

ELLORA: Enabling Low Resource Languages with Technology

Kalika Bali, Monojit Choudhury, Sunayana Sitaram, Vivek Seshadri

Microsoft Research Labs

Bangalore, India

{kalikab, monojitc, susitara, visesha}@microsoft.com

Abstract

Language technology has had a huge impact on the way language communities communicate and access information. However, this revolution has bypassed over 5000 languages around the world that have no resources to develop technology in their languages. ELLORA, with its mission to empower every person and every organization using underserved languages to achieve more, is a program for enabling low resource languages through language technology. In this paper we describe how through innovative methodologies and techniques that allow systems to be built in resource constrained settings, ELLORA seeks to positively impact the underserved language communities around the globe..

Keywords: Low resource languages, language technology, language data, endangered and minority languages

Résumé

भाषा प्रौद्योगिकी के माध्यम से डिजिटल क्रांति की अगुवाई में भाषा समुदायों को जानकारी पहुँचाने के तरीके पर भारी प्रभाव पड़ा है। जबकि कुछ प्रमुख संसाधन संपन्न भाषाओं के उपयोगकर्ता हर दिन ऐसी तकनीक का लाभ उठा रहे हैं, इस क्रांति से दुनिया भर की अधिकांश भाषाएँ उपेक्षित हैं। प्रौद्योगिकी विकास की डेटा-भूखी दुनिया में, 5000 से अधिक भाषाओं को अपनी भाषाओं में प्रौद्योगिकी विकसित करने के लिए कोई संसाधन नहीं हैं। एलोरा (ELLORA) का मिशन प्रत्येक व्यक्ति और प्रत्येक संगठन को अपनी भाषा में आगे बढ़ने के लिए प्रोत्साहित करना है। यह प्राकृतिक भाषा प्रसंस्करण, वार्तालाप और भाषण प्रौद्योगिकी के माध्यम से कम संसाधन भाषाओं को सक्षम करने के लिए एक कार्यक्रम है। इस शोध पत्र में हम बताते हैं कि कैसे नवीन पद्धतियों और तकनीकों के माध्यम से ELLORA दुनिया भर के भाषा समुदायों को सकारात्मक रूप से प्रभावित करना चाहता है।

1. Introduction

Technology pervades all aspects of society and continues to change the way people access and share information, learn and educate, as well as provide and access services. Language is the main channel through which such transformational technology can be integrated into the socio-economic processes of a community.

However, this benefit is still limited to a subset of the world's language communities and large populations worldwide are bereft of access to technology in their own languages. Most languages in the world lack the linguistic resources to build large data-driven (e.g., Deep Neural Net) models. To be able to truly support speech and language systems that can enable everyone on the planet, methodologies and techniques to build systems in resource constrained settings are essential.

ELLORA-a program for enabling low resource languages through Natural Language Processing, Conversation and Speech technology was established at Microsoft with the view to address the needs and aspirations of language-users currently unable to access such technologies. ELLORA aims to impact underserved communities through enabling language technology by creating economic opportunities, building technological skills, enhancing education and preserving local language and cultures for future generations. The approach taken is to start with the low resourced languages of India and then scale it to the low resource languages in the rest of the world.

In recent times, there have been a number of breakthroughs

in the field of NLP, which is primarily due to the use of deep neural networks and availability of large amount of language resources as well as computational power. Languages resources include both linguistic datasets that are used for training language processing systems, and basic language processing tools such as stemmers, morphological analyzers, parsers etc., which enable other language processing technologies.

In September 2017, Microsoft announced a speech recognition system that could achieve better-than-human performance in speech transcription (Xiong et al., 2017), which used 200M transcribed words from conversational speech. In 2018, human-parity was achieved for English-Chinese Machine translation, again training on 18M bilingual sentence pairs (Hassan et al., 2018). These achievements hold a lot of promises, particularly in making information and technology accessible to the speakers of a language. Unfortunately, these technologies work only when there is large amount of training data. For most languages in the world, there is hardly any resources available.

In a 2008 study of the Linguistic Data Consortium (LDC) portal, Choudhury (2008) shows that resource distribution across languages follow power-law, with only four languages – Arabic, Chinese, English and Spanish – having a large amount of resources and only a handful having some, leaving 90% of the world languages in the long tail of this distribution with hardly any resources (Fig 1) to train any useful NLP system. Choudhury (2008) predicts that since availability of language resources propels creation of technology and more resources, and attracts more

researchers and technologists toward that language, this digital divide between languages will widen further with time, unless intervened aggressively and strategically. Indeed, today, after a decade, while technology has grown more and more data intensive, LDC catalogue does not list any resource for Javanese and Kannada, which are the 12th and 32nd most spoken languages with 98M and 43.7M speakers respectively.

