Low resource language identification

- Supervisor: Molly Kennedy
- Examiner: Prof. Hinrich Schuetze
- Open for: BSc, MSc
- General Topic Area: Language Identification

Prerequisites:

- Enthusiasm
- Good programming skills particularly in Python -
- Preferably good understanding of ML/NLP
- Familiar with libraries like numpy, pandas, and scikit learn

Summary

Extensive research has been carried out on high resource languages, languages that are widely spoken and for which there is an abundance of data and resources such as English and Spanish. However, research of low resource languages is much more challenging due to data and tool constraints.

This project aims to enhance the capability of language identification systems to better recognise and understand low resource languages.[1, 2] This would in turn improve communication and information access for people who speak these languages and contribute to the preservation and study of linguistic diversity [3]. Possible tasks include:

- Investigating how models trained on high-resource languages can be adapted to low-resource languages through transfer learning. This includes studying the effectiveness of multilingual models like mBERT or XLM-R in identifying low-resource languages.
- Exploring how cross-lingual word embeddings and sentence representations can aid in language identification, especially for languages that lack extensive labeled datasets.
- Utilizing crowdsourced data collection and annotation efforts, possibly in collaboration with native speakers and communities.
- Gathering data in low resource languages, cleaning and preprocessing data to ensure quality and consistency.
- Exploring and implementing machine learning algorithms suited for LID, particularly those efficient in handling small datasets.

References

[1] S. Gaikwad, T. Ranasinghe, M. Zampieri, and C. M. Homan, "Cross-lingual offensive language identification for low resource languages: The case of marathi," *arXiv preprint arXiv:2109.03552*, 2021.

[2] T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification for low-resource languages," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, pp. 1–13, 2021.

[3] J. Basu, S. Khan, R. Roy, T. K. Basu, and S. Majumder, "Multilingual speech corpus in low-resource eastern and northeastern indian languages for speaker and language identification," *Circuits, Systems, and Signal Processing*, vol. 40, pp. 4986–5013, 2021.