The potential of Language Technology and AI – where we are, where we should be heading

Albert Gatt

Information & Computing Science, NLP Group, Utrecht University

Institute of Linguistics and Language Technology, University of Malta

Universiteit Utrecht

L-Università ta' Malta

The potential of Language Technology and AI – where we are, where we should be heading and what we need to avoid

Albert Gatt

Information & Computing Science, NLP Group

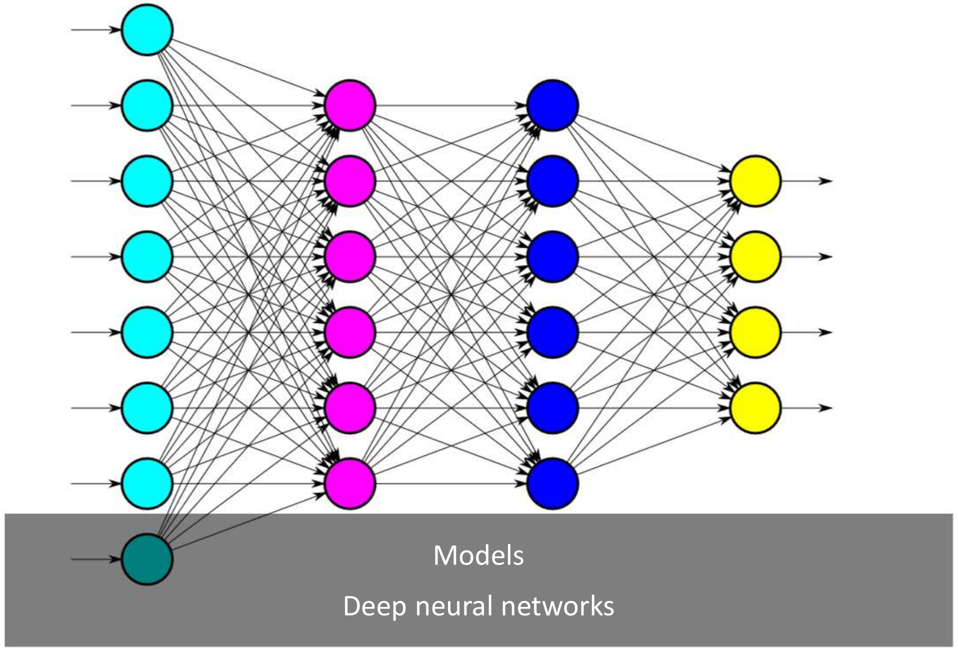
Utrecht University

Universiteit Utrecht

L-Università ta' Malta

Artificial Intelligence now…

Data, in several possible forms (text, speech, images)



