Application of Natural Language Processing to Preserve Minority, Regional and Oppressed Languages

Can Natural Language Processing be used to preserve minority, regional and oppressed languages given the lack of data?

ITGS Extended Essay

Word Count: 3881

Introduction

My grandparents live in a small town in India where Awadhi is the local language — a language with a rich culture and history. Whenever I visit them, I enjoy listening to people speak Awadhi. I like learning languages, so I wanted to be able to converse with the people in the town in their native Awadhi. However, when I attempted to learn it, I realized that there were very few resources available. More shocking to me was that Awadhi-speaking people seemed to have given up on passing their language and heritage to their future generations. Rather, they encourage their children to learn English as they consider it to be the key to success in their lives and careers. I also learned that only Hindi is taught in primary schools and English is added in middle school. The business of the state and central government is conducted in Hindi and English. As I began exploring this phenomenon further, I realized that there is more to it than meets the eye. More importantly, I grew interested in exploring how technology could be used as a tool to promote and preserve regional and minority languages like Awadhi.

The question, "**Can Natural Language Processing be used to preserve minority, regional and oppressed languages given the lack of data?**", is worthy of investigation. This is because, as Franco-American polyglot George Steiner once said, "when a language dies, a way of understanding the world dies with it, a way of looking at the world." Additionally, the Sapir-Whorf hypothesis suggests that the language a person speaks natively, influences their perception of reality and the structure of their thoughts. It is well understood and acknowledged that it is important to be able to incorporate diverse perspectives and opinions for making the best decision, whether commercial, political, or social. The significance of languages is best summarized by the United Nations: "Through language, people preserve their community's history, customs, and traditions, memory, unique modes of thinking, meaning and expression.

They also use it to construct their future. Language is pivotal in the areas of human rights protection, good governance, peace building, reconciliation, and sustainable development" ("The Role of the Language"). So it is critical that we protect and preserve minority, regional and oppressed languages.

In this research paper, I will explore the IT systems involved, their current and potential applications to this research topic, and the social and ethical issues that need to be considered in such applications. For deeper analysis, I will explore what global companies like Google, Microsoft, and Facebook are doing in this space, the challenges and limitations of the solutions/models they have developed, and how these models can be leveraged or enhanced to accomplish language preservation. Since these models depend upon the availability of a significant amount of text data, I will explore emerging technologies, how alternative sources of data can be used, and the ethical issues that need to be dealt with in tapping into those data sources. For my research, I have relied on academic and industry research, opinions of thought leaders, self-study (secondary sources), and interviews with industry executives (primary sources) who are developers or users of applications of NLP in their businesses, and as such, delve into related limitations and ethical issues. In this paper, I use minority, regional, and oppressed languages interchangeably.

IT System - Natural Language Processing (NLP)

What is NLP?

NLP is a branch of Linguistics and Artificial Intelligence (AI) that focuses on how computers can interact with natural languages. This happens through computers analyzing swathes of data, determining patterns, and then applying those patterns to translate and summarize texts, extract information from texts, read texts, and understand them, amongst other things.

How do NLP Models Work?

According to SAS, a data analytics company that works in this area summarizes the procedure of NLP models as follows: NLP models generate linguistic text outlines that identify the relevant words in the texts, followed by some advanced analytics to figure out overall meaning and themes; after that, the models evaluate the contexts of the texts and derive moods or opinions in the texts; following this, the models make the necessary text to speech or speech to text conversions; finally, they produce text summaries and can translate texts between different languages. Daniel Nelson, an engineer, data analyst, and writer of machine learning and deep learning topics, classifies NLP techniques as either “syntactic,” focusing on structure, or “semantic,” focusing on meaning (Nelson).

