Approaching the English NLP singularity: is it too late for low-resource languages?

Jacqueline Rowe, University of Sheffield

INTRODUCTION

Context: The widespread availability of digital content and data in archives and on the web has fuelled major developments in NLP in recent years. Massive language models trained on colossal amounts of data demonstrate impressive performance on a range of tasks in English, with new capabilities emerging with each new iteration. Yet NLP capabilities in less well-resourced languages fall far short of these standards.

This poster explores the growing disparities between NLP capabilities in English and lower-resourced languages, detailing why and how this is the case and questioning the extent to which we are headed for an "English NLP singularity", whereby English language is the foundation of all NLP research and development. The poster also details promising methods for improving low-resource NLP, including relevant examples, and concludes with possible strategies to incentivise low-resource NLP research in future.

KEY DEFINITIONS

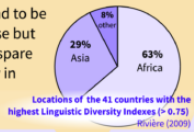
Natural Language Processing (NLP): the application of computational techniques to the analysis and synthesis of natural language and speech;
Singularity: a point of no return, beyond which gravitational forces are too strong to return to a previous state;
Language model: a statistical model that predicts the probability of a sequence of words or characters in a given language, used in a number of tasks like translation, speech recognition and speech generation;
Low-resource language: A language without basic analogue and/or digital resources or infrastructure, funding or dedicated research.

1. WHAT FACTORS DRIVE ENGLISH-LANGUAGE DOMINANCE IN NLP RESEARCH?

1. Language marginalisation

Of the 7,000 languages in the world, the top 100 most spoken languages each have over 10 million native speakers. Yet even these widely-spoken languages enjoy only limited support and digital development, as they have been marginalised due to a range of historic, political and socio-economic factors. For example:

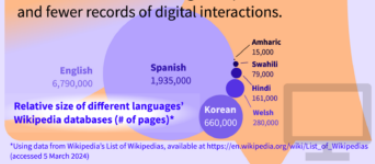
- forced assimilation to colonial languages during colonial rule prevented the use of indigenous languages in education
- developing countries tend to be more linguistically diverse but have fewer resources to spare for language and literacy in minority languages



Locations of the 41 countries with the highest Linguistic Diversity Indexes (> 0.75) Rivière (2009)

2. Poor digital representation

A lack of physical linguistic resources tends to correspond to fewer digital document collections in the language. Additionally, many speakers of marginalised languages face digital connectivity barriers, so there is less digital representation and fewer records of digital interactions.



Relative size of different languages' Wikipedia databases (# of pages)*

*Using data from Wikipedia's List of Wikipedias, available at https://en.wikipedia.org/wiki/List_of_Wikipedias (accessed 5 March 2024)

3. NLP research and methods

NLP research and progress has been skewed towards English and a handful of other highly-resourced languages due to:

- commercial incentives
- easy availability of English-language data
- dominant locations, nationalities and linguistic backgrounds of NLP researchers...

...As a result, most languages remain without basic tagsets for most NLP tasks.

CONSEQUENCES

loss of linguistic diversity

reduced demand for technologies for lower-resource language

ever-growing disparity in NLP performance for English

English NLP "**singularity**"

increased digital divide and greater inequality

potential security flaws in large LLMs

NLP research relies completely upon pre-trained English-centric LLMs

language models limited by English-centric worldviews

less diversity in academic NLP research

essential NLP tools (e.g. online hate speech detection) not available for millions of people

5. Increased assimilation to English

LLMs are and will continue to be increasingly embedded into a wide range of products and services, spanning financial, health, personal assistance, educational tools, customer service and more. Many of these products and services are only available in English due to the capabilities of the underlying language models, providing huge advantages to those with fluency in English and driving assimilation towards English to access the benefits of the technology. This, in turn, reduces demand for NLP tools in non-English languages, driving a vicious circle.

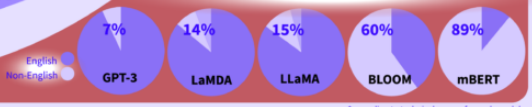
4. Limitations of multilingual LLMs

The drive towards big data and AI has focused much NLP research on building large language models (LLMs), which can perform a range of tasks. Many LLMs tout multilingual capabilities; but:

- training sets and tokenisation strategies usually favour English above other languages;
- model capabilities in most non-English languages have not been properly tested, due to a lack of evaluation data
- strategies to reduce toxicity and bias, such as reinforcement learning from human feedback, are rarely applied to a model's non-English outputs

These factors mean that non-English content produced by LLMs usually falls short of its English-level performance and safety standards.

Proportion of models' training data which is in non-English languages*



*according to technical papers for each model

2. WHAT METHODS CAN WE USE TO IMPROVE LOW RESOURCE NLP?

Data-collection

The most reliable way to improve NLP tools for low-resource languages is to create more resources for model training. This includes unstructured data like recordings, web content or news reports as well as labelled tagsets and datasets for specific NLP tasks, such as Part-of-Speech tagging or Named Entity Recognition, or evaluation benchmarks. Collecting and labelling language data requires working with local language experts, and while costly it is essential for building NLP capacity.

EXAMPLES

Adelani et. al (2021) created **MasakhaNER,** the first large publicly available high-quality dataset for named entity recognition (NER) in ten African languages.

Kann et al (2022) created the **Americas NLI corpus,** which extended Conneau et. al's (2018) XLNI corpus for language inference to 10 Indigenous American languages.

