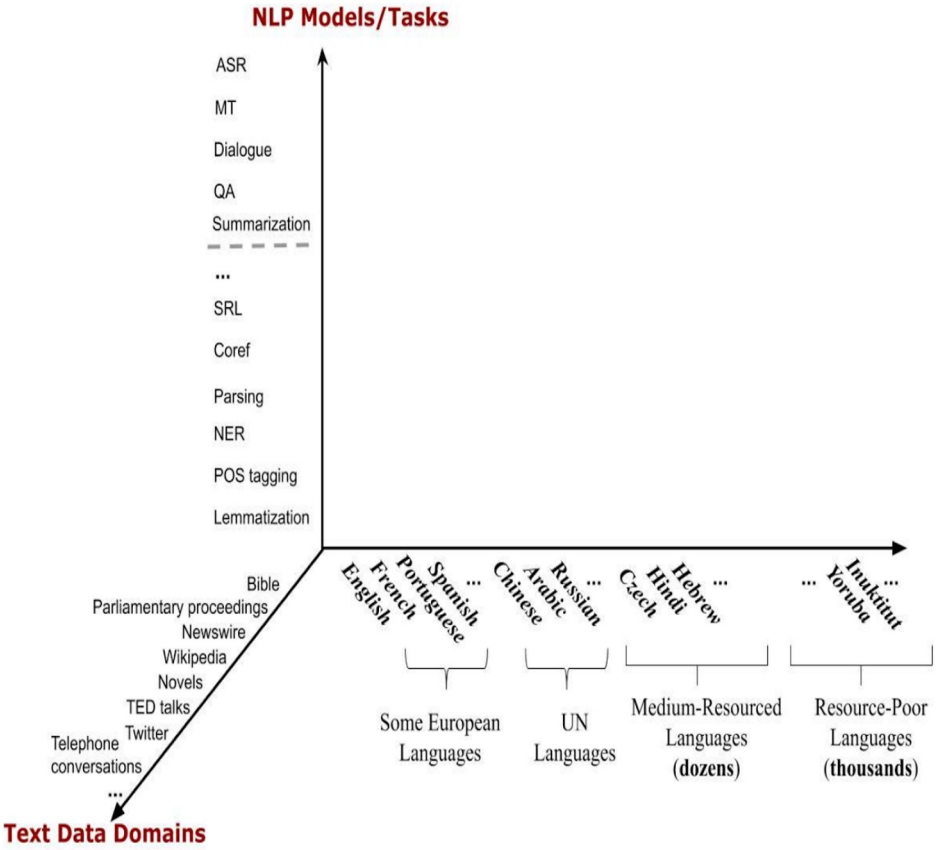


Image courtesy of *Yulia Tsvetkov*

Data Creation/Curation

- What types of data? (monolingual? multilingual? annotated?)
- Where can we get it? (annotated data sources? curated text collections? scraping?)
- Can we create data? (efficient, high-quality creation strategies)
- How do we deal with the ethical issues? (working with communities, language ownership)

Parallel Corpora

			Sm.	Lg.
		CLASSIC SOUPS		
清燉雞湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞飯湯	58.	Chicken Rice Soup	1.85	3.25
雞麵湯	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
蕃茄雪湯	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞湯	62.	Regular Wonton Soup	1.10	2.10
酸辣湯	63.	Hot & Sour Soup	1.10	2.10
雪花湯	64.	Egg Drop Soup	1.10	2.10
雲雪湯	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66.	Tofu Vegetable Soup	NA	3.50
雞玉米湯	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮湯	69.	Seafood Soup	NA	3.50

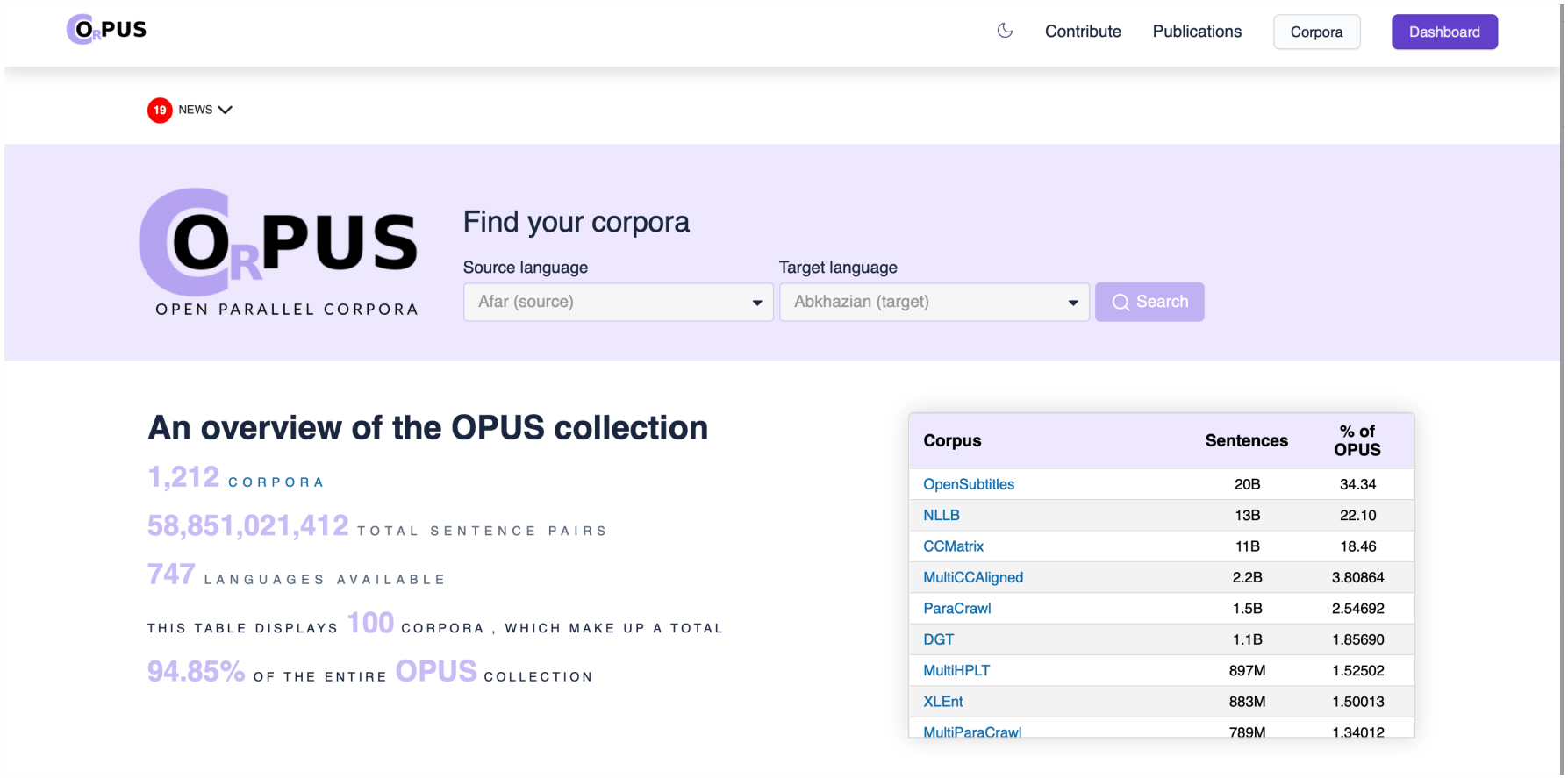
Do We have enough Parallel data?

