Web Tool for Phonemes Week-4

University at Buffalo The State University of New York

Agenda

- Word Dataset
- Phoneme Dataset
- Category Data
- Articulation Data
- Semantic Data
- Integration of all data

Word Dataset

- Scraped around 200k+ words

https://raw.githubusercontent.com/matthewreagan/WebstersEnglishDictionary/master/WebstersEnglishDictionary.txt

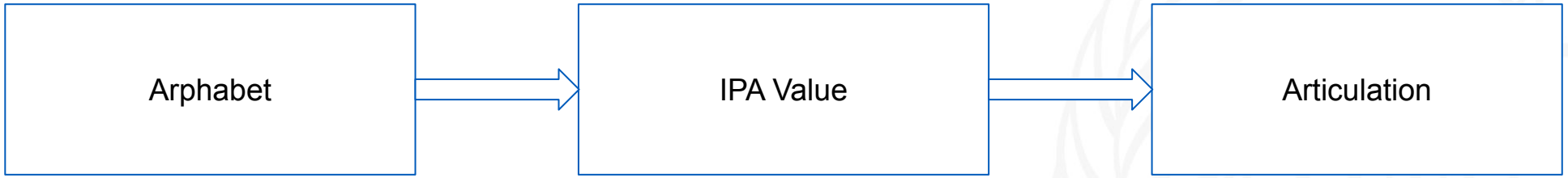
Phoneme Data

- G2P model to generate the Phonemes for the words

Category Data

- Wordnet

Articulation Data



From G2P Model

Convert to IPA

Get Articulation

Example :

R OW Z AH Z

Semantic Data

- Wordnet
- Spacy

  When given an isolated word without context, spaCy's tagger can only guess its most likely part of speech

  whereas WordNet can list all possible parts of speech associated with that word.

Minimal Pairs

- sample queries
- Query in Solr

Integration of all data

Data management is crucial

Future uses

modification of dataset or methods to collect various data

documentation

Questions

1. GPU