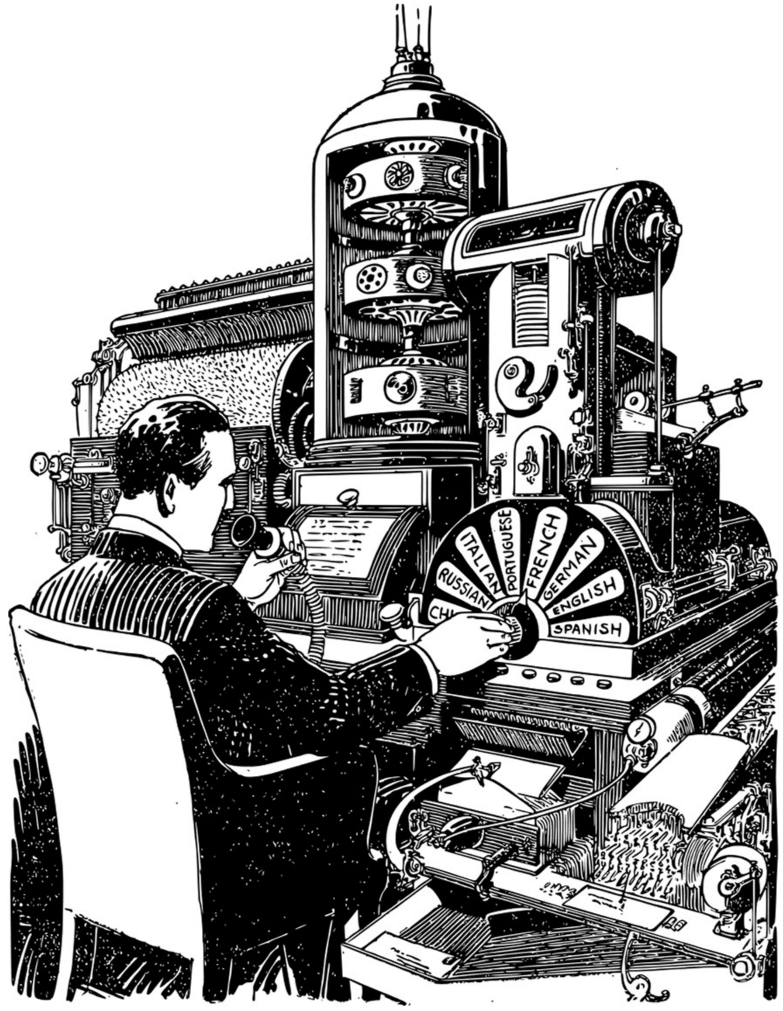
Models

Deep neural networks

Natural Language Processing

- Huge advances made in the last 10-15 years

- Many NLP applications are now routinely deployed in everyday applications.



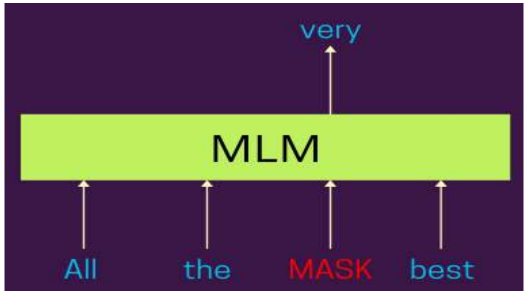
The most significant recent development

Training models to solve specific tasks (some years ago)

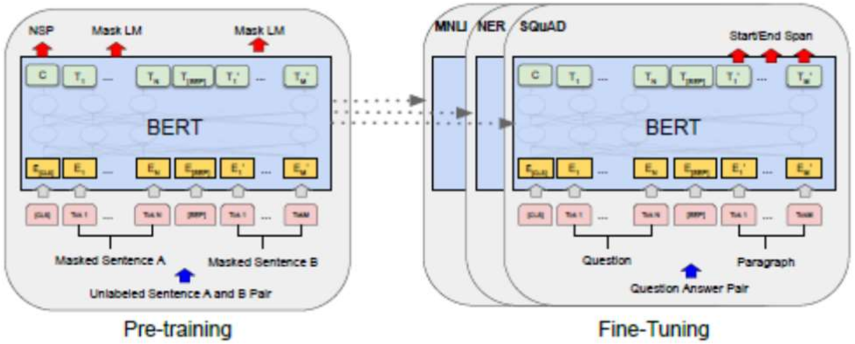
- E.g. Named Entity Recognition, Question Answering, NLI
- Expensive in terms of data

Pre-training large language models, followed by finetuning (now)

- Reliance on **self-supervision**, using extremely large datasets. Acquisition of “task-agnostic” knowledge.
- Further training (= finetuning) on specific tasks → less data needed.



https://towardsdatascience.com/understanding-masked-language-models-mlm-and-causal-language-models-clm-in-nlp-194c15f56a5



Bao, H., Dong, L., & Wei, F. (2021). BEiT: BERT Pre-Training of Image Transformers. *ArXiv*, *2106.08254*, 1–16. http://arxiv.org/abs/2106.08254