There are various models developed by different companies. The most well-known are: BERT, developed by Google, and GPT-3, developed by OpenAI. OpenAI was formed in 2015 by Elon Musk, Sam Altman, and others, and it was subsequently invested by Microsoft. NLP models such as BERT and GPT-3 can learn languages, taking into account register, text type, and audience, amongst other factors. They can produce text at the same levels as people, including some of the nuances of human expression. However, they are not perfect and do make mistakes

from time to time. Irrespective of these mistakes, a problem with these systems is that they need an abundant amount of training data to develop an understanding of language at this level. For example, GPT-3 was trained on a data set of almost a trillion words gleaned from across the internet, including many major corpora of texts (Brown et al.).

This poses a problem in the development of models for minority languages as they have a negligible internet presence or other sources of digitized data (websites, emails, text messages, chats, social media posts, and ebooks) used by NLP models. Consequently, there is a limited amount of data, which is the critical raw material for training NLP models.

Defining Common Terms

"Model architectures" are general approaches to solving AI problems, rather than a specific model. A "neural network" is a common model architecture that is designed similarly to how biological nervous systems work. A "transformer" is a model architecture in many NLP models that relies on a component used to model long-range interaction to draw global dependencies between input and output. Previous model architectures were based on various neural networks.

A "parameter" is a variable in a model or a neural network that is fine-tuned with data to help the model or neural network make decisions and create outputs. Having more parameters improves the performance of a model. A "hyperparameter" is a special parameter that is controlled and set before the model is run. Depending on how the hyperparameter is set, the parameters will adjust differently and can take different values. "Tokens" are the smallest unit of a text. Usually, they are words that are important and contribute meaning to the text. Training a model on more tokens gives it better quality results. "Training data" are the colletion of tokens, corpora, and other training material used to train models.

A “vector space” or “embedding space” is a hypothetical space with many dimensions where words are placed and represented as vectors which are easier for NLP models to work with. The placement and proximity of the vectors can help determine relationships between words. An “encoder” is used to create a vector for a word into a vector space, and a “decoder” takes a vector from the vector space and translates it into a word.

“Machine Learning Operations” (“ML Ops”) is designed to streamline the release cycle of machine learning and software applications. It provides automated testing of machine learning artifacts (e.g. data validation, ML model testing, and ML model integration testing). Agile principles can be applied to machine learning projects using MLOps. MLOps must be independent of language, framework, platform, and infrastructure.

Applications of NLP to Preserve Minority, Regional, and Oppressed Languages

Common NLP Models and Techniques

BERT (Bidirectional Encoder Representations from Transformers)

- BERT is a transformer model that is used by Google for Google Search and since late 2020 is being used in almost every English search.

GPT-3 (Generative Pre-trained Transformer 3)

- GPT-3 is another transformer model with 175 billion parameters. Until May 2021, it was the largest NLP model.

Word2vec

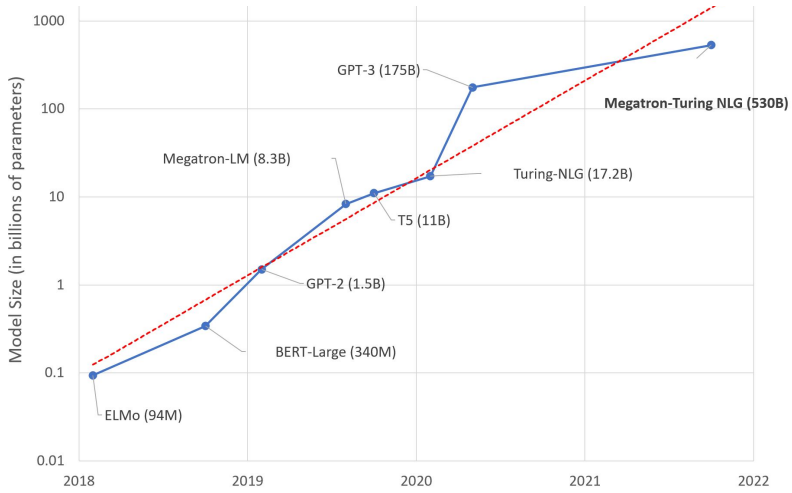
- Models like word2vec take in texts and output vector spaces, each typically containing hundreds of dimensions. Each word within a text corresponds to a vector in the respective vector space.
- Many word relationships can be determined or solved by applying simple algebra to the vectors. For example, vector(“husband”) - vector(“man”) + vector(“woman”) will result in a vector that is very close to vector(“wife”). Similarly, vector(“shortest”) - vector(“short”) + vector(“tall”) will result in a vector that is very close to vector(“tallest”).