Agic & Vulic (2019) used the Jehovah's Witness website, which contains documents translated into hundreds of low resource languages, to create the **JW300 corpus**.

Data augmentation

We can also leverage existing machine learning techniques for low-resource contexts by augmenting the small amount of training data we may have. This includes techniques like oversampling, bootstrapping, or building systems for synthetically generating new data either through a rule-based system (e.g. combine parts of this phrase and parts of that phrase to make a new sentence) or through a generative model trained on existing text or structural data.

EXAMPLES

Sennrich, Haddow and Birch (2016) improved low-resource **Turkish-English machine translation** (by +2.1-3.4 BLEU) by synthetically generating translation data through "backtranslation" of monolingual data.

Qin et al (2020) developed an algorithm for generating multi-lingual codeswitching data via replacement, leading to substantial performance gains across 5 tasks in **19 different languages** when fine-tuning mBERT.

Transfer learning

Transfer learning allows us to adapt a pre-trained model from a different context, leveraging the knowledge gained in a higher-resource setting and applying it to a lower-resource one. The model is pre-trained on a large dataset, some of its weights are fixed, and then it is fine-tuned on more specific data or prompts. While this can be effective, success may depend on the typological similarities of the languages in question.

EXAMPLES

Nag et. al (2023) adapted mBERT for entity-detection and relation classification in **Hindi, Bengali and Telugu** using a dataset of 21k tagged sentences in each language, obtaining F1 scores of 86.43-94.32.

Adelani et al (2022) adapted multilingual pre-trained models (MT5, ByT5, mBART, M2M-100) to **16 low-resource African languages** using 2k-7k sentences, improving machine translation for those languages by up to 4.1 BLEU compared to model baselines.

Zero-shot learning

Zero-shot learning is related to transfer learning, but here the model must generalise to the novel ask or language without *any* training data at all. This approach is highly scalable and does not require expensive data. However, like transfer learning, its effectiveness appears to depend on the similarity of the languages seen in the training data with the unseen target language. Furthermore, performance of a zero-shot model is much lower than a fully-trained, bespoke model, limiting their utility.

EXAMPLES

Lent et al (2023) applied pre-trained multilingual models to unseen **Creole** languages, showing better zero-shot performance for Named-Entity-Recognition and Part of Speech tagging compared to sentence analysis or inference.

Üstün et al (2024) trained **Aya** on 101 languages, of which 50% are lower-resourced. They show that Aya's accuracy on four zero-shot unseen discriminative tasks in 31 languages (including 7 low- and 7 medium-resourced languages) exceeds other model baselines by an average of 19.8%.

3. HOW CAN WE INCENTIVISE MORE INVESTMENT IN AND ATTENTION FOR LOW RESOURCE NLP?

The safety argument: Low-resource NLP is vital for detecting hate speech and disinformation online in non-English languages, which is increasingly a regulatory requirement for large online platforms. LLMs should also be tested in lower resource languages to ensure that they are not vulnerable to manipulation or toxicity in those languages.

The market incentive: The top 100 most spoken languages in the world each have over 10 million speakers and 75% of the world's population does not speak English at all. There is a huge market for technology in non-English languages, particularly for translation software to enable cross-cultural communication and knowledge-sharing.

The climate agenda: Finding ways of training language models on less data not only allows us to better model low-resource contexts, but also paves the way for more computationally efficient methods, reducing carbon footprints of LLMs and driving progress towards AI sustainability.

The research imperative: The field of linguistics shows us that comprehensive models of human language systems require input data from very many languages. Limiting NLP research to a handful of high-resource languages prevents us from understanding universals of human communication and likely hinders progress towards general-purpose AI.

References

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., ... & Osei, S. (2021). MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9, 1116-1131.

Adelani, D. I., Alabi, J. O., Fan, A., Kreutzer, J., Shen, X., Reid, M., ... & Manthalu, S. (2022). A few thousand translations go a long way! leveraging pre-trained models for african news translation. *arXiv preprint arXiv*:2205.02022.

Agic, Z., & Vulic, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. *Association for Computational Linguistics*, 57, 3204–3210.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv*:1809.05053;

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv*:2004.09095;

Kann, K., Ebrahimi, A., Mager, M., Oncevay, A., Ortega, J. E., Rios, A., ... & Vu, N. T. (2022). AmericasNLI: Machine translation and natural language inference systems for Indigenous languages of the Americas. *Frontiers in Artificial Intelligence*, 5, 995667;

Lent, H., Tatariya, K., Dabre, R., Chen, Y., Fekete, M., Ploeger, E., ... & Bjerva, J. (2023). CreoleVal: Multilingual Multitask Benchmarks for Creoles. *arXiv preprint arXiv*:2310.19567;

Nag, A., Samanta, B., Mukherjee, A., Ganguly, N., & Chakrabarti, S. (2023). Transfer learning for low-resource multilingual relation classification. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2), 1-24;

Qin, L., Ni, M., Zhang, Y., & Che, W. (2020). Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv preprint arXiv*:2006.06402;

Rivière, F. (Ed.). (2009). Investing in cultural diversity and intercultural dialogue (Vol. 2). *UNESCO* p.305-307;

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv*:1511.06709.

Üstün, A., Aryabumi, V., Yong, Z. X., Ko, W. Y., D'souza, D., Onilude, G., ... & Hooker, S. (2024). Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv*:2402.07827.