Exploring Language Representation through a Resource Inventory Project

Carolyn Jane Anderson
Wellesley College
Wellesley, MA
carolyn.anderson@wellesley.edu

1 Introduction

The increasing scale of large language models has led some students to wonder what contributions can be made in academia. However, students are often unaware that LLM-based approaches are not feasible for the majority of the world's languages due to lack of data availability. This paper presents a research project in which students explore the issue of language representation by creating an inventory of the data, preprocessing, and model resources available for a less-resourced language.

Students are put into small groups and assigned a language to research. Within the group, students take on one of three roles: dataset investigator, preprocessing investigator, or downstream task investigator. Students then work together to create a 7-page research report about their language.

2 Course Context

This assignment is the midterm research project for an advanced undergraduate Natural Language Processing course. Before the project, the class covers text processing and non-neural NLP techniques roughly corresponding to the first six chapters of Jurafsky and Martin (in prep.). Students have two weeks to work on this project before presenting their findings to the class.

3 Learning Goals

This assignment is designed to engage students with issues of linguistic representation. The primary goal is for students to explore the availability of data resources for a less-resourced language. Along the way, students build useful skills in how to locate and evaluate data and model resources, which are applicable outside of the context of low-resource languages as well.

The learning goals for this project are as follows:

- Explore issues of language representation
- Gain familiarity with dataset and model hubs
- Analyze the quality and availability of software artifacts
- Collaborate with classmates to present research findings
- Practice scientific writing and presentation

4 Language Selection and Assignment

For this project, it is important to select languages that have some available data and model resources. For this reason, I refer to the languages as "less-resourced" rather than "low-resource".

Table 1 shows the languages used in a prior semester. To create the language groups, I surveyed students on their language backgrounds and when possible, seeded each group with a student literate in the language or a related language. Unless all group members were literate in another writing system, I assigned only languages that used the Latin alphabet.

5 Assignment Structure

Each student was assigned one of three roles: dataset investigator, preprocessing investigator, or downstream task investigator. Each team worked together to prepare a 7-page report and a 3-minute in-class presentation of their findings. Students were responsible for writing a two-page section of the report on their individual topic, as well as for collaborating on a one-page introduction to their language. As a result, the project gives students a chance to practice teamwork and collaboration, while allowing the instructor to assess their effort individually.

5.1 Language Introduction

Each report was required to begin with an introduction to the language and its context. This section described the language's communities of use and social context, its writing system, and its morphology and syntax, including some sentences with

Language	Writing System	Language Expertise
Indonesian	Latin	1 Indonesian-literate student
Somali	Latin	1 Arabic-literate student (significant lexical borrowing from Arabic)
Japanese	Kanji/Kana	All students literate in Japanese
Afrikaans	Latin	2 German-literate students (significant lexical borrowing from German)
Haitian Creole	Latin	All students literate in French (significant lexical borrowing from French)
Romanian	Latin	None
Portuguese	Latin	2 students literate in Portuguese

Table 1: Example language groups

word-by-word translations.

5.2 Dataset Investigator

Dataset investigators were required to explore a number of platforms to investigate the availability of data resources for their language.

Students were asked to evaluate Wikipedia as a language resource, reporting on the existence, size and robustness of Wikipedia in their assigned language. Students were also required to search Kaggle, Hugging Face, and Github for other datasets in their language. They were asked to describe each dataset that they found and to consider multiple aspects of its utility: accessibility, quality, and size.

Students reported many challenges around accessibility: they found research papers reporting the use of a dataset, but couldn't find the dataset itself, or they found the dataset's website, but the links were broken. Quality was the most challenging aspect for them to evaluate, especially for groups without a member who was literate in the language.

5.3 Preprocessing Investigator

Preprocessing investigators explored text processing tools for their language. The suggested tasks included tokenization, segmentation, part-of-speech tagging, and parsing.

Students were asked to evaluate whether the popular NLP libraries NLTK (Bird and Klein, 2009) and SpaCY (Honnibal et al., 2020) provided any tools for their language. Students were also asked to look for other preprocessing tools on Github or other websites. Several students leveraged search tools for academic research papers, and found relevant systems via a literature search, an effective approach that I hadn't anticipated.

For each preprocessing tool that they discovered, students described how it worked, what task it was designed for, what data it was trained on, and assessed its usability. Students had the most trouble finding information about the data used to train the tools. They also ran into many instances where the code could no longer be downloaded or run.

5.4 Downstream Investigator

Downstream task investigators looked into the availability of systems in their language for downstream tasks such as named-entity recognition, event recognition, language modeling, sentiment analysis, question-answering, and machine translation.

Students were required to look for models on Github and Hugging Face, and were also encourage to do a general Web search. As above, for each preprocessing tool that they discovered, students described how it worked, what task it was designed for, what data it was trained on, and assessed its usability.

6 Scaffolding for the Final Project

This project serves as scaffolding for the course's final research paper (due at the end of the term), which is individual. Although most students come up with interesting and challenging topics on their own, a handful of students struggle to do so each semester. I encourage these students to build something for the language they investigate in the resource inventory project. This has worked well and led to final projects on sentimental analysis and named entity recognition for Indonesian and hate speech identification and speech recognition for Portuguese.

7 Conclusion

This assignment allows students to explore issues of language representation by conducting a resource inventory for a less-resourced language. The hope is that this project highlights how much work remains for researchers to do on NLP for the breadth of the world's languages.

References

Edward Loper Bird, Steven and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Daniel Jurafsky and James H. Martin. in prep. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd edition.