Challenges and Limitations of Current NLP Models

All these models rely on a significant amount of text data that these companies have gathered over the years from web pages, emails, social media, and queries. In all my secondary research, the competition was who used more data rather than who produced better results with fewer data. Even the latest models like Microsoft and Nvidia’s Megatron-Turing Natural Language Generation model (MT-NLG) released in October 2021, rely on an enormous amount

of data. MT-NLG has over 530 billion parameters (see figure 1) — more than three times the parameters of GPT-3 — and combines "data, pipeline, and tensor-slicing based parallelism" methods to efficiently use memory, compute, and parallel GPUs (Naik).

Figure 1:



Source: "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model." Microsoft Research, 11 Oct. 2021, https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/.

These companies pride themselves on how much more they can achieve with an increasing amount of data (see table 1). Unfortunately, not even a fraction of this quantity of data is available to train the models in minority languages.

Table 1

Model	Total train compute (PF-days)	Total train compute (flops)	Params (M)	Training tokens (billions)
T5-Small	2.08E+00	1.80E+20	60	1,000
T5-Base	7.64E+00	6.60E+20	220	1,000
T5-Large	2.67E+01	2.31E+21	770	1,000
T5-3B	1.04E+02	9.00E+21	3,000	1,000
T5-11B	3.82E+02	3.30E+22	11,000	1,000
BERT-Base	1.89E+00	1.64E+20	109	250
BERT-Large	6.16E+00	5.33E+20	355	250
RoBERTa-Base	1.74E+01	1.50E+21	125	2,000
RoBERTa-Large	4.93E+01	4.26E+21	355	2,000
GPT-3 Small	2.60E+00	2.25E+20	125	300
GPT-3 Medium	7.42E+00	6.41E+20	356	300
GPT-3 Large	1.58E+01	1.37E+21	760	300
GPT-3 XL	2.75E+01	2.38E+21	1,320	300
GPT-3 2.7B	5.52E+01	4.77E+21	2,650	300
GPT-3 6.7B	1.39E+02	1.20E+22	6,660	300
GPT-3 13B	2.68E+02	2.31E+22	12,850	300
GPT-3 175B	3.64E+03	3.14E+23	174,600	300

Source: Brown, Tom B., et al. “Language Models Are Few-Shot Learners.” ArXiv:2005.14165 [Cs], July 2020. arXiv.org, http://arxiv.org/abs/2005.14165.

While existing NLP models available for a dominant language could theoretically be used for a related minority language in the same language family, it can only help so much. Due to the effects of language evolution, the relationships between two languages — if there even is one — are not always clear. This could be because many languages from different sub-families may be very similar and have begun to resemble one another through long-term close contact. Dialect continua cause further confusion, as neighboring dialects may be mutually intelligible, but dialects at the beginning and end of a continuum may not be. If we don't know the relationships between languages, we cannot express that, or take that into account, when training multiple languages in NLP models.

Current Industry Applications - Interview Findings and Analysis

I interviewed several people from the industry – Amazon, Meta (formerly Facebook), Mashreq Bank, PeopleStrong and ShareChat to get their expert views on my research topic. All of them acknowledged that it is a very hard problem to crack which they are seeking to solve for to increase their customer base and also to develop differentiated capabilities compared to their competitors.

As per Mr. Walia of PeopleStrong, he is dealing with this problem in his company which want to expand into South East Asia for non-English speaking customers. Without enormous amount of training data, he cannot build the desired NLP models. To overcome the problem, he needs to first develop a lexical dictionary of the language; the writing systems don't really matter. Since he doesn't have one, he usually gets this with the help of Google Translate which achieves around 60 to 70% accuracy. Without the help of Google Translate or a lexical dictionary of any language, the NLP models cannot be developed. He called it the "Cold Start" Problem".