Parallel Corpus	Sentences	Parallel Corpus	Sentences
Romanian-English	399,375	Greek-English	1,235,976
Bulgarian-English	406,934	Swedish-English	1,862,234
Slovene-English	623,490	Italian-English	1,909,115
Hungarian-English	624,934	German-English	1,920,209
Polish-English	632,565	Finnish-English	1,924,942
Lithuanian-English	635,146	Portuguese-English	1,960,407
Latvian-English	637,599	Spanish-English	1,965,734
Slovak-English	640,715	Danish-English	1,968,800
Czech-English	646,605	Dutch-English	1,997,775
Estonian-English	651,746	French-English	2,007,723

Europarl parallel data: http://www.statmt.org/europarl/

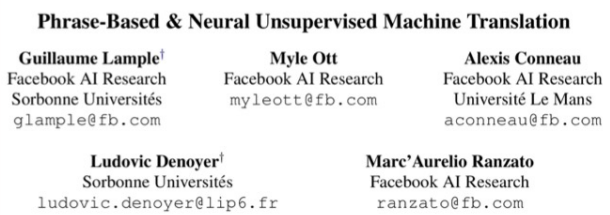
The Open Parallel corpus

opus.nlpl.eu



What if don't have parallel data?

https://arxiv.org/pdf/1804.07755.pdf



Phrase-Based & Neural Unsupervised Machine Translation

Guillaume Lample[†]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott
Facebook AI Research
myleott@fb.com

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

Ludovic Denoyer[†]
Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc'Aurelio Ranzato
Facebook AI Research
ranzato@fb.com

https://arxiv.org/pdf/1711.00043.pdf

UNSUPERVISED MACHINE TRANSLATION
USING MONOLINGUAL CORPORA ONLY

Guillaume Lample † ‡ , **Alexis Conneau** † , **Ludovic Denoyer** ‡ , **Marc'Aurelio Ranzato** †
† Facebook AI Research,
‡ Sorbonne Universités, UPMC Univ Paris 06, LIP6 UMR 7606, CNRS
{gl,aconneau,ranzato}@fb.com, ludovic.denoyer@lip6.fr

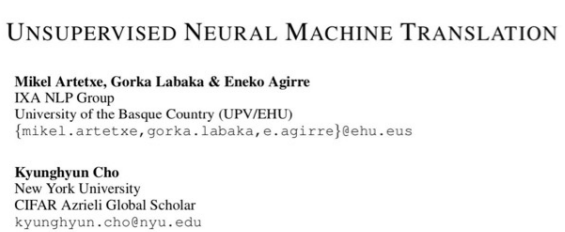
https://arxiv.org/pdf/1901.07291.pdf

Cross-lingual Language Model Pretraining

Guillaume Lample*
Facebook AI Research
Sorbonne Universités
glample@fb.com

Alexis Conneau*
Facebook AI Research
Université Le Mans
aconneau@fb.com

https://arxiv.org/pdf/1710.11041.pdf



UNSUPERVISED NEURAL MACHINE TRANSLATION

Mikel Artetxe, Gorka Labaka & Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

Kyunghyun Cho
New York University
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu


Ollscoil na Gaillimhe
University of Galway
Technology
University
ofGalway.ie

NLP History

1950-69
- Machine Translation

1970-92
- NLP demo systems (syntax and reference in human languages)
- hand-based rules to knowledge-based AI

1993-2012
- Empirical ML models of NLP
- Construction of annotated linguistic resources

2013-Present
- ANN/Deep Learning (2013-18)
- Transformers/BERT/LLMs

NLP Applications

- Text classification
- Token classification
- Question answering
- Causal language modelling
- Masked language modelling
- Translation
- Summarization
- Multiple choice

Language Models

Conditional probability

$$p(w_1, w_2, w_3, \ldots, w_N) =$$
$$p(w_1)\, p(w_2|w_1)\, p(w_3|w_1, w_2) \times \ldots \times p(w_N|w_1, w_2, \ldots w_{N-1})$$

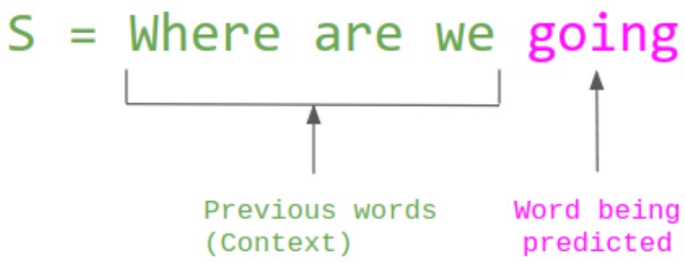
Sentence: “the cat sat on the mat”

$$P(\text{the cat sat on the mat}) = P(\text{the}) * P(\text{cat}|\text{the}) * P(\text{sat}|\text{the cat})$$
$$* P(\text{on}|\text{the cat sat}) * P(\text{the}|\text{the cat sat on})$$
$$* P(\text{mat}|\text{the cat sat on the})$$

Implicit order

Language Models

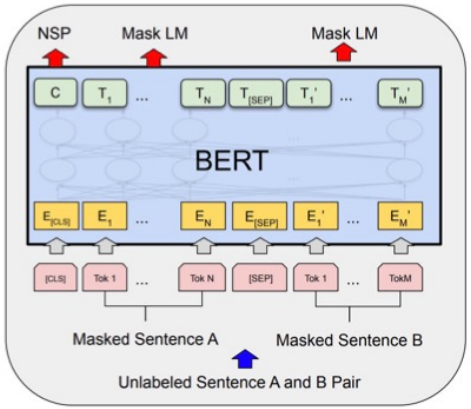
$$P_{(w_1,w_2,\ldots,w_n)} = p(w_1)p(w_2|w_1)p(w_3|w_1,w_2)...p(w_n|w_1,w_2,..,w_{n-1})$$
$$= \prod_{i=1}^{n} p(w_i|w_1,...,w_{i-1})$$