A study by Caribou Digital Analysis (Will et al., 2019) summarizes the economic aspect of this divide as follows: “Analysis shows there is a clear income gap in access. Google Translate is available in the languages spoken by just 54% of those living on less than \$1.90 per day. The picture is even more stark for Natural Language Understanding frameworks such as Dialogflow, which supports languages spoken by only 3% of those living on less than \$1.90 per day.”

Based on the availability of digital resources and access to technology, the languages of the world can be broadly classified into four groups, as illustrated by the pyramid in Fig 1. Only the top of the pyramid, which has 10-15 economically important languages, are positively affected by the breakthroughs in AI and NLP. Research and development of technology for low resource languages have always been in the fringes. The recent advances in language technology are mostly beneficial to languages in the second and third tier, which together cover another 100 languages. The remaining 5000+ languages, i.e., the bottom of the pyramid, have no resources, and consequently the speakers do not have access to technology in their native languages.

2. Enabling Data and Technologies

Language technology enables access to information as well as broader technology, such as through local language smart phone interfaces. This can be viewed as a multi-layer approach to enabling a language, where the more complex AI technologies build upon fundamental and simpler layers. Input/Output Support form the first layer of language technology, that include basic font and keyboard support, and more advanced features like text prediction, and spelling and grammar correction. Speech input/output mechanisms include speech-to-text and text-to-speech systems and are more relevant for languages without script and where a large fraction of low-literate users. Local language UI is the next layer of support, which ranges from availability of particular OS or apps in a language to generic technological support for building UI in a language. Information access is enabled by text and voice based search technologies. However, languages that have little content on the Web will particularly benefit from machine translation techniques that can be used to translate content from other languages. Finally, digital assistants and conversational interfaces are useful for ease of interfacing with the devices, various other technologies and are enablers of businesses in that language.

Building some of these technologies require linguistic expertise, while others are data intensive. Some technologies, such as speech-to-text, require transcribed speech data, which is labor intensive and requires native speakers to label/transcribe data, whereas other technologies such as text prediction can be built by training

on any corpus of the language and requires no further human labeling or processing. This makes some of the technologies more expensive, time intensive and challenging to build for low resource languages. Table 1 summarizes the data/expertise requirement for various technologies and their availability.

As can be seen from the table, technologies that require high to moderate amounts of labeled data are typically unavailable for No and Low resourced languages. Languages in the bottom of the pyramid also lack technologies which require only unlabeled data. This is not surprising as creation of labeled data is expensive and there may not be any financial incentive to invest in these languages. While Low-resourced languages have the basic enabling technologies, they are deprived of high-quality advanced tools, especially those that require labeled data

3. ELLORA Data Initiatives

As discussed in the above sections, building language technologies requires significant data resources in the desired language. As a commercial initiative, there has been little interest in developing these resources for indigenous and minority languages, as well as the languages of the Global South. Data thus, remains the single biggest bottleneck as language technology models become more and more data hungry. Data remains at the heart of several activities undertaken through Ellora, ranging from innovative data collection initiatives to providing seed data for nurturing research.

Microsoft India released speech training and test data for Telugu, Tamil and Gujarati in Sept 2018. This is the largest publicly available Indian language speech dataset aimed at helping researchers and academia build Indian language speech recognition for all applications where speech is used. This data was used as the basis of the Low Resource Speech Recognition Challenge, held as a part of Interspeech 2018. Participants were able to create high quality speech recognition models using the Microsoft Indian language speech corpus, thus validating the efficacy of the corpus (Srivastava et al, 2018)

As data collection remains an expensive exercise, an effort to collect high quality data at low cost is one of the goals for ELLORA. Karya (Chopra et al, 2019) a crowdsourcing platform to provide digital work to low-income workers, presents an ideal vehicle for this. The motivation for Karya lies in the fact that roughly half the Indian population lives under \$5 a day. By identifying training needs and imparting digital literacy to low-income population, Karya enables microtasking in an efficient manner on a mobile platform. Thus, it significantly increases the current daily wage of the workers while at the same time reduces the cost of task completion which may allow data collection at scale. Currently, an experimental project in collaboration with Indian Institute of Technology, Bombay, to determine the feasibility of using Karya successfully for speech data collection is underway in rural and semi-urban parts of the Indian state of Maharashtra. While, the final results are awaited, the initial feedback seems encouraging.