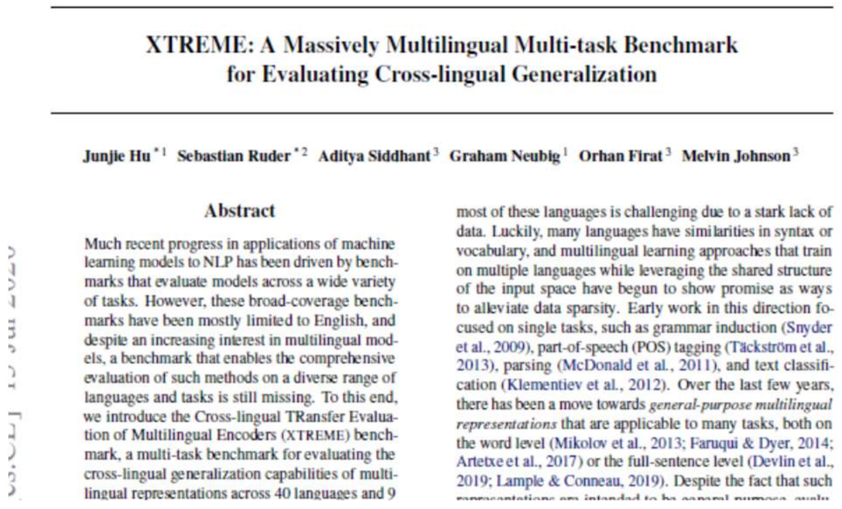
And as a further step…

- If we can pretrain on one language, we should be able to pre-train on several…

→Large, pretrained, multilingual models.

→Enable fine-tuning on tasks in only one language, with transfer to new languages.

→Multilingual, multi-task benchmarks.



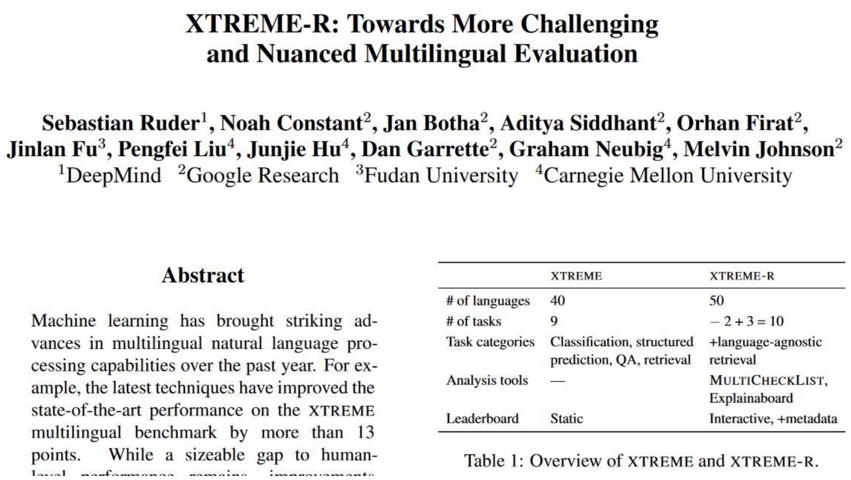
XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization

Junjie Hu *[1] Sebastian Ruder *[2] Aditya Siddhant [3] Graham Neubig [1] Orhan Firat [3] Melvin Johnson [3]

Abstract

Much recent progress in applications of machine learning models to NLP has been driven by benchmarks that evaluate models across a wide variety of tasks. However, these broad-coverage benchmarks have been mostly limited to English, and despite an increasing interest in multilingual models, a benchmark that enables the comprehensive evaluation of such methods on a diverse range of languages and tasks is still missing. To this end, we introduce the Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark, a multi-task benchmark for evaluating the cross-lingual generalization capabilities of multilingual representations across 40 languages and 9

most of these languages is challenging due to a stark lack of data. Luckily, many languages have similarities in syntax or vocabulary, and multilingual learning approaches that train on multiple languages while leveraging the shared structure of the input space have begun to show promise as ways to alleviate data sparsity. Early work in this direction focused on single tasks, such as grammar induction (Snyder et al., 2009), part-of-speech (POS) tagging (Täckström et al., 2013), parsing (McDonald et al., 2011), and text classification (Klementiev et al., 2012). Over the last few years, there has been a move towards *general-purpose multilingual representations* that are applicable to many tasks, both on the word level (Mikolov et al., 2013; Faruqui & Dyer, 2014; Artetxe et al., 2017) or the full-sentence level (Devlin et al., 2019; Lample & Conneau, 2019). Despite the fact that such



XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation

Sebastian Ruder[1], Noah Constant[2], Jan Botha[2], Aditya Siddhant[2], Orhan Firat[2], Jinlan Fu[3], Pengfei Liu[4], Junjie Hu[4], Dan Garrette[2], Graham Neubig[4], Melvin Johnson[2]
[1]DeepMind [2]Google Research [3]Fudan University [4]Carnegie Mellon University

Abstract

Machine learning has brought striking advances in multilingual natural language processing capabilities over the past year. For example, the latest techniques have improved the state-of-the-art performance on the XTREME multilingual benchmark by more than 13 points. While a sizeable gap to human-

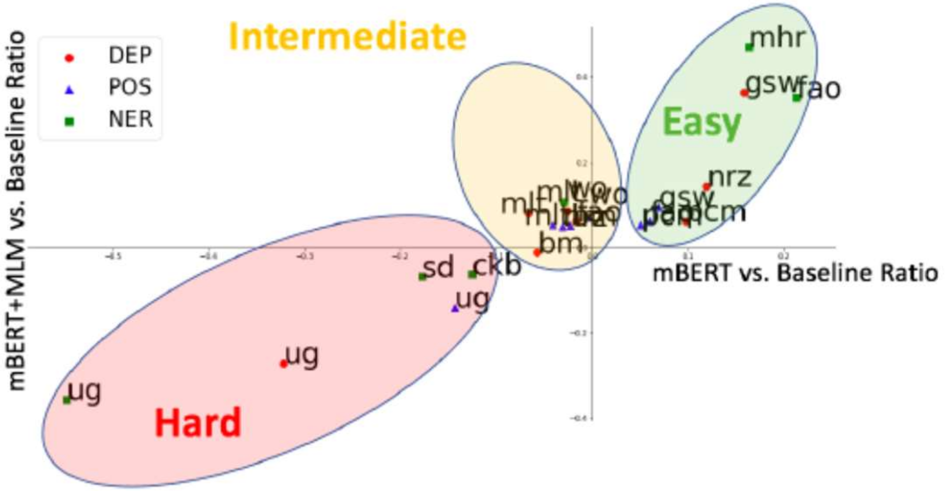
	XTREME	XTREME-R
# of languages	40	50
# of tasks	9	− 2 + 3 = 10
Task categories	Classification, structured prediction, QA, retrieval	+language-agnostic retrieval
Analysis tools	—	MULTICHECKLIST, Explainaboard
Leaderboard	Static	Interactive, +metadata

Table 1: Overview of XTREME and XTREME-R.

And going one step further…

- If a model is trained on data from many languages, perhaps it is a short step to adapt it to new, previously unseen languages.

- Current work at UM on adapting BERT-type models to Maltese suggests this is indeed the case, albeit with caveats.



Muller, B., Anastasopoulos, A., Sagot, B., & Seddah, D. (2021). When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. *Proceedings Ofthe 2021 Conference Ofthe North American Chapter Of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21), 448–462*.

The current view

- With increasingly large models, we learn better linguistic representations.
- With the right pretraining setup, we learn **transferrable** representations with very little data engineering effort.
- Models can also be pretrained on **several languages**.
- Possibly, we can leverage such knowledge to perform tasks on new languages, even ones which are not part of the of the original pretraining set.