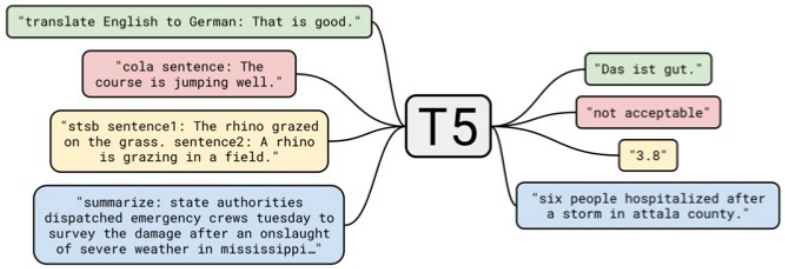


P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

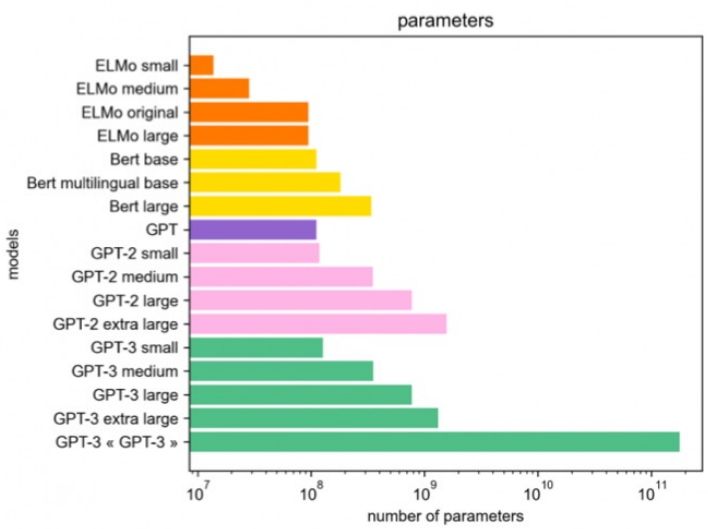
Language Model

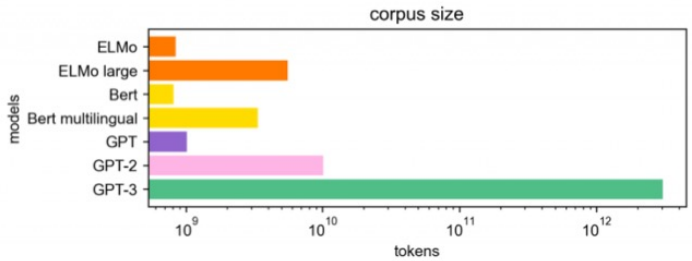
- Decoder-only models (GPT-x models)
- Encoder-only models (BERT, RoBERTa, ELECTRA)
- Encoder-decoder models (T5, BART)





How large are ‘large’ LMs?





More recent models: PaLM (540B), OPT (175B), BLOOM (176B)…

Image source: https://hellofuture.orange.com/en/the-gpt-3-language-model-revolution-or-evolution/

How large are ‘large’ LMs?

- Today, we mostly talk about two camps of models:
  - Medium-sized models: BERT/RoBERTa models (100M or 300M), T5 models (220M, 770M, 3B)
  - “Very” large LMs: models of 100+ billion parameters

What we can do in Multilingual Setup?

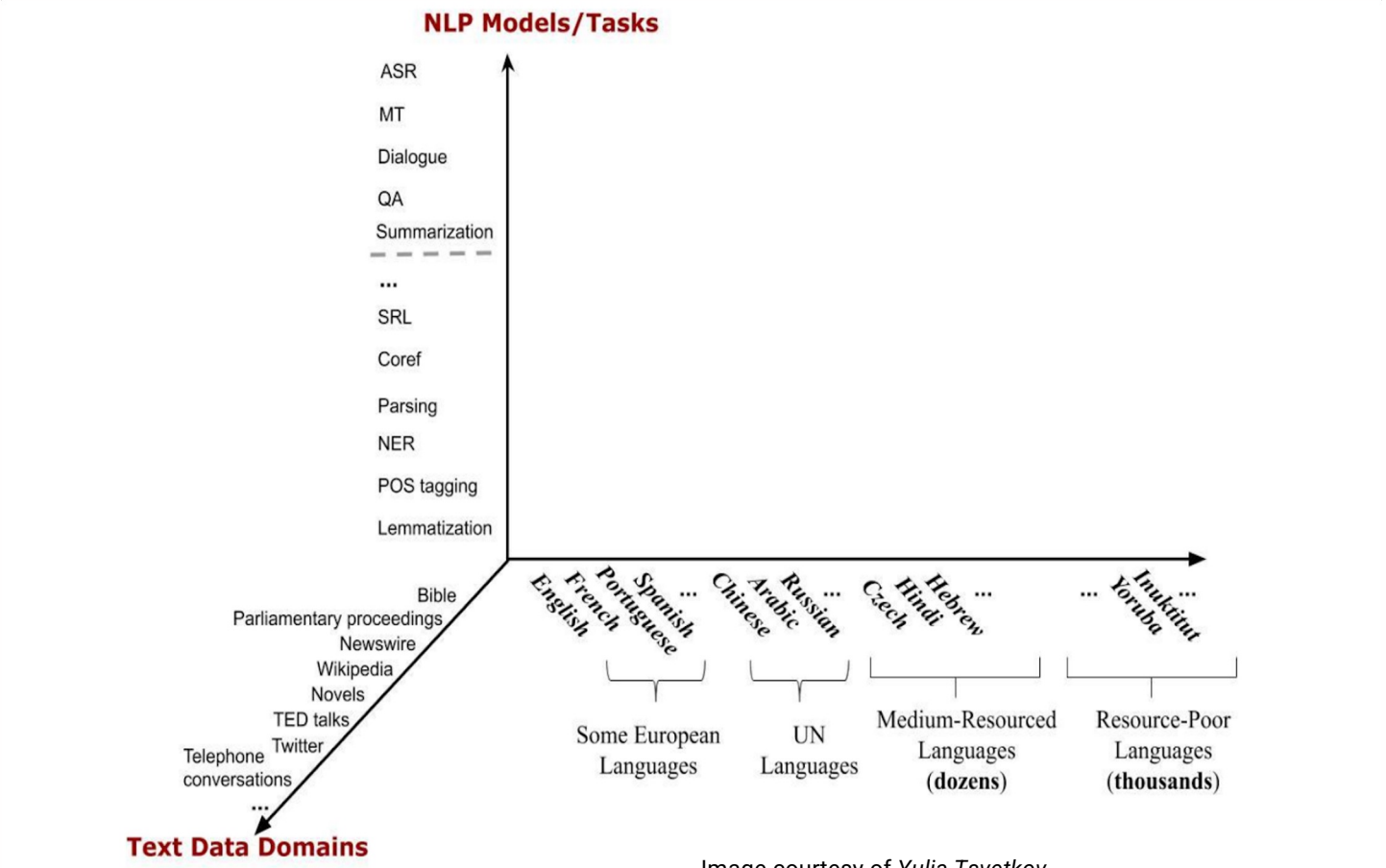
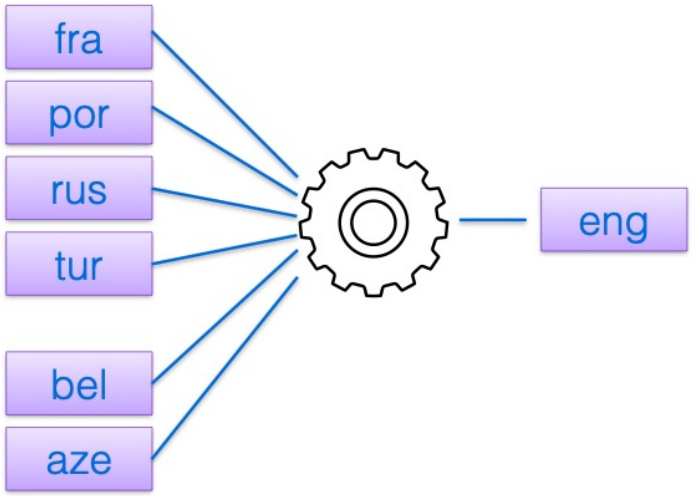