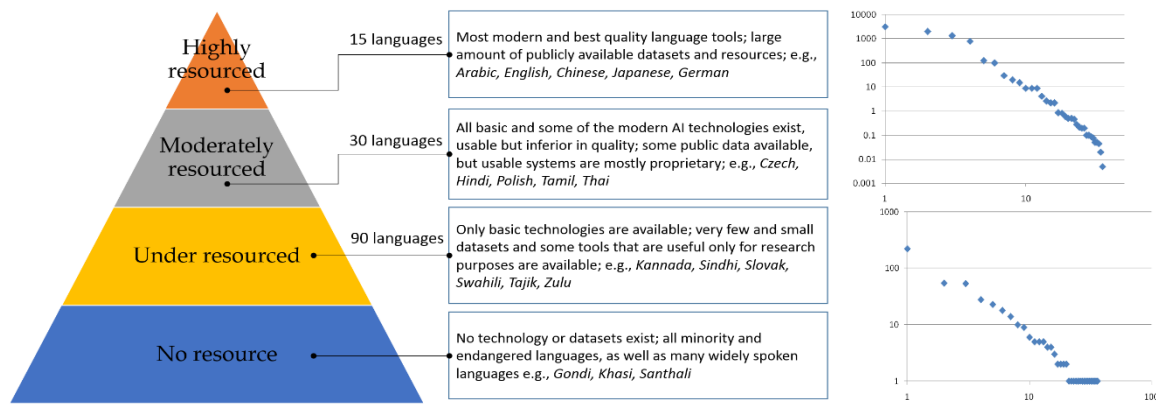


Figure 1: Classification of languages according to the availability of language technology, tools and resources (left) based on the power-law distribution of the resources across the languages of the world (right).

4. Enabling Minority Languages

Some of the most disadvantaged socio-economic groups in the world are also the ones with the least access to language technology. ELLORA’s impact can only be measured if the communities using languages with no technological support can be provided access through technology in their own language. To understand the viability of such a social impact, ELLORA is working with CGNet Swara on building and deploying Gondi language technology.

Gondi is a South-Central Dravidian language and is in the ‘vulnerable’ category on UNESCO’s Atlas of the World’s Languages in Danger (Mosley, 2010). Spoken by nearly 3 million people (Indian Census, 2011) in the Indian states of Chhattisgarh, Andhra, Odisha, Maharashtra and Karnataka, it is heavily influenced by the dominant state language. CGNet Swara provides a citizen journalism portal for the tribal regions of Chhattisgarh and home to the Gondi language community, by making local stories accessible through mobile phones. As there is absolutely no language technology support for Gondi, most of the content is created, moderated and edited manually. Targeted language technology applications can increase the scale, and hence the access to information for a community that is marginalized and lives in areas of civil unrest.

A meeting was held in April 2018 to understand the potential impact of Language Technology for Gondi on the community and brainstorm on transformational technology applications. The discussions involved stakeholders and experts from CGNet Swara, academic institutes like IIIT Raipur, IIT-KGP, Jadavpur University etc, Microsoft and other non-profit organizations like Pratham Books. Subsequently, a workshop was organized at IIIT Naya Raipur in collaboration with Pratham Books, CGNetSwara and Microsoft Research. The Gondi speakers who participated in the workshop translated approximately 200 books on Storyweaver from Hindi to Gondi. Not only was this the first step towards creating parallel data for Gondi-Hindi that can be of use in building Machine Translation systems for Gondi, it made available children’s books for the first time in the language.

Adivasi Radio, a Mobile News App for Gondi has also been developed through this collaboration. The first version released uses Text-to-Speech synthesis in Gondi to read out news and articles available on the CGNetSwara site on the users’ phones. Future development envisages the incorporation of a Machine Translation system that allows news articles and other content in Hindi to be translated and read out in Gondi. This would have a major impact on the community by providing access to news in local language, while also producing content in Gondi.

Technology	Availability of technology for the resource status of a language				Data/Expertise Requirement		
	High	Mode-rate	Low	No	Linguistic Expertise	Unlabel-ed Data	Label-ed Data
Input/Output Support							
Font & keyboard	☆☆☆	☆☆☆	☆☆☆	☆☆	☆☆☆		
Speech-to-text	☆☆☆	☆☆			☆	☆☆	☆☆☆
Text-to-speech	☆☆☆	☆☆	☆		☆☆☆		☆☆
Text prediction	☆☆☆	☆☆☆	☆☆			☆☆☆	
Spell checker	☆☆☆	☆☆☆	☆☆		☆☆☆	☆☆	
Grammar checker	☆☆☆	☆☆			☆☆	☆☆☆	☆☆
Local language UI							
	☆☆☆	☆☆☆	☆		☆☆☆		
Information access							
Text search	☆☆☆	☆☆	☆		☆	☆☆☆	☆☆
Voice to Text search	☆☆☆	☆				☆	☆☆☆
Voice to speech search	☆☆	☆			☆	☆☆☆	☆☆☆
Machine translation	☆☆	☆	☆		☆☆		☆☆☆
Conversational systems							
	☆☆	☆			☆☆☆	☆☆☆	☆☆☆

