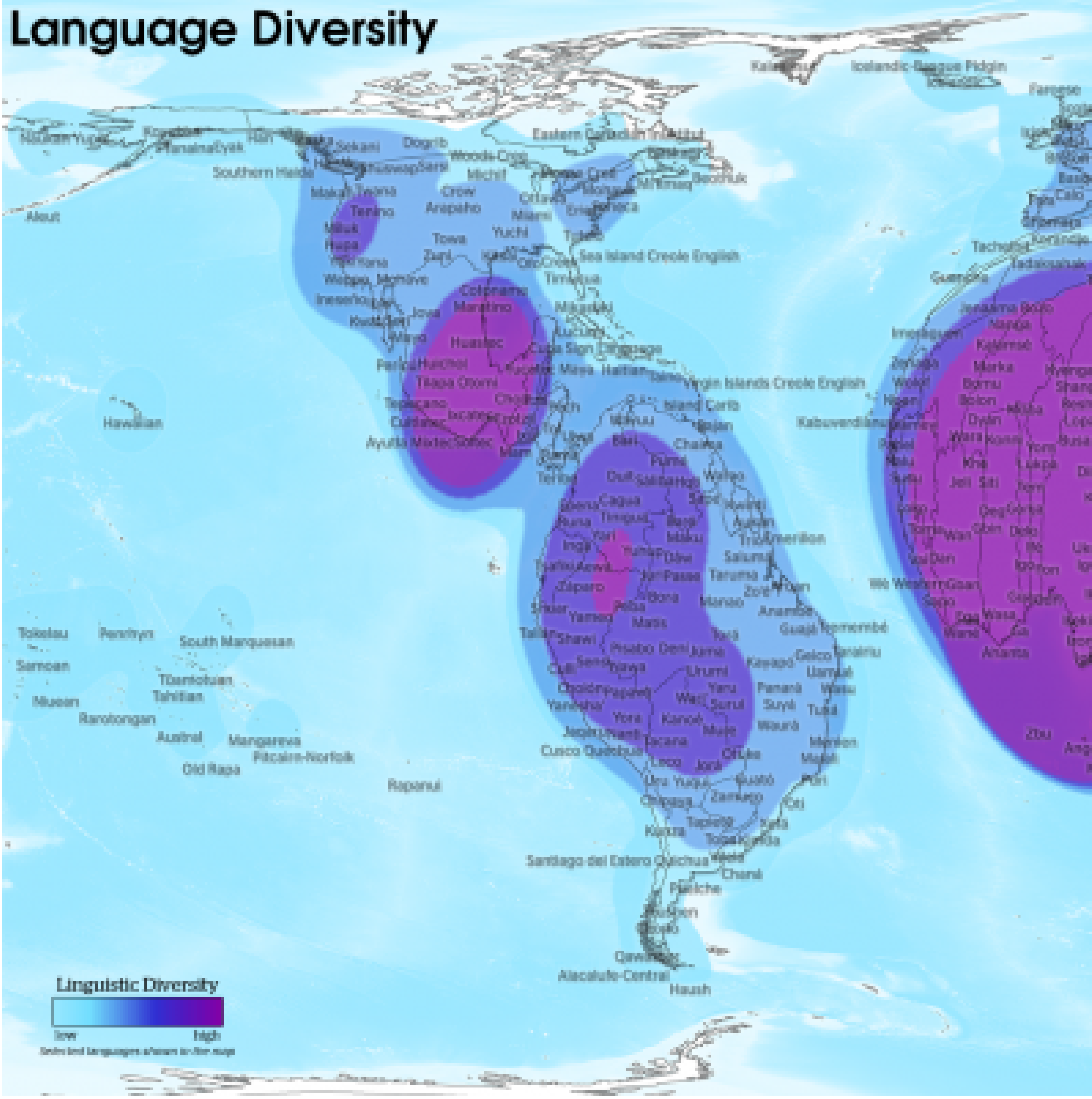
# Multilingualism in Natural Language Processing targeting low resource Indian languages