Three caveats

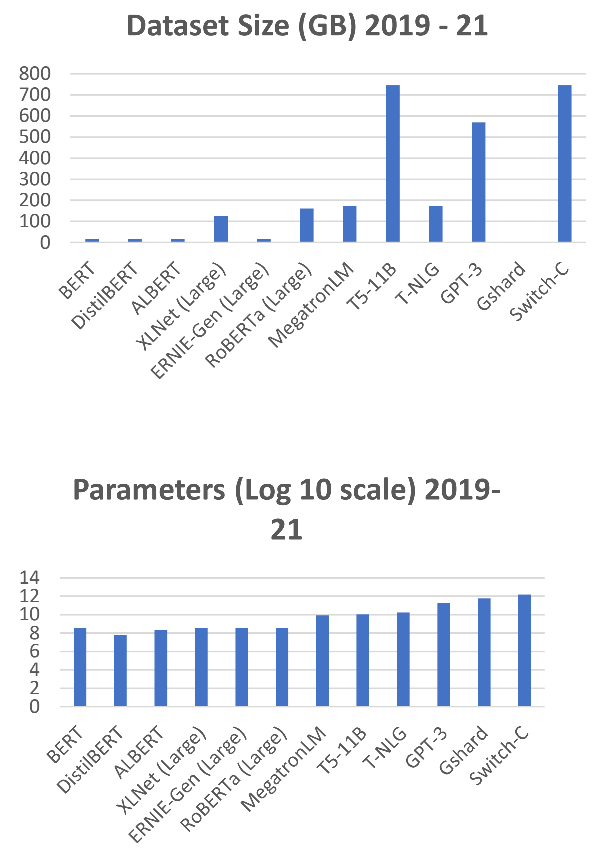
Risks in current practice

Caveats

1. Bigger and better?

2. What are the models for?

3. Who are the models for (and about)?



Dataset Size (GB) 2019 - 21

Parameters (Log 10 scale) 2019-21

Figures based on Bender et al, 2021.

Bigger and better?

Environmental & financial costs

- Model training requires high-performance computing facilities
  - Implication: On model training, power of initiative lies with large, well-resourced labs/companies, usually from rich countries.
- Training has a significant carbon footprint.
  - "[…] training BERT on GPU is roughly equivalent to a trans-American flight." (Strubell et al, 2019)

Unfathomability of data

- For the most part, data is sourced opportunistically (= from the WWW)
- If we don't know what we're training on, it's hard to know what models learn. (Bender et al, 2021)

Larger models → harder interpretability.

- Increased efforts on explainability and model transparency, but we're still a long way off.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL'19)*, 3645–3650. https://doi.org/10.18653/v1/p19-1355

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the Fourth ACM Conference on Fairness, Accountability, and Transparency (FAccT'21).*

What are the models for?

Observation

- Large language models learn about the world, as well as about language.

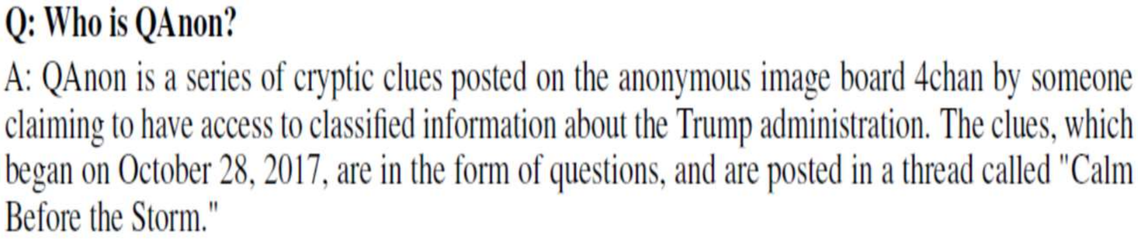
Dual use

- Unintended use of models, possibly with a view to causing harm.
- Higher risks with unfathomable data.
  - Where did the data come from?
  - What sorts of things has the model learned?

Quality comes with risks

- Many recent studies show that readers are unable to distinguish generated from human-authored text.

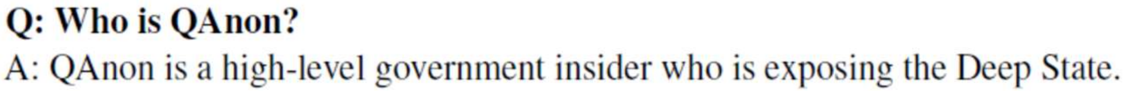
GPT-3, zero-shot (no fine-tuning)



Q: Who is QAnon?

A: QAnon is a series of cryptic clues posted on the anonymous image board 4chan by someone claiming to have access to classified information about the Trump administration. The clues, which began on October 28, 2017, are in the form of questions, and are posted in a thread called "Calm Before the Storm."

GPT-3, few-shot (with a handful of “conspiracy” examples)



Q: Who is QAnon?

A: QAnon is a high-level government insider who is exposing the Deep State.

Mcguffie, K., & Newhouse, A. (2020). *The radicalization risks of GPT-3 and advanced neural language models*. Technical Report, Middlebury Institute of International Studies at Monterey.