So the first challenge to solve for developing any NLP model is to have a lexical dictionary – either it has to be developed by humans or, if one exists in text form, it needs to be digitized. The second problem is to gather data to train the models. From my interviews I learned that many speakers of minority languages only use spoken language – they cannot read or write. Even if there were text-based NLP models, it will not solve the problem for them. So many companies are now building conversational NLP models. They are also using voice, video, and gamification to capture data as collecting text-based data was impractical and will take several decades.

Mr. Mukherjee of ShareChat confirmed that they have developed social media and interaction platform in 10 major regional languages of India, but it required significant

investment running in hundreds of millions of dollars and human involvement. They also use video and voice recordings to capture the data and train their models. To overcome the data quantity problem, for data gathering, they are applying principles of MLOPs so they focus on capturing quality of data rather than quantity of data. This requires more and active human engagement. Mr. Rebello of Mashreq Bank gave similar feedback. During my interview with Ms. Sinha of Amazon's Alexa team, she said that Alexa has a skill called "Cleo" that utilizes conversational gamification in order to teach Alexa new languages. However, she also mentioned that the process of gamification has its own set of challenges pertaining to reliability and accuracy.

Since it will be hard to generate lots of text data through traditional methods, companies are trying to convert speech captured from video and audio recordings into text that is understandable by machines. Ms. Banda of Facebook informed me that in September 2021, Facebook unveiled a speech-based NLP model called the Generative Spoken Language Model (GSLM). GLSM inputs raw audio signals without tags or labels. It doesn't need large datasets like text-based NLP models, and can even access an extensive part of human expression that is not clearly transmittable through text. Most of the world's languages don't have data sets that are this varied. This also has the added benefit of including languages lacking a written form or a significant corpus of text to be preserved and also creates diversity and nuance in the data collected (Textless NLP). GSLM uses a speech encoder called wav2vec, analogous to word2vec for text, "with just 10 minutes of transcribed speech and [53 thousand] hours of unlabeled speech, [has]... a word error rate (WER) of 8.6 percent on noisy speech and 5.2 percent on clean speech".

I conducted a brief experiment in the supervision of Mr. Walia to see the effects of having a lack of training data to train a model. I picked a video of people speaking Awadhi that lasted around five minutes — less than 0.01% of the amount of data used to train the encoder and many times more data than what a language even has to offer — and split it into audio files that were 20 seconds long. After that, I ran code from the GSLM model — which can achieve more with less data — and ran the audio files through it to train the model. When I tried interacting with the model I had just trained, the results came out as gibberish and made no sense whatsoever. This reinforced my finding that if a language doesn't have enough data, an NLP model cannot be developed.

Social and Ethical Significance for Stakeholders

Social Impacts

The internet has proliferated in most parts of the world, however, English and a few major world languages are the dominant languages of the internet. This excludes a large part of the world population that does not speak these languages. Over the coming years, a majority of internet users will be native speakers of a regional language (Jaiswal). It is important to keep up with this trend because, as with any advancement, this will help determine the shape that the future of the internet — and Web 3.0 — will take, the skills people will need to acquire to be future-ready, and how people will interact with one another. The benefit of the internet has to be fully leveraged to educate them, employ them, and also engage them in social and political discourse, making integrating minority languages on the internet essential. This can only be done with the application of NLP.

According to researchers Dhana L. Rao, Ethan Smith, and Venkat N. Gudivada, learning in one's mother tongue is the best way to acquire knowledge and skills, especially at a young age. This, and constantly being exposed to their mother tongue, "helps in the mental, moral, and emotional development of children." Preserving these languages and using them in education will help with literacy rates (Rao et al. 355-356).

Furthermore, this issue is time-sensitive: About every fourteen days a language dies (Russ Rymer), and there are fewer and fewer opportunities to save these languages. Through the loss of language comes the loss of speakers' associated cultures and identities.