UNCATEGORIZED

## Introduction

Language is a systematic form of communication that can take a variety of forms. There are approximately 7,000 languages believed to be spoken across the globe. Despite this diversity, the majority of the world's population speaks only a fraction of these languages. In Spite of such a rich diversity Languages are still evolving across time much like the society we live in. While the English language is uniform, having the distinct status of being the official language of multiple countries there are varieties of English too like American English, British English and so on. For example Colloquial Singaporean English, better known as Singlish, is an English-based creole language spoken in Singapore. The term Singlish is a blend of Singaporean slang, English and other local delicacies which has unique grammar different from that of the English and mandarin Chinese. And when we build an NLP system we should also be able to provide it to the speakers of Singlish language which constitutes millions of people. Next example is African languages in Africa which is a continent with very high and rich linguistic diversity there are around 1.5 – 2 K African languages which are spoken by 1.33 billion people and almost all of the African languages are resource poor. And NLP technologies are the key to connect billions of people around the globe to the access and avail the facilities, provisions, learning resources, knowledge and benefits of the internet from education to communication and many other applications that can be provided to the future generations in their language of conveniences.



This map shows linguistic diversity index around the world. Linguistic Diversity Index (LDI) or Greenberg's diversity

India is the epicenter of diversity which is why there are 456 languages spoken in India. Constitutionally, 22 are official languages India is perhaps the richest language hub on the earth. India can boast of being the home for a number of dominant languages of the world, and a large number of small languages that are in the brink of extinction. Hindi is the most widely spoken language (40% of the population). English, although spoken by a smaller percentage of the population (10%) is important in business (easier to get better jobs if one speaks English) and education (most of the universities only teach in English).

## Table of content

    1. Transfer learning
        - Zero Shot Learning
    2. Joint multilingual learning
        - Multilingual Embeddings
    3. Crosslingual transfer learning
    4. Unsupervised Techniques
        - Unsupervised word translation
        - Unsupervised Multilingual Machine Translation

3. Google translate for Indian languages

## Challenges in using NLP for low resource languages

This article is more of a survey looking at recent trends and technologies that benefit multilinguality. There are 7000 major languages in the world 120 major languages in India and this gives rise to the divide about the availability of resources of training data and benchmarks which are not available for the majority of world's languages and therefore the benefits of the natural language technology which has been taken for granted in developing various systems and applications in English and other resource rich languages have not reached many of the other users yet.

The objective of our interest is to see how one can benefit or build systems which can work in all languages especially in low resource languages.

*If you look at the study of classical NLP language has different layers starting from sound, word, Morphology, POS, Syntax, Semantics etc.*

As in many other domains there has been several paradigm shifts in NLP in the initial days there were logic base and rule based language modelling system, and the major emphasis shifted in the 90's to statistical NLP and in the mid 2010's along with all other domains Neural Networks started to form a very important component in many NLP systems.

While this has been the case it has also meant in all of these scenarios to the most part these machine learning models have not really benefited so much of low resources languages but this is changing in recent years and that is what this article is focused on.

**The reason why standard NLP techniques cannot be applied to low resources languages is because the NLP techniques either require linguistic knowledge that can be only developed by experts and some by speakers of that language.**

**Or they require a lot of labeled data which is again expensive to generate.**

An naïve attempt would be to collect manually data for each individual resource poor language. But obviously we all understand that this is prohibitively expensive and infeasible.

## Work around Low resource languages so far

So there are 5 major approaches in enable the use of NLP technology in languages with low resources

1. Manual curation and annotation of large scale resources for thousands of languages is infeasible and prohibitively expensive
2. Unsupervised Learning
3. Cross-lingual transfer learning

4. Zero and One shot learning
5. Joint multilingual learning

## Transfer learning (Zero shot and one shot translation)

In transfer learning we transfer either models or sometimes resources in which we can take labeled or unlabeled data which have been created in one language and transfer it to another language so as to quickly develop labeled data in the target language on which you can build your models or you can learn a model in a source language and apply it directly to the target language so both transfer of annotations and transfer of models can benefit low resource languages. And in some cases zero and one shot learning has become a reality.

Zero-shot-learning

It includes an encoder, decoder and attention module, remains unchanged and is shared across all languages. In a multilingual NMT model, all parameters are implicitly

shared by all the language pairs being modeled. This forces the model to generalize across language boundaries during training. It is observed that when language pairs with little available data and language pairs with abundant data are mixed into a single model, translation quality on the low resource language pair is significantly improved.

