Bridging Language Gaps: Cross-Linguistic Translation Strategies for Indic Languages

Diksha Seth & Sachin Kumawat & Sofia Sunam
Indian Institute of Science
Bengaluru, KA, India
{dikshaseth,sachinkumawa,sofiasunam}@iisc.ac.in

1 Problem Statement

India is a linguistically diverse country with numerous local languages. Despite this diversity, these local languages are low-resourcedAdelani et al. (2022) and there are limited resources available for machine translation among these languages. This project aims to address this gap by developing a better machine translation model that translates Hindi (*may be changed later) to other local Indic languages.Ramesh et al. (2021)

Language translation is the key to make **News Media**Adelani et al. (2022) more accessible and spreading awareness .It's especially useful for people living in areas where their native language is not commonly spoken, as it allows them to engage with and understand local news more effectively. In context of India's **Electoral Campaigns**, it is crucial for political parties to spread their messages to a linguistically diverse audiences. A potential strategy can be to develop a universal message and then translate it into numerous regional languages, optimizing the communication efforts. Machine translation plays a significant role in **Movie Captioning** Team et al. (2022), especially for local Indic languages. Many **Healthcare Workers** Team et al. (2022) are not educated enough to communicate in English or it is difficult for the patients to convey their issues in English. So **Doctors** could use the MT to know the problems. Healthcare manuals could be prepared in one language and translated to others. Ultimately, **Improved Communication** leads to greater **harmony**, as it helps bridging the gap among people.

Moreover, within the expansive Indian film industry, the necessity to broaden audience reach prompts the need for dubbing local language films. Leveraging machine translation (MT) technology offers a viable solution to streamline and economize this process. Similarly, the dissemination of educational content demands localization in various regional languages to ensure inclusivity and accessibility. This will lay the foundations of better India.

Furthermore, in the healthcare sector, language barriers between patients and health workers pose significant challenges. Many health professionals may lack proficiency in English, hindering effective communication. Conversely, patients may struggle to articulate their concerns in a language unfamiliar to them. Implementing MT facilitates communication between doctors and patients, thereby enhancing diagnostic accuracy and patient care. Additionally, the preparation of health manuals in one language with subsequent translation mitigates language-based impediments in healthcare delivery.

Ultimately, the widespread adoption of language translation technologies fosters improved communication and understanding among diverse linguistic communities. This advancement contributes to societal harmony by bridging linguistic divides and fostering inclusive dialogue.

2 Related Work

Machine Translation is a promising field of research in NLP and people are pouring efforts to efficiently and effectively translate both high and low resource languages. Few of the research papers are discussed below:

- No Language Left Behind: Scaling Human-Centered Machine Translation
  This paper forms the basis of our project, providing the foundations for our exploration into the challenges of machine translation.Team et al. (2022). It created dataset and improved performance gaps between high and low resource languages by developing a conditional compute model based on Sparsely Gated Mixture of Experts. It is trained on novel data obtained from mining techniques for low-resource languages.
- Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages
  The demand for quality, publicly available translation systems in a multilingual society like India is obvious. Ramesh et al. (2021) While MT has been favourable for resource-rich languages, there has been limited benefit for resource-poor languages which lack parallel corpora, monolingual corpora, and evaluation benchmarks. This paper improved the existing corpora to currently large dataset supporting 49.6M sentence pairs between English to Indian Languages.

- **A Few Thousand Translations Go A Long Way! Leveraging Pre-trained Models for African News Translation** Adelani et al. (2022) :
  This paper focused on two questions:
  1) How can pretrained models be used for languages not included in the initial pre-training? and 2) How can the resulting translation models effectively transfer to new domains?Adelani et al. (2022)
  Their contributions were the following:
  - They created a new African news corpus for machine translation covering 16 African languages.
  - They adapted several multilingual pre-trained models (MT5, ByT5, mBART, M2M-100) to these largely unseen languages, and evaluated their quality on news translation.
  - They quantified the effectiveness of small in-domain translation sets by measuring domain transfer effects and comparing fine-tuning strategies.
- **Findings of the 2021 Conference on Machine Translation (WMT21)**Akhbardeh et al. (2021)
  This paper presents the results of the news translation task, the multilingual low-resource translation for Indo-European languages, the triangular translation task, and the automatic post-editing task organised as part of the Conference on Machine Translation (WMT) 2021.Akhbardeh et al. (2021)
  This conference presented several key findings:News Translation Task,Multilingual Low-Resource Translation,Traingular Translation Tast,Automatic Post-Editing Task
  Other areas of focus included large-scale multilingual machine translation, machine translation using terminologies, metrics, quality estimation, and unsupervised and very low-resource translation.