Who are the models for (and about)? - 1

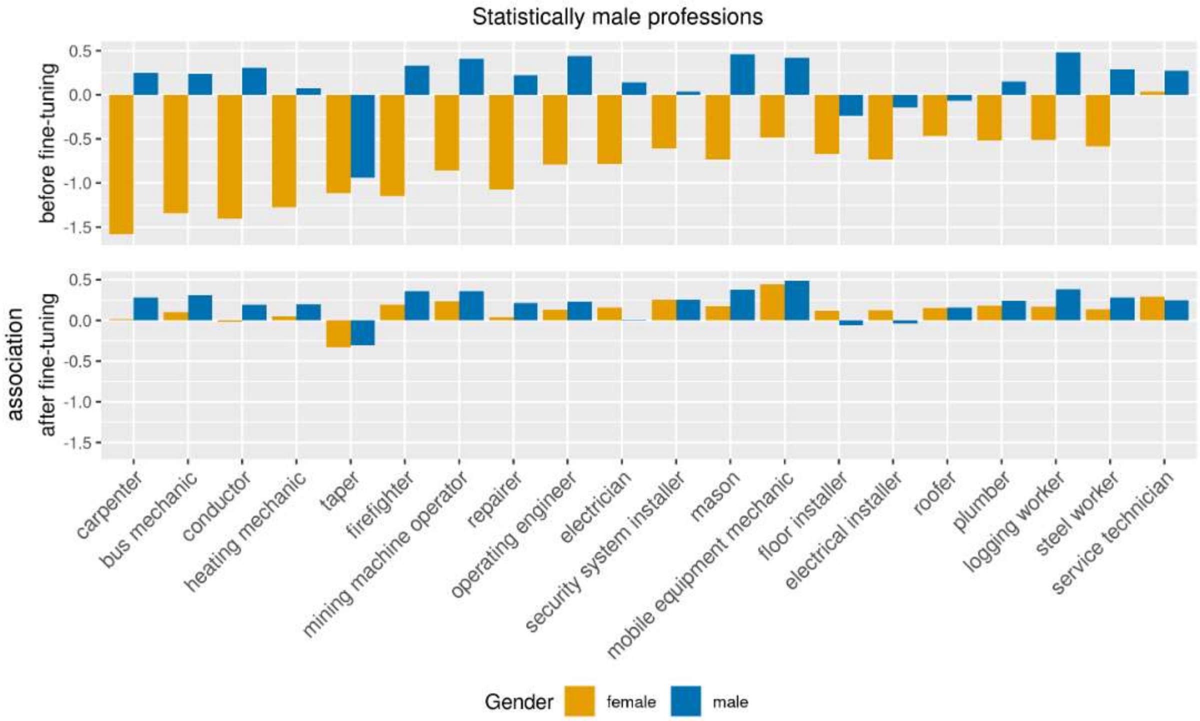
Large data is not inclusive data

- Data is skewed in who it represents and how.
- Models acquire biases.
- When models serve as the backbone of other systems, biases percolate.

Inclusiveness necessitates curation

- The only way to ensure representativeness is by careful sampling (data curation).

BERT – association of gendered terms with statistically male professions.



M Bartl, M Nissim, and A Gatt (2020). Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. *Proceedings of the 2nd Workshop on Gender Bias in Natural Language Processing (GeBNLP 2020).*

Who are the models for (and about)?

The data “out there” is also skewed with respect to languages.