Also, as per the Education Review Office of New Zealand, linguistic diversity can lead to social and economic benefits. Economic innovation and entrepreneurialism are possible results of such diversity (Addressing Cultural and Linguistic Diversity | Education Review Office).

However, it is not just beneficial to society; it signifies a privilege that not everyone gets the luxury of having. Nevertheless, because of dominant languages, such as English, digital tools, products, and services are unavailable in many native languages. This barrier excludes millions of people from using or having access to them. As global and local companies are looking to either expand their customer base by including erstwhile excluded from their customer base because these potential customers (millions in number) only know regional and minority languages and many don't even know how to read and write. Given these language barriers, they have been attempting to leverage NLP technology to develop solutions in vernacular languages.

Ethical Considerations

Surveillance and Data Privacy

As companies are developing alternative approaches to capture data in the absence of communication in text format (recording audio and video interactions and gamification), some of the biggest issues that need to be addressed are the invasion of personal privacy and surveillance. While this is also relevant for text-based data sources, for audio and video it becomes particularly acute. The government should have appropriate laws and companies must have strict policies to prevent abuse. For example, NAYAN, a vision technology company in Dubai that provides video-based traffic surveillance, has a policy framework to ensure that the data is anonymized and deleted automatically after a certain amount of time. Data that is relevant for traffic violations are retained for longer and can only be accessed only with 2 people authenticating simultaneously. Similarly, any text, video, and audio data should be retained in an anonymized form and access should be regulated. Peoples' consent should be obtained and they should have the option of deleting any data, similar to how Google and Microsoft users can delete data gathered on them.

Generation and Spread of Misinformation

NLP models can be used to create convincing, false, and misleading information better than humans can, and at scale. The most glaring example of this problem is the 2021 Facebook investigation, where technology is only able to capture only a limited amount of false and misleading information. Facebook has hired thousands of employees to manually scan for inappropriate posts, as machines cannot do this effectively yet. This problem is even worse when it comes to non-English languages. A live case study is all the challenges Facebook is facing in India, where it is used to spread hate speech and communal violence. “Facebook’s problems on the [Indian] subcontinent present an amplified version of the issues it has faced throughout the world, made worse by a lack of resources and a lack of expertise in India’s 22 officially recognized languages” (Frenkel and Alba).

Bias

As new NLP models are developed on the back of existing models for related predominant languages, the issues of inherent bias and loss of perspective and uniqueness that makes two languages (even if related) distinct remain. NLP models even for dominant languages have not been able to solve many of these biases (gender, race, community, religious, etc.) and actually perpetuate them as they pick the biases from the data they are trained on.

According to Stanford University’s Human-Centered Artificial Intelligence, “GPT-3 can exhibit undesirable behavior, including known racial, gender, and religious biases.”

The historical misclassification of regional languages and the spread of newer technologies, such as “radios, televisions, film, and now the Internet,” have had a long-lasting, detrimental effect on these regional languages. In India, Hindi and English have become the dominant languages of these media, as well as in education, government, and commerce, which

further reinforces these misconceptions and makes it hard to preserve minority languages. This follows a common pattern and leads to the dominating language permeating homes and communities.

Digital Divide

Making more accessible helps create equal access to opportunities to the under-represented communities of minority language speakers. This helps companies serve larger target markets, which in turn, makes their products and services available to these communities. This also allows more people to access the internet, which gives people who speak minority languages to learn more, and share their own knowledge that may be unique to them. Another benefit is that it may bring together communities that have been separated by geography due to globalization or past historical events.