Table : Enabling language technologies, their availability and quality

5. Discover, Design, Develop and Deploy

While it is clear that a large number of languages in the world require intensive investment in resource creation for technology enablement, it seems highly unlikely that such an investment can be delivered readily and easily in a short span of time. It is thus imperative that the investment is done in a manner that ensures maximum benefit for a community through language technology. To enable language technology to deliver a positive social impact on the low-resource language communities around the world, we propose the use of a modified version of the 4-D design thinking methodology of Discover, Design, Develop and Deploy.

Discover the problem that language technology can address for a particular language community. This observation led approach can help target resources where they are most needed.

Design for the users and their language. Understand the diversity in linguistic properties of languages and their usage. Avoid a majority language (usually English) led approach where all effort is spent on adapting an existing technology for a dominant language. The cost of ignoring pertinent language features in such a process often results in a less than optimal technology development.

Develop rapidly and Deploy frequently. An iterative process can ensure early failures lead to success.

Such a user-centric approach will not only deliver a more functional language technology but also ensure a more equitable and beneficial distribution of resources.

This approach, at the heart of all ELLORA activities, is well-illustrated in a project on Mundari language. This collaboration between ELLORA and IIT-Kharagpur has a team of scholars led by Prof Dripta Piplai and Prof Manjira Sinha have been spending time with the Mundari-speakers in the Jhargram villages in the eastern state of West Bengal observing and recording the nuances of the language and its use (Mitra, 2019) This information will feed into a mobile based app for teaching, learning and communication in Mundari. According to Dr Piplai, “The close interactions have thrown up interesting facts. One of them is that the younger generation does not use it in its pure form. They communicate in a mix of Bengali, Mundari and Oriya”

The digital revolution spear-headed by language technology has had a huge impact on the way language communities communicate and access information. However, this revolution seems to have bypassed the 5000+ languages at the bottom of the pyramid with zero resources available to them. ELLORA aims to bridge this gap by supporting language technology systems and applications to enable everyone on the planet. Through innovative methodologies and techniques that allow systems to be built in resource constrained settings, ELLORA seeks to positively impact the underserved language communities around the globe.¹

6. Acknowledgements

We would like to acknowledge Manu Chopra, Shubhranshu Choudhary, Sandipan Dandapat, Rupesh Mehta, Niranjana Nayak, Dripta Piplai, Manjira Sinha, Brij Mohan Lal Srivastava, IIT Bombay, IIIT Naya Raipur, Pratham Books, Gondi workshop participants and the people of Amale village for their contribution to ELLORA.

7. Bibliographical References

- Awadalla, H. et al. (2018). Achieving human parity on automatic Chinese to English News Translation. arXiv:1803.05567
- Census (2011). Primary Census Abstracts, Registrar General of India, Ministry of Home Affairs, Govt of India
- Chopra, M. et al (2019). Exploring crowdsourced work in low-resource setting. ACM CHI Conference on Human Factors in Computing Systems.
- Choudhury, M. (2008). Breaking the Zipfian barrier of NLP. Invited Talk. . In the Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages.
- Gericke, K and Blessing, L (2011). Comparisons of design methodologies and process models across domains: a literature review. In DS 68-1: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Vol. 1: Design Processes, Lyngby/Copenhagen, Denmark.
- Mager, M. et al (2018). Challenges of language technologies for the indigenous languages of the Americas. Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, August 2018.
- Mitra, D. (2019). IIT set to launch app in Mundari to keep indigenous language relevant. Times of India, February 2019.
- Mosley, C. (ed.) (2010). Atlas of the World’s Languages in Danger, 3rd edn. Paris, UNESCO Publishing.
- Srivastava, B.M.L. et al (2018). Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages 29-31 August 2018, Gurugram, India
- Willis, A, Barrie, G, and Kendall, J (2019). Conversational interfaces and the long tail of languages in developing countries. <https://dfslab.net/wp-content/uploads/2019/01/NLP-Language-Divide-Report-.pdf>
- Xiong, W. et al. (2017). Toward Human Parity in Conversational Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing | September 2017, Vol 25: pp. 2410-2423