- **On the use of Comparable Corpora to improve SMT performance**Abdul-Rauf & Schwenk (2009): Statistical Machine Translation uses the Noisy Channel model build on small amounts of parallel texts for translation. It may happen that the alignment of the source and targeted languages are not same. Thus sentence aligned parallel corpora is required for SMT translation.
- **TICO-19: the Translation Initiative for COvid-19**Anastasopoulos et al. (2020): Given the pandemic times, it was highly crucial to convey the information across multiple countries and language. This paper was an initiative to translate the information on Covid-19 across 26 low resourced languages. This discusses on creating of corpora for later use by other researchers.
- **The AMARA Corpus: Building Parallel Language Resources for the Educational Domain**Abdelali et al. (2014): This paper highlights the need of translation of high resource languages to low resource languages in the educational sector. It has created 20 monolingual and 190 parallel bilingual corpora to enhance translations in eductaional domain. In house crawler was used. Training phase included building

word aligments and using grow-diag-final and heuristics to make it symmetric. The dataset was tuned based on BLUE score and monotone-at-punctuation decoding was done.

- **The Effect of Domain and Diacritics in Yorùbá –English Neural Machine Translation**
  The difficulty of evaluating MT models on low-resource pairs is often due to lack of standardized evaluation datasets. Adelani et al. (2021) Low-resource languages such as Hi, Or, Be etc, one can only find few thousands of parallel sentences online.
- **Machine Translation for African Languages: Community creation of datasets and models in Uganda**
  They created a parallel text corpus, SALT for 5 Ugandan languages and various methods were explored to train and evaluate translation models. Akera et al. (2022)

3 Evaluation of Results

Evaluation of models with Machine Translation is still an active field of research. A few of the existing evaluation criteria are discussed below to see the accuracy of MT models. The model shall be evaluated based on the BLEU (Bilingual Evaluation Understudy) score and TER (Translation Error Rate)Akhbardeh et al. (2021). Human Evaluation can also be a part of this criteria.

Bleu measures the similarity of two sentences. It will be used to find the similarity between the translated sentence and the test data from parallel corpora.

Translation Error Rate (TER) measures the amount of changes that someone would perform so that machine translation matches with real translations.

For the Human Evaluation, we shall put a survey so that the people can verify the translations on a graded scale.

Contributions

The authors have equally contributed to the writing of the proposal. Other tasks performed for this proposal were researching and finalising topics and collecting bibliographies.

The following papers have been read by each one of us. Please find the reference of each in the Reference section.
Sachin-

- A Few Thousand Translations Go A Long Way! Leveraging Pre-trained Models for African News Translation
- Findings of the 2021 Conference on Machine Translation (WMT21)
- Machine Translation Systems for Indian Languages: Review of Modelling Techniques, Challenges, Open Issues and Future Research Directions

Sofia-

- On the use of Comparable Corpora to improve SMT performance
- TICO-19: the Translation Initiative for COvid-19
- The AMARA Corpus: Building Parallel Language Resources for the Educational Domain

Diksha-

- Towards a Cleaner Document-Oriented Multilingual Crawled Corpus
- The Effect of Domain and Diacritics in Yorùbá –English Neural Machine Translation

- Machine Translation for African Languages: Community creation of datasets and models in Uganda

All -

- No Language Left Behind: Scaling Human-Centered Machine Translation
- Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. The AMARA corpus: Building parallel language resources for the educational domain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 1856–1862, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/877_Paper.pdf.

Sadaf Abdul-Rauf and Holger Schwenk. On the use of comparable corpora to improve SMT performance. In Alex Lascarides, Claire Gardent, and Joakim Nivre (eds.), *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pp. 16–23, Athens, Greece, March 2009. Association for Computational Linguistics. URL https://aclanthology.org/E09-1003.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3053–3070, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL https://aclanthology.org/2022.naacl-main.223.

David I. Adelani, Dana Ruiter, Jesujoba O. Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Awokoya, and Cristina España-Bonet. The effect of domain and diacritics in yorùbá-english neural machine translation, 2021.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. Machine translation for african languages: Community creation of datasets and models in uganda. In *3rd Workshop on African Natural Language Processing*, 2022. URL https://openreview.net/forum?id=BK-z5qzEU-9.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. Findings of the 2021 conference on machine

translation (WMT21). In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (eds.), *Proceedings of the Sixth Conference on Machine Translation*, pp. 1–88, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.1.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federman, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. Tico-19: the translation initiative for covid-19, 2020.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Deepak Kumar, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *CoRR*, abs/2104.05596, 2021. URL https://arxiv.org/abs/2104.05596.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.