Conclusion

NLP can certainly play a significant role and perhaps is the only hope in the preservation of minority languages. However, it cannot solve the problem of preserving minority languages in the absence of abundant data. While this is a major challenge, recent breakthroughs and developments (some still in progress) by companies who want to increase their customer base are solving both the data capture problem and trying to develop efficient NLP models that can be developed with fewer data. They are utilizing speech and video to capture non-text data, and applying clever techniques like multi models, ML Ops, metaverse, and gamification. This gives me optimism that the data and technical challenges will be overcome in the near future. Irrespective, proactive human engagement will still be necessary, not only to address the technical problems, but more importantly, for the effective capturing and preserving of the unique nuances and perspectives of minority languages, and for preventing its misuse. This can only happen by continuous human vigilance which has no substitute. Given the significance of preserving languages, instead of succumbing to the economic and other pressures, as is in the case of Awadhi, it is incumbent on individual communities to actively leverage and make the best use of existing and evolving NLP and other technologies to preserve their heritage and culture — and stay actively engaged to overcome and prevent its shortcomings.

Works Cited

"Addressing Cultural and Linguistic Diversity." *Education Review Office*, https://www.ero.govt.nz/publications/responding-to-language-diversity-in-auckland/addressing-cultural-and-linguistic-diversity/#benefits-and-challenges-of-diversity.

Baevski, Alexei, et al. *Wav2vec 2.0: Learning the Structure of Speech from Raw Audio*. https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/.

Brown, Tom B., et al. "Language Models Are Few-Shot Learners." *ArXiv:2005.14165 [Cs]*, July 2020. *arXiv.org*, http://arxiv.org/abs/2005.14165.

Census of India 2011. Office of the Registrar General & Census Commissioner, 2011, p. 52.

D'Monte, Leslie. *'Search Provides a Very Fundamental Benefit to Society': Google Search Rankings Head, Pandu Nayak*. 18 Nov. 2021, https://www.techcircle.in/2021/11/18/search-provides-a-very-fundamental-benefit-to-society-google-search-rankings-head-pandu-nayak/.

Frenkel, Sheera, and Davey Alba. "In India, Facebook Grapples With an Amplified Version of Its Problems." *The New York Times*, 23 Oct. 2021. *NYTimes.com*, https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html.

Grierson, George A. *Linguistic Survey of India*. 1911. *Internet Archive*, http://archive.org/details/LSIV0-V11.

Iyer, Brijesh, et al., editors. *Applied Computer Vision and Image Processing: Proceedings of ICCET 2020, Volume 1*. Springer Singapore, 2020. *DOI.org (Crossref)*, https://doi.org/10.1007/978-981-15-4029-5.

Linguistic Society of America. *"Monolingual Fieldwork" Demonstration - Daniel Everett*. *YouTube*, https://www.youtube.com/watch?v=sYpWp7g7XWU.

Nelson, Daniel. "What Is NLP (Natural Language Processing)?" *Unite.AI*, 8 Nov. 2019, https://www.unite.ai/what-is-natural-language-processing/.

Rani, Sujata, and Parteek Kumar. "A Journey of Indian Languages over Sentiment Analysis: A Systematic Review." *Artificial Intelligence Review*, vol. 52, Aug. 2019, pp. 1–48. *ResearchGate*, https://doi.org/10.1007/s10462-018-9670-y.

Rymer, Russ. "Vanishing Voices." *National Geographic*, July 2012, https://www.nationalgeographic.com/magazine/article/vanishing-languages.

Tamkin, Alex, and Deep Ganguli. "How Large Language Models Will Transform Science, Society, and AI." *Stanford HAI*, 5 Feb. 2021, https://hai.stanford.edu/news/how-large-language-models-will-transform-science-society-and-ai.

Textless NLP: Generating Expressive Speech from Raw Audio. https://ai.facebook.com/blog/textless-nlp-generating-expressive-speech-from-raw-audio/.

"The Role of the Language." *2019 - International Year of Indigenous Language*, https://en.iyil2019.org/role-of-language/.

"Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model." *Microsoft Research*, 11 Oct. 2021, https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/.

Visengeriyeva, Larysa, et al. "Ml-Ops.Org." *MLOps*, https://ml-ops.org/.

"What Is Natural Language Processing?" *SAS*, https://www.sas.com/en_nz/insights/analytics/what-is-natural-language-processing-nlp.html.