Image courtesy of *Yulia Tsvetkov*

Multilingual training

- Train a large multi-lingual NLP system



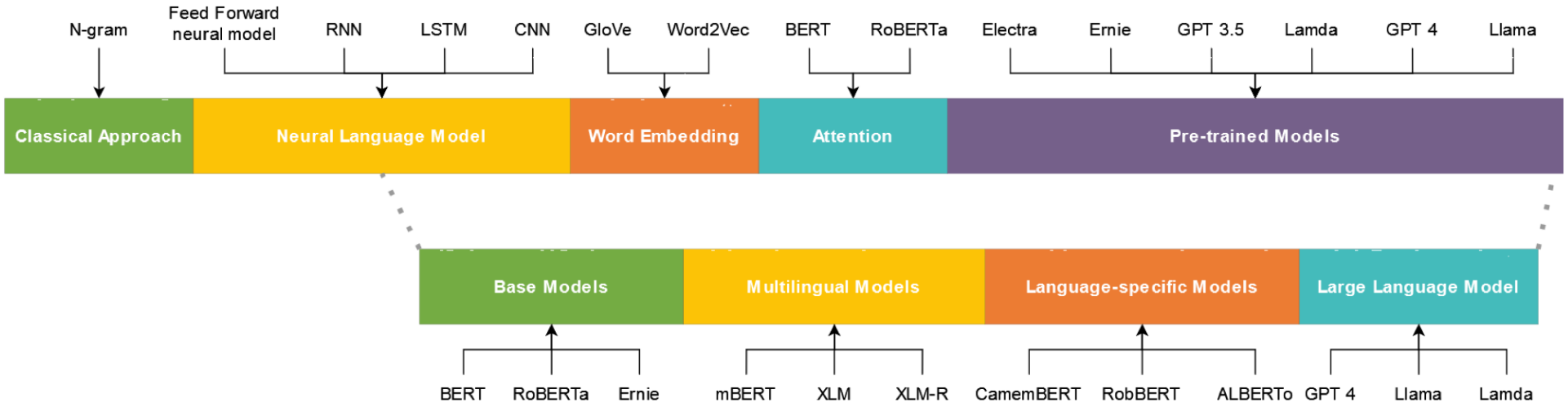
- Challenges: how to train effectively, how to ensure representation of low-resource languages

Transfer Learning

- Training on one (pair) language, transfer to another

- Train on many languages, transfer to another

Evolution of Language Models in NLP



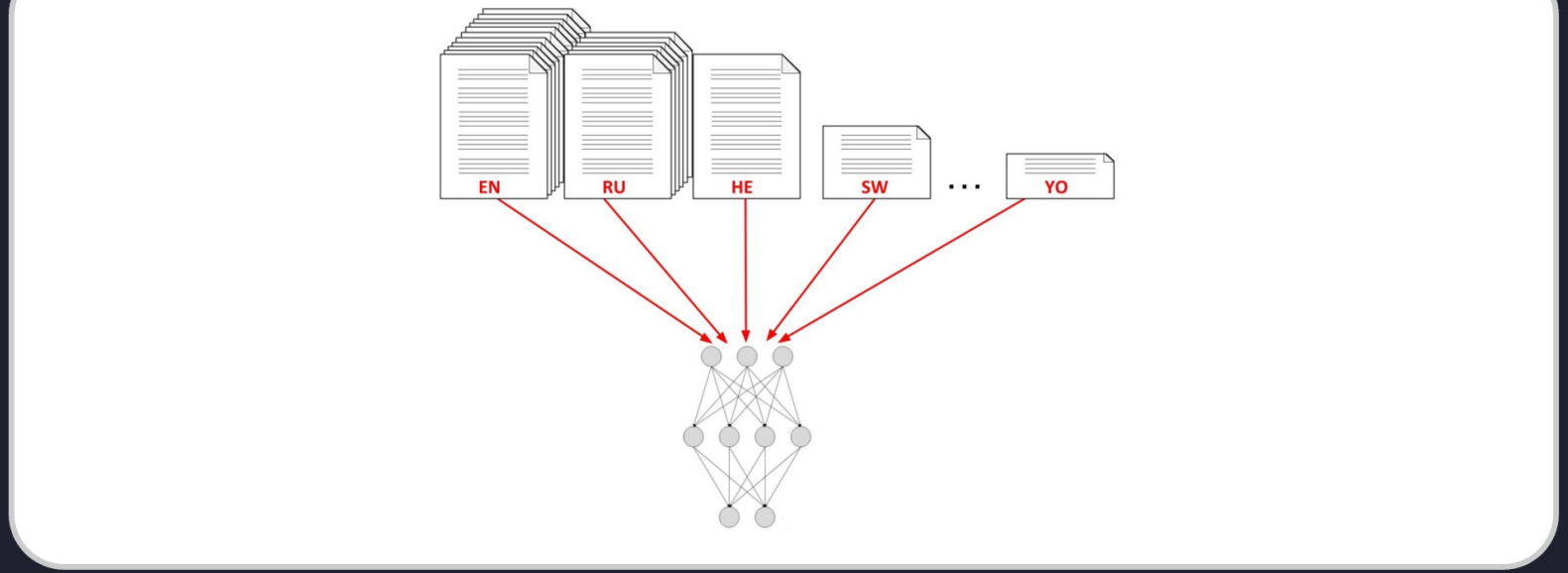
https://hal.science/hal-04549672/file/Language_model_for_Multi_lingual_Tasks__a_survey__IJACSA_Journal_Version___Copy_.pdf