A surprising benefit of modeling several language pairs in a single model is that the model can learn to translate between language pairs it has never seen in this combination during training (zero-shot translation) a working example of transfer learning within neural translation models. For example, a multilingual NMT model trained with Hindi→English and English→Marathi examples can generate reasonable translations for  Hindi→Marathi although it has not seen any data for that language pair.

Training the model in one domain and assuming it generalizes more or less out-of-the-box in a low resource domain without annotated data of target language. One shot learning. Training the model in one domain and using only a few samples from low resource domain to adopt it.

## Joint multilingual learning

In this approach we train a single model on a mix dataset in all languages to enable data and parameters sharing wherever possible. It can be implemented in different ways some of them are listed below

### Multilingual Embeddings

If we want to use embedding in a way to benefit low resource languages one way of doing this is to come up with a shared representation because if you can have shared representation where words in 2 or more languages they can share the same vector space such as 2 words will be close to each other in the vector space if their meaning is similar irrespective of from what language they belong to!

This will also benefit Transfer learning, cross lingual information retrieval, Machine Translation

There has been gradual progress in it starting from simple Embeddings by Mikolov et al, 2013 who worked on getting some linear transformation from one space to another to more advanced work in recent years some of which we'll discuss in this article. They believed that there exist some linear relationships between the words like

King – Queen = Man – Women

And it does not require a complex nonlinear model although it is not True always but when talking about Multilingual embeddings the case is not the same because when you look upon vocabulary of 2 languages there is a core vocabulary which is common here is some peripheral vocabulary that may or may not be common because the words can be very culture specific or region specific.

There have been methods that depend on certain resources like a small bilingual dictionary or sentence aligned parallel data or document aligned corpora.

## Crosslingual transfer learning

However there has been a lot of research in the last several years which involve various machine learning frameworks which enable the quick development of systems in low resources languages for example there has been a lot of interest in transfer learning. In transfer learning you can transfer a task which has been developed for some domain to another domain and you can also do a transfer of knowledge from a resource rich language to resource poor language or from one NLP task to another NLP task so both cross domain and cross lingual transfer and has in general benefited NLP Systems so that features and systems developed for one task could also benefit other tasks. And which also set to benefit cross lingual transfer of knowledge between languages. In cross-lingual transfer learning we train a model for resource poor language but using resources from resource languages we can transfer annotations we can build a word or phrase alignments and this would serve as bridges between resource rich and resource poor language and we can also project annotations from resource rich language to resource poor language.

## Unsupervised word translation

In the last 2 years there has been a lot of good work in unsupervised word translation that does not depend on either a dictionary or parallel corpus which are only based on monolingual corpora.

So one example of this work is done by this group Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou

Paper Link : https://arxiv.org/abs/1710.04087

So what they do is that they take monolingual corpora in 2 or n language and they independently learn the world embeddings for lets say English and separately the embeddings for let’s say Italian. And then they try to learn a rotation matrix to roughly align the 2 domains firstly they focus on frequent words at random from each language and then they project one of the 2 and then try to rotate the embeddings to make sure the distributions match they also use a method so as to maintain the shapes of the distribution and maintain some orthogonality criteria of the mapping using the most frequent words and then they iterate this after which the are able to get common embeddings. They claim that results on word translation with this technique even outperforms supervised approaches which rely on small dictionary and parallel corpora by using more anchor points and lots of unlabeled data.

### Unsupervised Multilingual Machine Translation

As we know good quality machine translation using statistical methods or neural networks a large number of parallel sentences to get visible results and such resources are not available for most of the language pairs.

Machine translation has recently achieved impressive performance thanks to recent advances in deep learning and the availability of large-scale parallel corpora. There have been numerous attempts to extend these successes to low-resource language pairs, yet requiring tens of thousands of parallel sentences. In this work, they take this research direction to the extreme and investigate whether it is possible to learn to translate even without any parallel data. They propose a model that takes sentences from monolingual corpora in two different languages and maps them into the same latent space. By learning to reconstruct in both languages from this shared feature space, the model effectively learns to translate without using any labeled data. They demonstrate their model on two widely used datasets and two language pairs, reporting BLEU scores of 32.8 and 15.1 on the Multi30k and WMT English-French datasets, without using even a single parallel sentence at training time.

