Don’t necessarily keep the sections: don’t follow the standard template

ZHAN SHEN

Novelty not clearly stated in the novelty section.

The hypothesis behind this proposal is that different languages and speeches might have some at- tributes in common 🡪 maybe state an observable hypothesis, which is a consequence of this common patterns they have e.g. better performance?

I feel the background section is too long and detailed: piazza post saying should be around 0.5 pages. Details about LSTM not sure are necessary here as much as they are on the actual dissertation introduction.

Model architecture: maybe not in the program section?

In terms of datasets, in this project, there are two models need to be trained.

Not sure about all the possible integration methods.

Mapudungun is the ancestral language of the Mapuche people of south-central Chile and Argentina. Today Mapudungun is spoken mostly in parts of Chile by an estimated 144,000 speakers, while in Argentina, the number of self-reportedly competent speakers is around 8,400. In both countries, monolingualism is vanishingly rare, with the range of interlocutors and topics for the use of the language having grown progressively smaller [ref.]. For these reasons, Mapudungun represents an endangered language that is currently being studied and documented.

Our dataset contains a list of historical documents in Mapudungun, spanning the period 1621-1930. Only a small portion of this data is annotated: specifically, one document transcribed between 1895-1898 by the scholar Dr. Rudolf Lenz, who conducted some of the earliest explicitly academic studies of the language, and one document dated 1922 authored by the doctor and linguist Félix José de Augusta.

The challenges that this database poses are low-resource data , an intricate inflectional morphology, and a corpus of texts that were collected through a long historical period and thus present several linguistic variations, specifically in spelling. Training a system on a low-resource dataset represents a challenge as the system will be required to predict a large number of unseen words at test-time. Complex inflectional morphology means that the same lemma can have a wide variety of surface forms, making it harder for the system to pick up patterns and make the right prediction given the surface form. Lastly, the fact that documents have been collected and transcribed at different time periods, spanning from the 17th to the 20th century, means that the language transcription will not be completely consistent, with several spelling variations.

**More about the data set!!!**

**New dataset needs to be adapted to input expected by lematus. Annotated following TEF(?) format. Will parse XML to format training and testing data to the following format:**

**Input that we expect : inflected form with context**

**Output that I will get : lemmas**

**Look at WHETHER DATABASES ARE DIFFERENT and you gonna use them.**

**Maybe experiments**

Evaluation NO🡪 call it analysis

YUSHENG

This project focuses on translating speech from low-resource languages into text in high-resource languages.  
We’ll pre-train the model on multilingual ASR task rather than monolingual ASR task.

Would be nice to break up the introduction into subsections as indicated by the template. For example to highlight hypothesis, novelty, motivation etc. I think the info is all there but a bit scattered.

Honestly super good overall. Especially the experimental planning.

Too much detail in the background but not sure. Same as above.

TODO:

One of the main purposes of data-driven lemmatization is to handle unseen words at test time, yet

languages with differing morphological productivity will have very different proportions of unseen

words.

SPELLING CHANGES. IRREGULAR FORMS?

Require large datasets to work well. This issue has been addressed by employing a different type of attention: Hard attention: in each step, the model attends to a single input state. This modeling suits the natural monotonic alignment between the input and output, as the network learns to attend to the relevant inputs before writing the output which they are aligned to.

The input sequence is the inflected word, as a space-separated character representation of a word in its N-character left and right sentence context.