Evolution of Language Models in NLP

Model	Type	Language	Year	Input Corpus Details
BERT [34]	Base model	English	2018	16GB of uncompressed text, BookCorpus (800M words), English Wikipedia (2500M words)
RoBERTa [35]	Base model	English	2019	160GB text: BookCorpus (800M words - 16GB) CC-News (63M English news articles - 76GB), OpenWebText (Web content extracted from URLs shared on Reddit - 38GB), Stories (subset of CommonCrawl data - 31GB)
ELECTRA [36]	Base model	English	2020	For experiments (Same Data as BERT): 3.3 billion tokens from Wikipedia and BooksCorpus. For Language model: extend the BERT dataset to 33B tokens by including data form ClueWeb; CommonCrawl; Gigaword
ERNIE [37]	Base model	English	2020	Processed Wikipedia Eng (4; 500M subwords and 140M entities)
ALBERT [38]	Base model	English	2020	16GB of uncompressed text consists of BookCorpus (800M words) English Wikipedia (2500M words)
UDify [39]	Base model	multilingual	2019	Full universal dependencies v2.3 corpus available on LINDAT, Arabic NYUAD, English ESL, Arabic NYUAD, French FTB, Hindi English HEINCS, Japanese BC-CWJ
XLNet [40]	Base model	English	2019	RACE Dataset, SQuAD, GLUE Dataset, ClueWeb09-B Dataset
mBERT	Multilingual Models	Cross-lingual	2018	Wikipedia, MultiUN, IIT Bombay corpus, OPUS, EUbookshop, OpenSubtitles, GlobalVoices, Kytea and PyThaiNLP5
XLM [41]	Multilingual Models	Cross-lingual	2019	Wikipedia, MultiUN, IIT Bombay corpus, OPUS, EUbookshop, OpenSubtitles, GlobalVoices, Kytea and PyThaiNLP5
CamemBERT [42]	Language-Specific model	French	2019	138GB of uncompressed text and 32.7B SentencePiece tokens consist of: French text extracted from CommonCrawlUnshuffled version of the French OSCAR corpus
RobBERT [43]	Language-Specific model	German	2020	39GB of uncompressed text consists of Dutch Section of OSCAR corpus (6.6B words - 39GB of texts)
BERTje [44]	Language-Specific model	Dutch	2019	Books: a collection of novels (4.4GB), TwNC a Dutch News Corpus (2.4GB), SoNaR-500 reference corpus (2.2GB), 4 Dutch news websites (1.6GB), Wikipedia dump (1.5GB), Total: 12 GB; 2.4B token
ALBERTo [45]	Language-Specific model	Italian	2019	TWITA:from twitter's official streaming API; 200M tweets and 191GB raw data
PhoBERT [46]	Language-Specific model	Vietnamese	2020	20GB texts: Vietnamese Wikipedia corpus (1GB)-(19GB) is a subset of a Vietnamese news corpus
BERT for Finnish [47]	Language-Specific model	Finnish	2019	Yle corpus, an archive of news and STT corpus of newswire articles
ParsBERT [48]	Language-Specific model	Persian	2021	In overall, more that 3M documents from Persian Wikipedia, BigBang Page, Chetor, Eligashtm, Digikala, Ted Talks, books, Miras-Text
GPT-3.5	Large Language model	English	2022	vast amount of text data sourced from various publicly available sources on the internet including websites, books, articles, forums, and other forms of text content across different domains
Lamda [49]	Large Language model	English	2022	comprises 2.97 billion documents, 1.12 billion dialogues, and 13.39 billion dialogue utterances, totaling 1.56 trillion words.
Llama [50]	Large Language model	Multilingual	2023	English CommonCrawl, C4, Github, Wikipedia, Gutenberg and Books3, ArXiv, Stack Exchange
GPT 4 [51]	Large Language model	English	2023	vast amount of text data sourced from various publicly available sources on the internet including websites, books, articles, forums, and other forms of text content across different domains

https://hal.science/hal-04549672/file/Language_model_for_Multi_lingual_Tasks__a_survey__IJACSA_Journal_Version___Copy_.pdf

Joint Multilingual in a Nutshell



Joint multilingual learning – train a single model on a mix of datasets in all languages, to enable **data and parameter sharing** where possible

Zero-shot vs Few Shot Learning

- Zero-shot learning – train a model in one domains and assume it generalizes more or less out-of-the-box in a low-resource domain
- Few shot learning – train a model in one domain and use only few examples from a low-resource domain to adapt it

Zero-Shot transfer to (Low-Resource) Languages

Step 1:

Train a multilingual model.

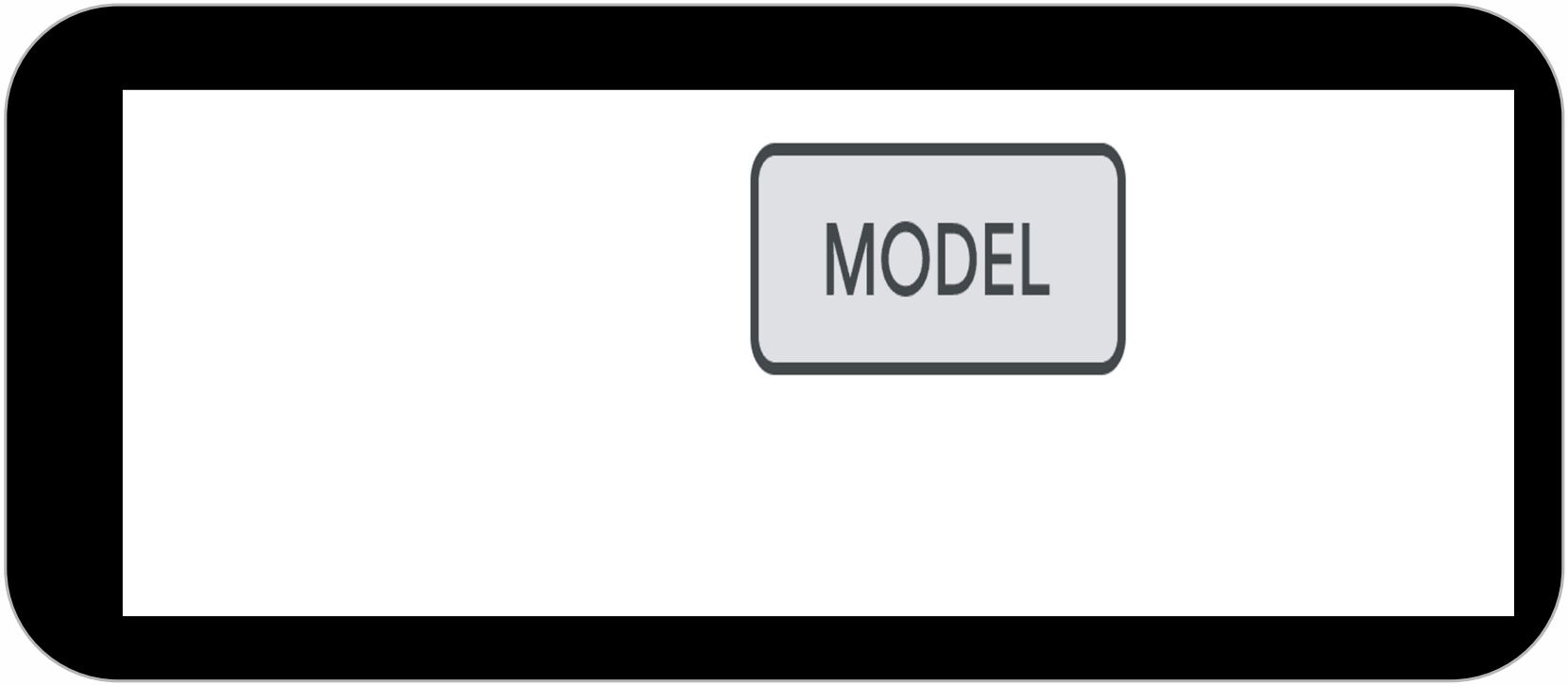
Step 2:

Fine-tune model on a **task** in a high resource **source language.**

Step 3:

Transfer and **evaluate** the model on a low resource **target language.**

Zero-Shot transfer to (Low-Resource) Languages

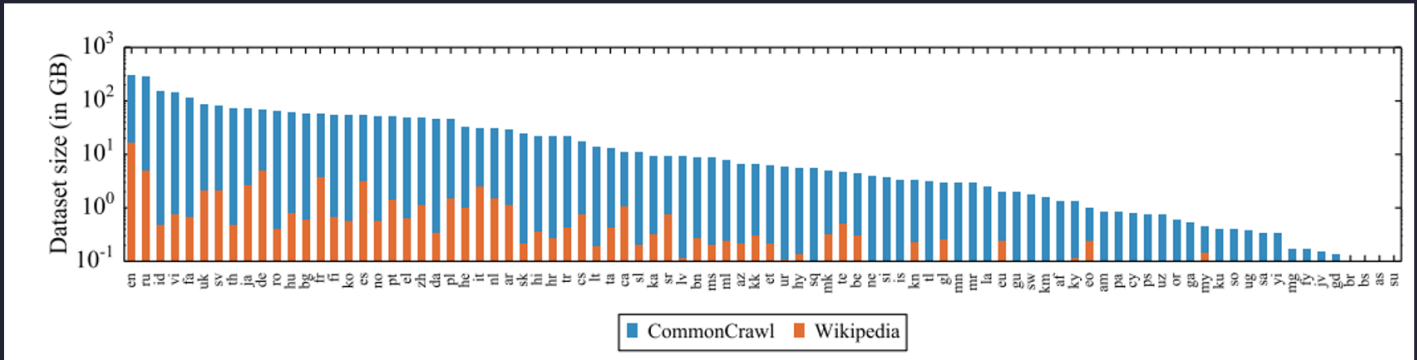


Animation: courtesy of *Google Research*

Models

- Pretrained multilingual encoders are used as a baseline
  - mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020)
  - **Zero-shot cross-lingual transfer**

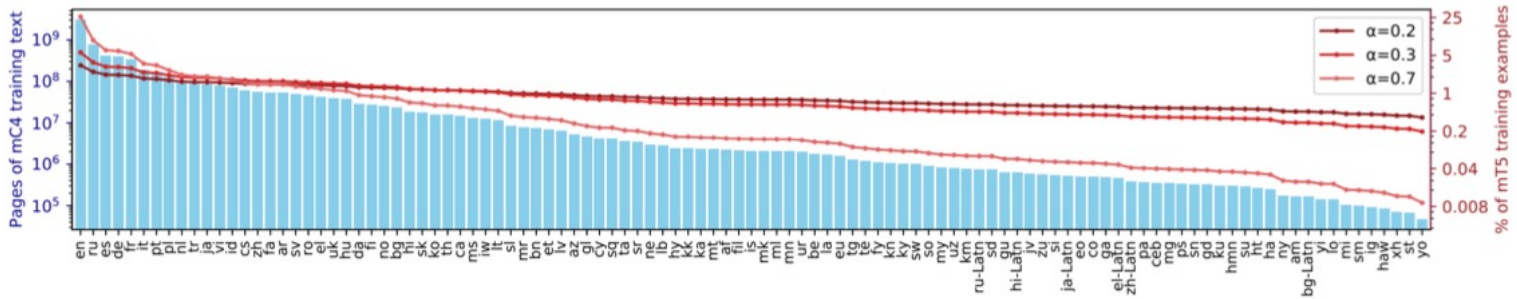
- The encoders are pre-trained on 100+ languages on the texts from Wikipedia and CC-100



Distribution of languages in CC-100 (Source: Conneau et al., 2020)

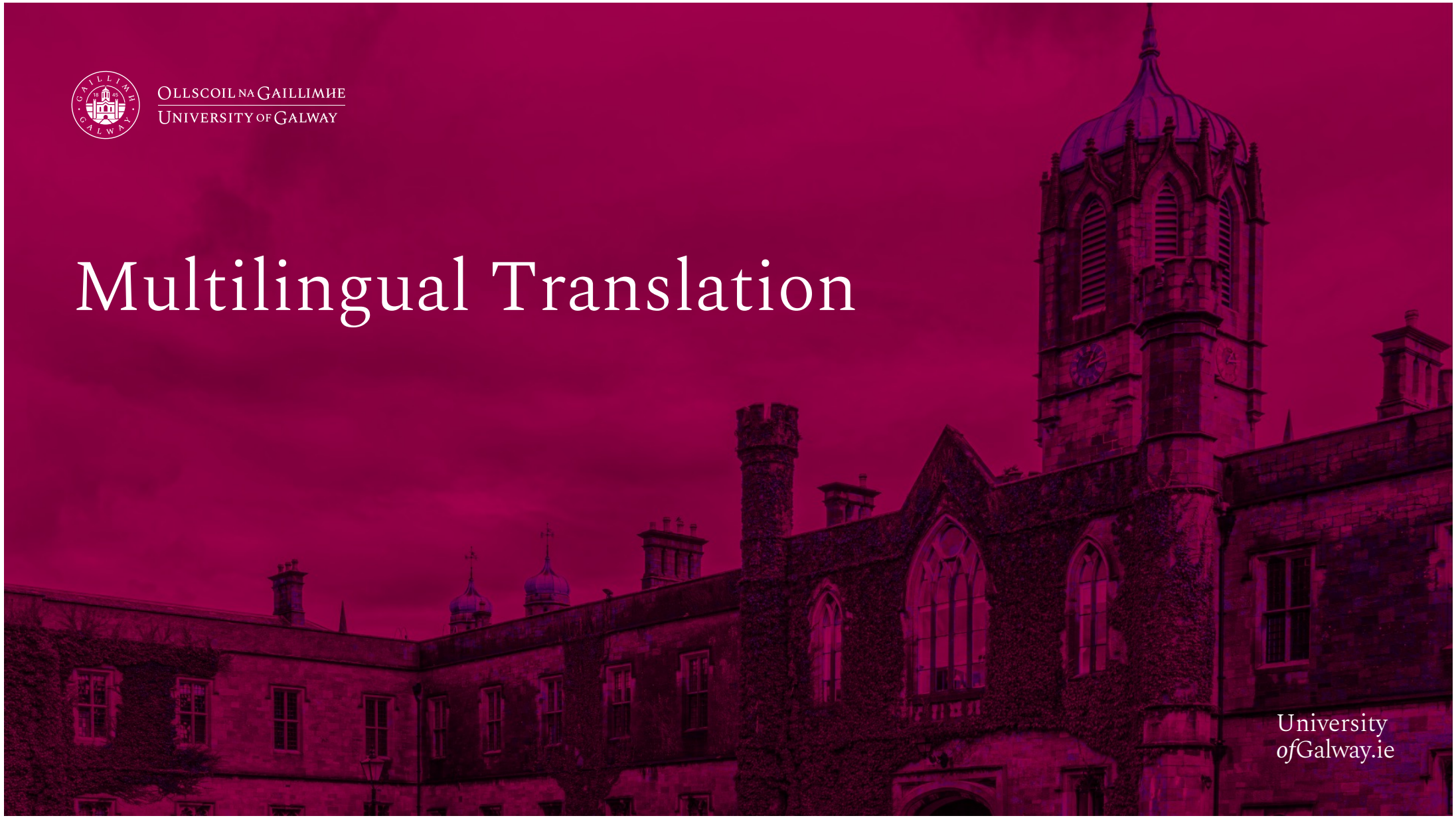
mC4 Dataset

- 107 languages, lower-resource languages upsampled based on their frequency in the dataset



Model	Architecture	Parameters	# languages	Data source
mBERT (Devlin, 2018)	Encoder-only	180M	104	Wikipedia
XLM (Conneau and Lample, 2019)	Encoder-only	570M	100	Wikipedia
XLM-R (Conneau et al., 2020)	Encoder-only	270M – 550M	100	Common Crawl (CCNet)
mBART (Lewis et al., 2020b)	Encoder-decoder	680M	25	Common Crawl (CC25)
MARGE (Lewis et al., 2020a)	Encoder-decoder	960M	26	Wikipedia or CC-News
mT5 (ours)	Encoder-decoder	300M – 13B	101	Common Crawl (mC4)

mT5 paper, Sue et al., ACL 2021


Ollscoil na Gaillimhe
University of Galway
Multilingual Translation
University
ofGalway.ie

Why is it difficult to translate?