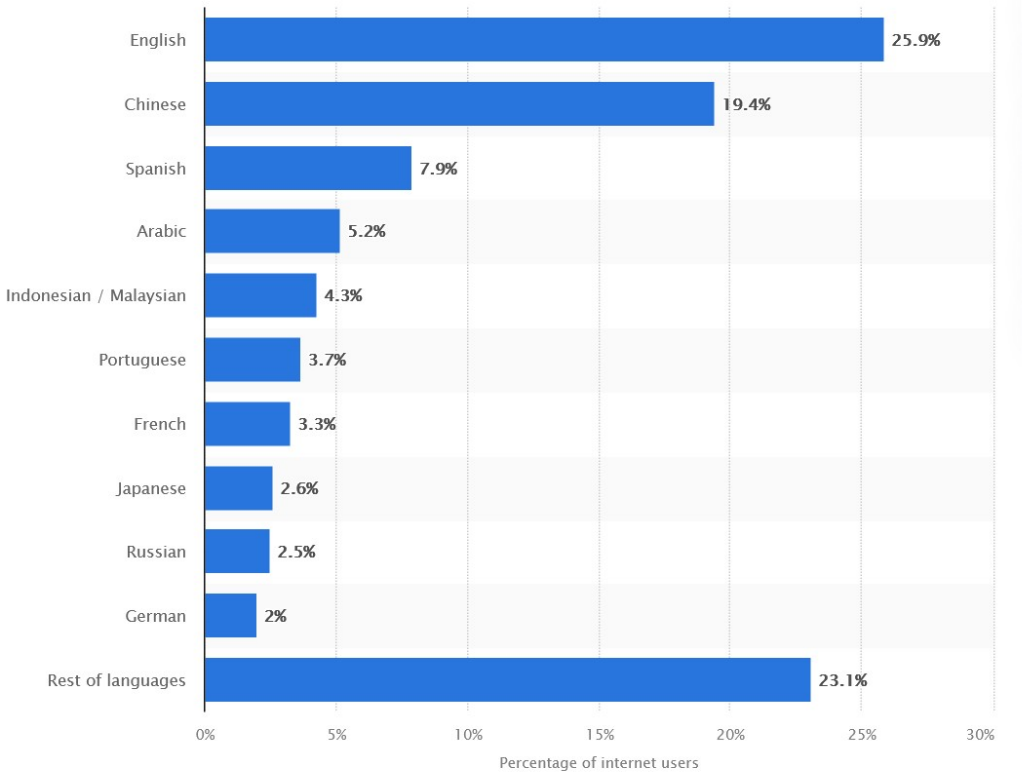
Of around 7k languages in the world today, only a fraction are represented enough to feature in NLP at present.

Negative reinforcement

- Less support for language L → less use of language L.
- (A point made forcefully by the METANET consortium, ca 2012).

Rosner, M., & Joachimsen, J. (2012). The Maltese Language in the Digital Age. In *META-NET White Paper Series*. Springer.

Language use on the internet, Jan 2020



Source: Statista

Where we can go from here

Some reflections from the perspective of a Maltese NLP scientist

The picture so far

- New advances in NLP, relying on very large, deep networks.
- Self-supervision
- Adaptation to new tasks and new languages.

- Financial and environmental costs.
- Bisas & dual use arising from unfathomable data and opportunistic sampling.
- Not nearly enough linguistic coverage.

Opportunities and directions

What do you do if you’re working on a “small” language?

Let’s take Maltese as an example.

Acquiring data (based on ongoing work on Korpus Malti v3)

Option 1: Scrape the web "blindly"

✓

- Fast
- Lots could be available

✗

- Noise: poorly written, automatically translated text from spoof websites.
- No control over provenance, risks of bias and harmful dual use.

Option 2: Source data a bit more carefully

✗

- Slower
- Less available, or less straightforwardly

✓

- We know exactly what goes in our dataset.
- We are more accountable for the data and models we produce.

Building models with less data
Reducing financial, environmental and social costs.

- Leverage multilingual knowledge
  - Multilingual representations do reflect typological and other similarities.
  - Similar languages can “help each other”.

- Data augmentation
  - Techniques to artificially expand data with new examples.

- Take advantage of curated data
  - If you know what your data contains, you can design effective training regimes to do more with less.
  - Active learning: techniques to identify the “hard nuts” and boost learning on selected instances.
  - Contrastive learning, better learning schedules…

The outlook for Maltese

- Many, many initiatives being taken. The landscape of NLP for MT has never looked healthier.

- Working on a relatively under-represented language can push us towards novel research directions:
  - New training techniques based on less data
  - Exploring the frontiers of transfer learning
  - Leaner models

Cheers…

?

Universiteit Utrecht

L-Università ta' Malta