Language Identification Challenges and Solutions

- Topic: Language Identification Challenges and Solutions
- Supervisor: Amir Hossein Kargaran
- Examiner: Prof. Hinrich Schütze
- Open for: MSc/BSc

Description

Goal: This project aims to identify challenges that language identification (LID) models face, experimentally demonstrate these challenges, uncover their root causes, and propose solutions.

Why LID: LID is a widely studied problem in NLP, but certain challenges remain yet unsolved. For instance, LIDs trained in specific domains may perform poorly in others, and variations in orthography can impact performane.

Solution: The project first part will start by experimentally identifying challenges faced by LID models. Examples include:

- Domain dependency: Training an LID on one domain source (e.g., religious) and testing it on other domains (e.g., wikipedia), reporting scores, analyzing failures, and identifying mistake roots such as code-switching, different vocabulary, unlucky frequent n-gram or orthographic variations.
- Coverage: Investigating how LIDs handle languages not covered in their training set, examining predictions for unsupported languages, and analyzing confidence scores.
- Not-a-Language: Analyzing LID responses to random characters, misrendered PDFs, or general noise.

The project second part involves proposing solutions to address these challenges, such as modifying training data or exploring different method (e.g., a hierarchical approach, sequence to sequence architectures).

Prerequisites

- Enthusiasm (for publishing results at a conference/workshop)
- Proficiency in speaking and writing English
- Good Python programming background (e.g., knowledge of numpy and pandas libraries)
- Basic knowledge of ML/NLP (e.g., understanding how a classifier works, knowledge of transformer architecture)
- Basic command of PyTorch and Transformers libraries is recommended

Supervisor

Hello, I am Amir. If you choose this project, I will be your supervisor. You will receive 30 minutes per week of guidance and help, accumulable up to 3 weeks. Our work begins with a comprehensive literature review of the task and available resources. I will ensure you receive the allocated time from my side. If I see potential in your work, I am willing to offer additional assistance. For any questions, contact me directly: amir@cis.lmu.de