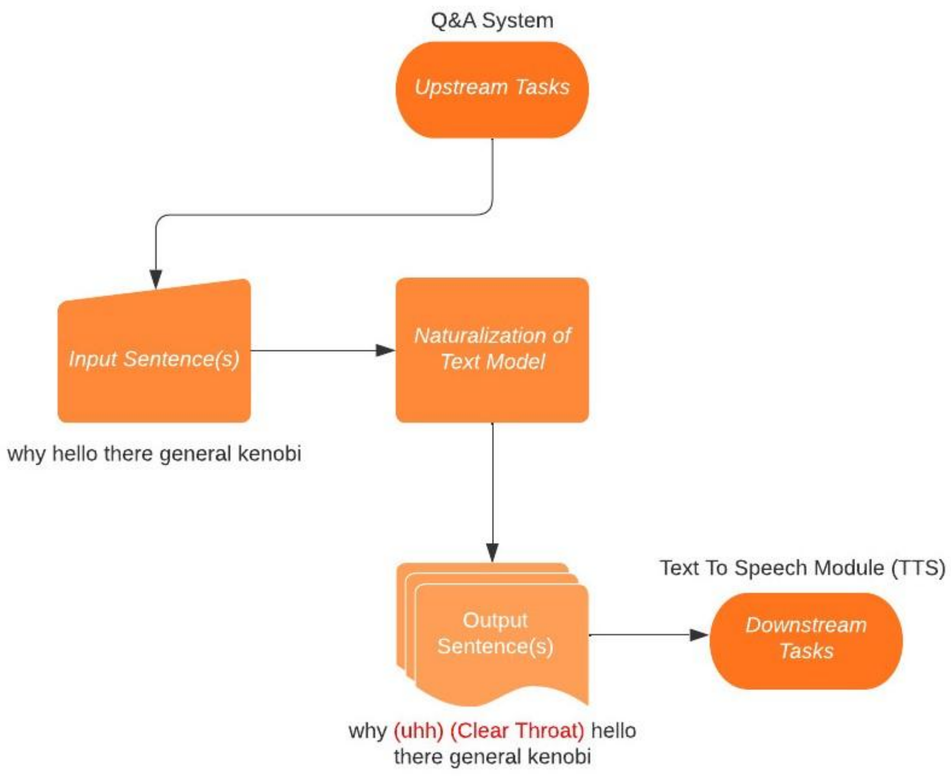
Naturalization of Text: Inserting Disfluencies

USC Viterbi School of Engineering

Alfianto Widodo, Bowman Brown, Ira Deshmukh, Parth Vipul Shah and Ryan Luu
University of Southern California, Los Angeles, CA, USA

