Web Tool for Phonemes Week-3

University at Buffalo The State University of New York

Agenda

- Word2Vec
- Finalising on Dataset
- Category Data
- Minimal Pairs
- Word2vec
- Paper
- g2p Model
- Educator Word Frequency Guide

Finalising on Dataset

Dataset :

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

Problems :

1. Nouns present - Eg : Abdul, John,etc
2. Plurals present
3. Large Dataset - More computation

Category Data

Wordnet

Problems :

Some of the categories too specific while some are too broad.

Minimal Pairs

Word2Vec

- Vector representations of the words
- Semantic similarity of the words is represented between their vectors
- Based on cosine similarity of word
- Classification and information retrieval
- dataset = [

  ['P', 'AH0', 'R', 'K'], # PARK

  ['K', 'AH0', 'R'], # CAR

  #]

https://radimrehurek.com/gensim/models/word2vec.html

can use it to analyze the phonemes in various ways, such as finding **similar phonemes**, phoneme analogy tasks, or visualizing phoneme vectors.

Paper on Minimal, Maximal and Multiple Oppositions

- Minimal pairs are pairs that differ in one phoneme
- Maximal Oppositions are the words that have contrasting phonemes which differ in as many distinctive features as possible
- Features are divided in non major class and major class
- Non Major class - Place, Articulation, Manner
- Major class - sonarant vs obstruent
- Multiple Opposition are sets of maximal opposition pairs
- https://www.speech-language-therapy.com/index.php?option=com_content&view=article&id=133&catid=9&Itemid=101

g2p Model

- To facilitate the pronunciation of words or phrases.
- G2P models are trained on linguistic data and often require dictionaries like CMU or datasets that contain word-to-phoneme mappings to learn the rules for converting graphemes to phonemes accurately

  https://huggingface.co/speechbrain/soundchoice-g2p

Educator Word Frequency Guide

- Study of word frequency, and the number of words analyzed (over 17 million tokens and 164,000 types)

- The guide is organized into four sections:
  1. Technical aspects.
  2. Words with frequencies of 1 or greater, including additional statistics by grade level.
  3. The third section covers words with frequencies less than 1
  4. It presents all words from the corpus in descending order of frequency.

Questions

1. Maximal and Minimal Pairs Length
2. Iphod data set - neighbourhood density
3. Word Frequency
4. Any support on Maximal and Minimal pairs
5. Educator Word Frequency Guide