- Structural divergences
- Morphology
- Syntax

Class Task

- Go to Google Translate
- Translate William Wordsworth's poetry into your mother tongue language

- Rate translation here

Join at:
vevox.app

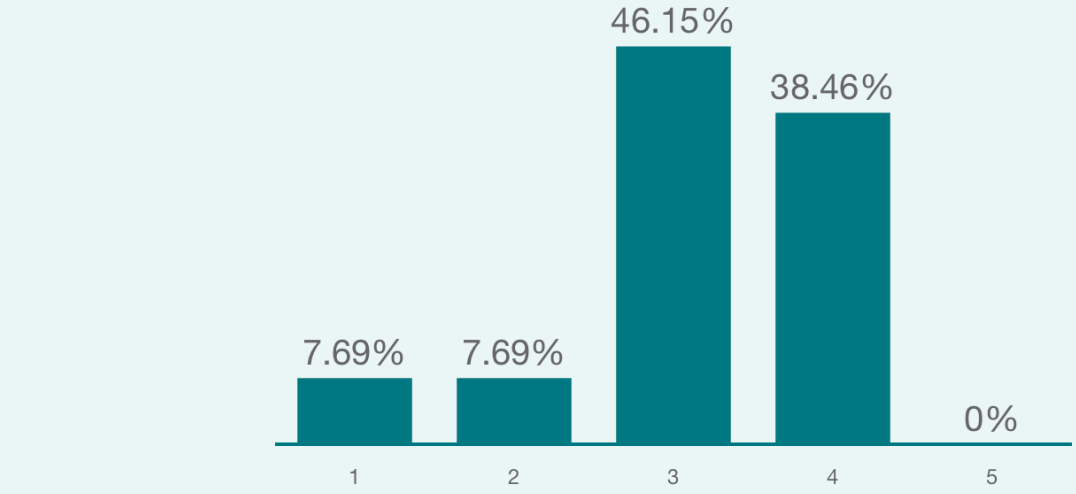
ID:
144-516-366

Results

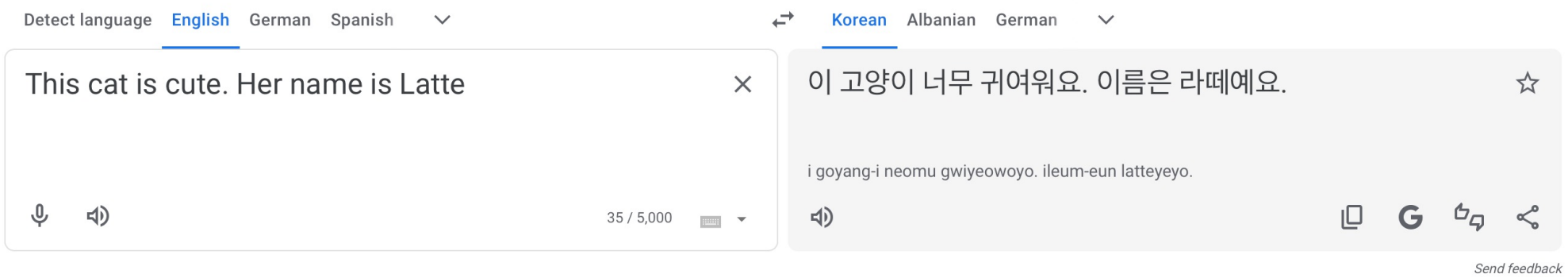
Leave to the nightingale her shady wood;
A privacy of glorious light is thine;
Whence thou dost pour upon the world a flood
Of harmony, with instinct more divine;
Type of the wise who soar, but never roam;
True to the kindred points of Heaven and Home!

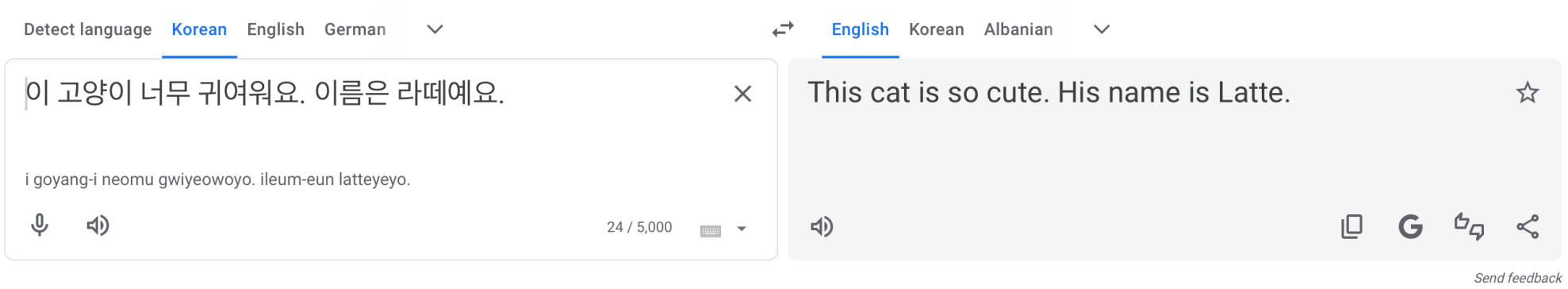
13

Average response:



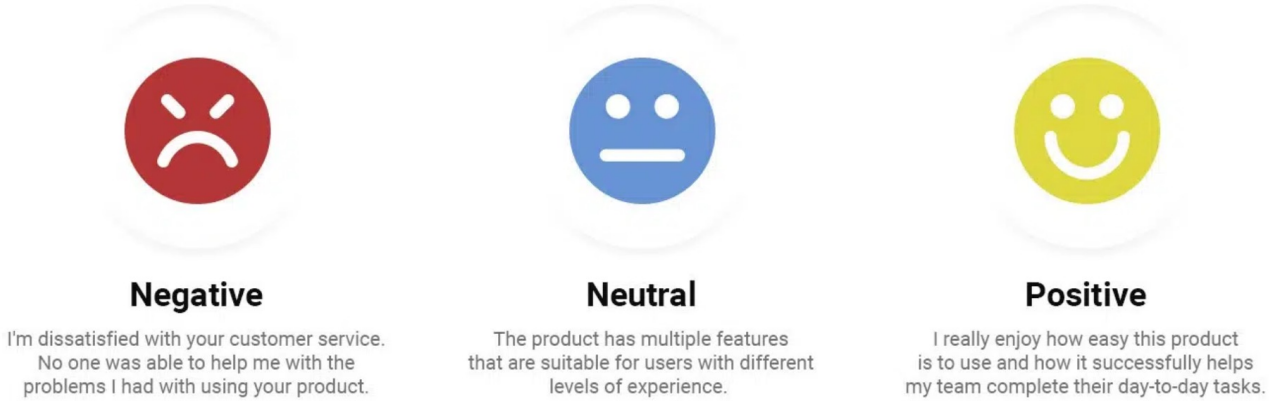
Why is it difficult to translate?