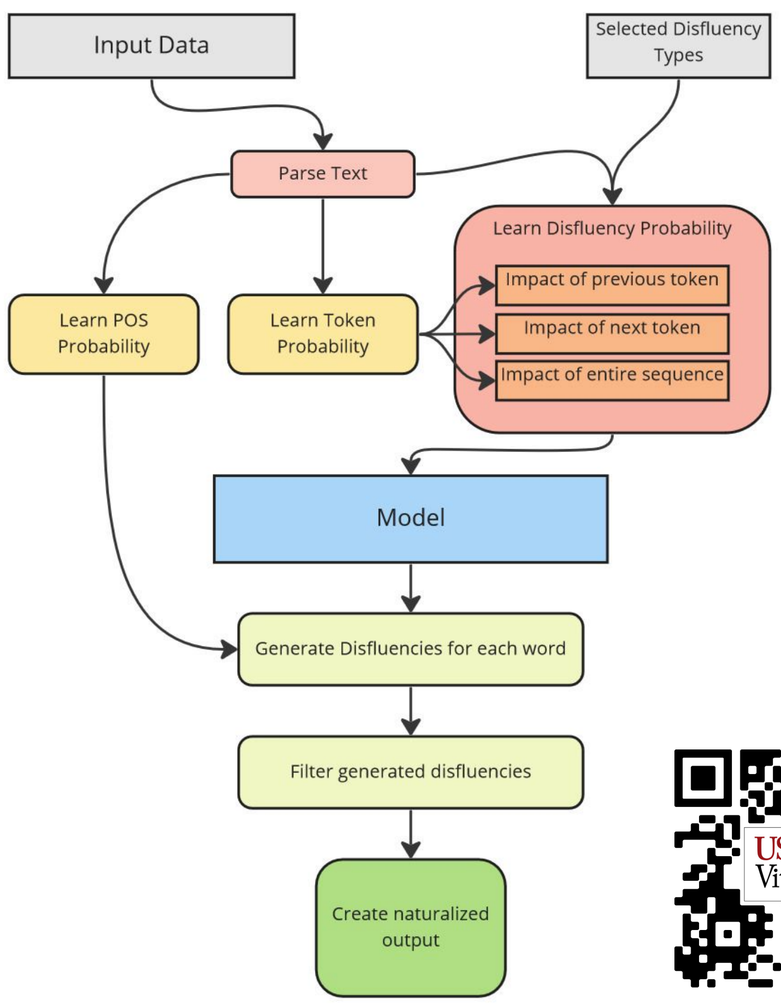
Introduction and Motivation

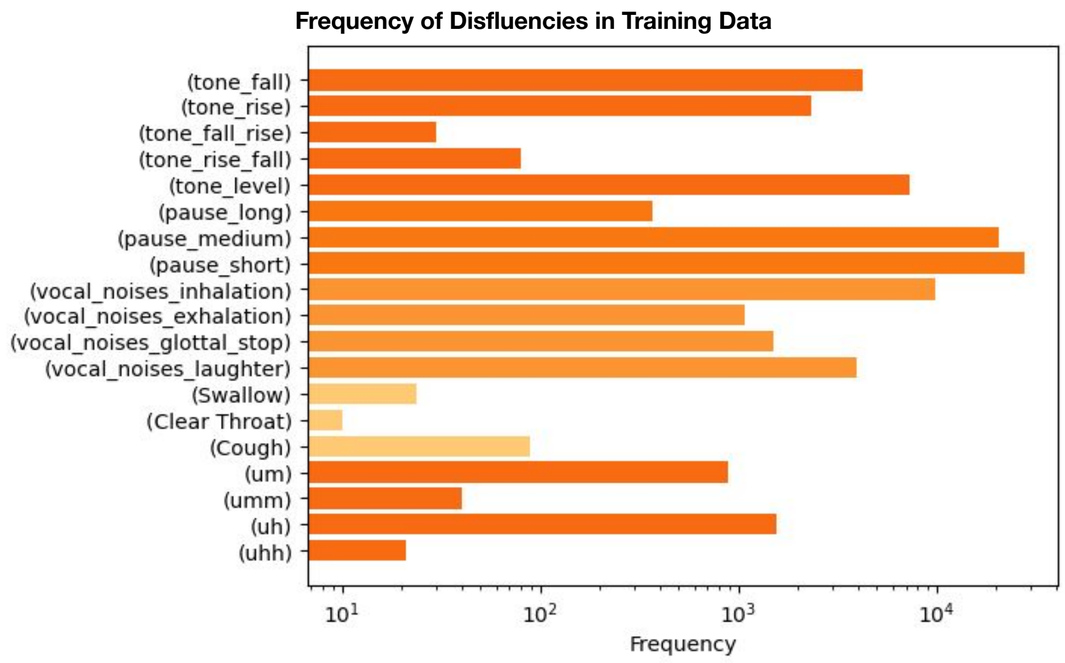
We aim to transform raw text documents of spoken dialogue / speech into its most natural-sounding version by augmenting these documents with disfluencies. Using various natural language processing techniques such as bi-grams and context analysis, we aim to create a model that can detect the most appropriate places in the document to insert these disfluencies and which form of it to insert. A successful implementation of the ideas examined in this project can aid in the efforts of naturalizing speech synthesis.



Methodology

Our method trained a model to make predictions based on the probabilities of disfluency usage and the tokens and corresponding parts-of-speech within the training corpus. We trained two types of models to insert disfluencies into text data. Our first model is a simple probabilistic model that uses bigram probabilities to predict and insert the disfluencies Our second model used a transformer based on the GPT-2 'Small' architecture that was fine-tuned on our disfluency training data. Despite using a more advanced architecture, the transformer-based model performed worse than the bigram model during disfluency selection.