## Google translate for Indian languages

Google’s free service instantly translates words, phrases, and web pages between English and over 100 other languages. Google Translate is using the Neural Machine Translation technology to translate between English and nine widely used Indian languages Hindi, Bengali, Marathi, Tamil, Telugu, Gujarati, Punjabi, Malayalam and Kannada. Neural translation offers a huge improvement over the old phrase-based system, translating full sentences at a time, instead of pieces of a sentence. Google also brought dictionary functionality on Google Search for users, and will offer Hindi dictionary results.

Increasing fluency and accuracy in translations instead of translating word by word, NMT allows translating sentences at a time, rather than just piece by piece. It uses this broader context to help it figure out the most relevant translation, which it then rearranges and adjusts to be more like a human speaking with proper grammar.

## NLP Tools for Indian Languages

As Indian languages pose many challenges for NLP like ambiguity, complexity, language grammar, translation problems, and obtaining the correct data for the NLP algorithms, it creates a lot of opportunities for NLP projects in India.

### NLTK and iNLTK

iNLTK provides support for various NLP applications in Indic languages. The languages supported are Hindi (hi), Punjabi (pa), Sanskrit (sa), Gujarati (gu), Kannada (kn), Malayalam (ml), Nepali (ne), Odia (or), Marathi (mr), Bengali (bn), Tamil (ta), Urdu (ur), English (en). iNLTK is like the NLTK Python package. It provides the feature for NLP tasks such as tokenization and vector embedding for input text with an easy API interface.

### Stanford NLP

StanfordNLP contains tools which can be used to convert a string containing human language text into lists of words and sentences. This library converts the human language texts into lists to generate base forms of those words, parts of speech and morphological features, and also to give a syntactic structure dependency parse. This Syntactic structure dependency parse is designed to be parallel among more than 70 languages using the Universal Dependencies formalism.

## Resources available for Indian languages

### Semantic Relations from Wikipedia

Contains automatically extracted semantic relations from multilingual Wikipedia corpus.

https://console.developers.google.com/storage/browser/wikipedia_multilingual_relations_v1/

### HC Corpora (Old Newspapers)

This dataset is a subset of HC Corpora newspapers containing around 16,806,041 sentences and paragraphs in 67 languages including Hindi.

https://www.kaggle.com/alvations/old-newspapers

### Sentiment Lexicons for 81 Languages

This dataset contains positive and negative sentiment lexicons for 81 languages which also includes Hindi.

https://www.kaggle.com/rtatman/sentiment-lexicons-for-81-languages

### IIT Bombay English-Hindi Parallel Corpus

This dataset contains parallel corpus for English-Hindi and monolingual Hindi corpus. This dataset was developed at the Center for Indian Language Technology.

http://www.cfilt.iitb.ac.in/iitb_parallel/

### Indic Languages Multilingual Parallel Corpus

This parallel corpus covers 7 Indic languages (in addition to English) like Bengali, Hindi, Malayalam, Tamil, Telugu, Sinhalese, Urdu.

http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/index.html

### Microsoft Speech Corpus (Indian languages)(Audio dataset)

This corpus contains conversational, phrasal training and test data for Telugu, Gujarati and Tamil.

https://www.microsoft.com/en-us/research/event/interspeech-2018-special-session-low-resource-speech-recognition-challenge-indian-languages/

### Hindi Speech Recognition Corpus(Audio Dataset)

This is a corpus collected in India consisting of voices of 200 different speakers from different regions of the country. It also contains 100 pairs of daily spontaneous conversational speech data.

https://kingline.speechocean.com/exchange.php?act=view&id=16389

## End Notes

As we have seen that unsupervised and semi-supervised language processing along with transfer learning is progressing a lot and this holds a great promise for developing systems in NLP involving low resources languages. We have also seen that the system can be greatly improved if you have limited supervised data which is a great promise in working on various tasks. Then the multilingual systems that simultaneously work in multiple languages taking these underlying principles go well beyond Machine Translation.

---

Article Url - https://www.analyticsvidhya.com/blog/2020/12/multilingualism-in-natural-language-processing-targeting-low-resource-indian-languages/

**Hrithik**