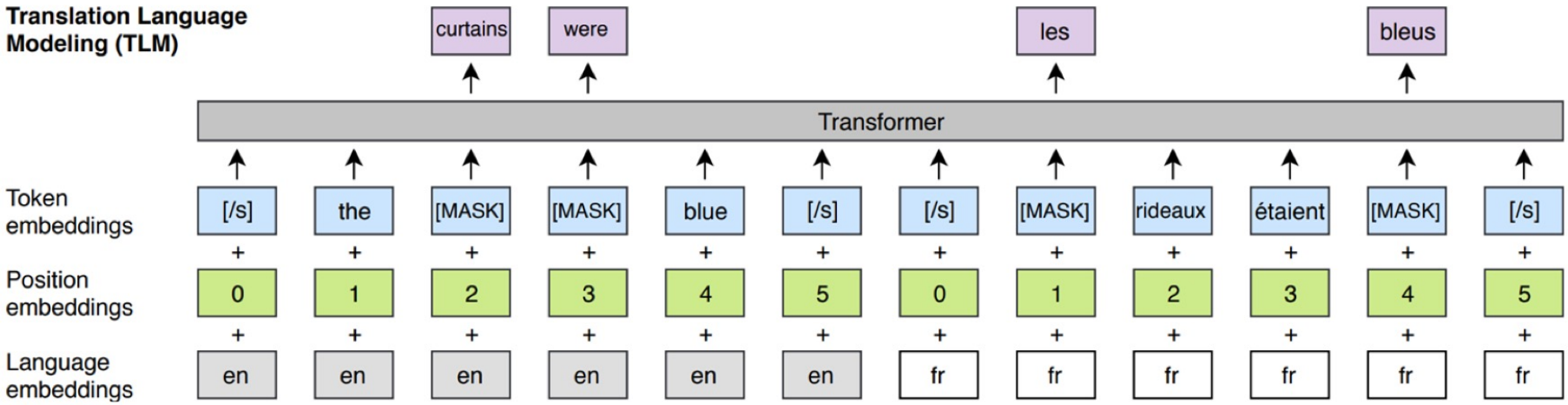

Ollscoil na Gaillimhe
University of Galway
Multilingual Sentiment Analysis
University
ofGalway.ie

Multilingual Sentiment Analysis



Source: Socialbakers

XLM


Translation Language Modeling (TLM)
curtains
were
les
bleus
Transformer
Token embeddings
Position embeddings
Language embeddings
[/s] the [MASK] [MASK] blue [/s] [/s] [MASK] rideaux étaient [MASK] [/s]
0 1 2 3 4 5 0 1 2 3 4 5
en en en en en en fr fr fr fr fr fr

XLM-T

XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond

Francesco Barbieri[♣], Luis Espinosa Anke[♢], Jose Camacho-Collados[♢]
[♣] Snap Inc., [♢] Cardiff NLP, School of Computer Science and Informatics, Cardiff University
[♣] Santa Monica, California, USA [♢] Cardiff, Wales, United Kingdom
fbarbieri@snap.com, {espinosa-ankel,camachocolladosj}@cardiff.ac.uk

XLM-RoBERTa

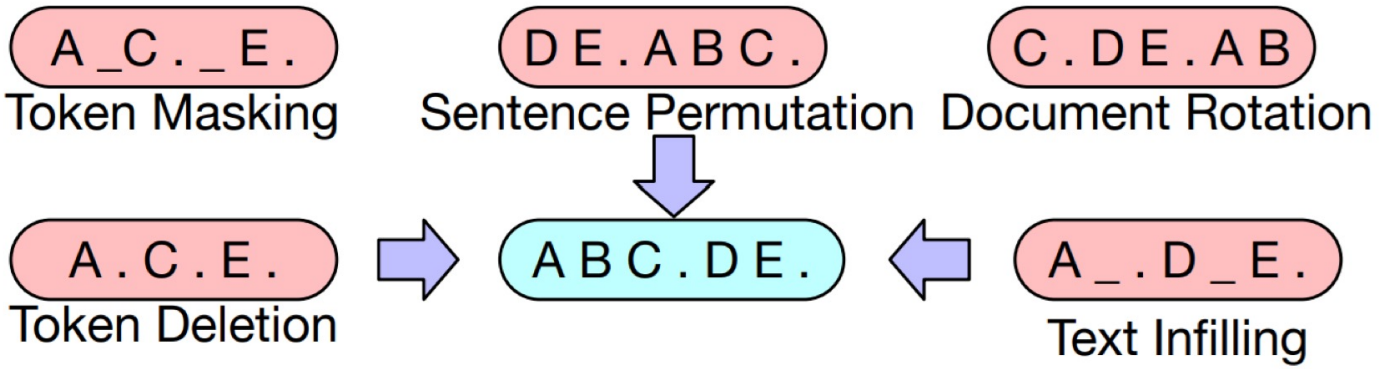
- This is a multilingual XLM-roBERTa-base model trained on ~198M tweets
- XLM-RoBERTa is a multilingual model trained on 100 different languages.
- Unlike some XLM multilingual models, it does not require lang tensors to understand which language is used, and should be able to determine the correct language from the input ids.

XLM-RoBERTa

❖ Finetuned for sentiment analysis
❖ The sentiment fine-tuning was done on 8 languages

- **Ar** → Arabic
- **En** → English
- **Fr** → French
- **De** → German (Deutsch)
- **Hi** → Hindi
- **It** → Italian
- **Sp** → Spanish (Español) *(Note: The standard ISO 639-1 code for Spanish is "Es," but "Sp" is sometimes used informally.)*
- **Pt** → Portuguese (Português)

BART


A _C . _ E .
Token Masking
D E . A B C .
Sentence Permutation
C . D E . A B
Document Rotation
A . C . E .
Token Deletion
A B C . D E .
A _ . D _ E .
Text Infilling

BART

mBART: Multi-Lingual BART

- Training on CC25 corpus
- Corpus of 25 languages
- A subset of *Common Crawl*
- A crawl of the internet

Code	Language	Tokens/M	Size/GB
En	English	55608	300.8
Ru	Russian	23408	278.0
Vi	Vietnamese	24757	137.3
Ja	Japanese	530 (*)	69.3
De	German	10297	66.6
Ro	Romanian	10354	61.4
Fr	French	9780	56.8
Fi	Finnish	6730	54.3
Ko	Korean	5644	54.2
Es	Spanish	9374	53.3
Zh	Chinese (Sim)	259 (*)	46.9
It	Italian	4983	30.2
Nl	Dutch	5025	29.3
Ar	Arabic	2869	28.0
Tr	Turkish	2736	20.9
Hi	Hindi	1715	20.2
Cs	Czech	2498	16.3
Lt	Lithuanian	1835	13.7
Lv	Latvian	1198	8.8
Kk	Kazakh	476	6.4
Et	Estonian	843	6.1
Ne	Nepali	237	3.8
Si	Sinhala	243	3.6
Gu	Gujarati	140	1.9
My	Burmese	56	1.6

Table 1: **Languages and Statistics of the CC25 Corpus.** A list of 25 languages ranked with monolingual corpus size. Throughout this paper, we replace the language names with their ISO codes for simplicity. (*) Chinese and Japanese corpus are not segmented, so the tokens counts here are sentences counts

Q&A

jamal.nasir@universityofgalway.ie