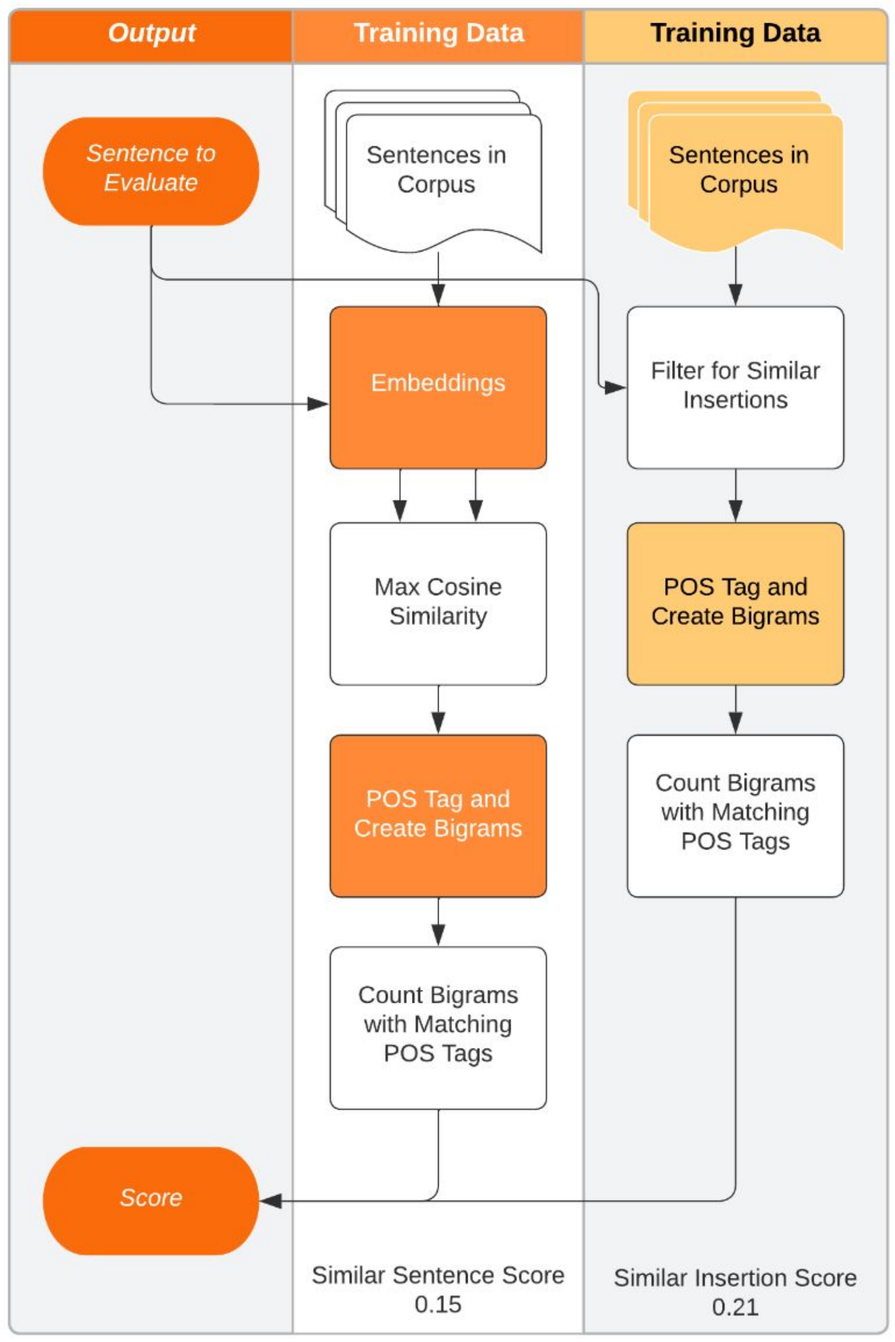


Evaluation & Results

The subject of this study is to produce text that sounds natural. Therefore, the ideal method to measure the model's performance would be to conduct a survey to receive feedback from real humans. Due to to scope limitations of the project, we implemented two automated scoring systems; similar sentence scoring and similar insertion scoring.

Test sentence: *It's just too bad*
Most similar corpus sentence: *It's too long*
Test sentence with insertions: *(pause_medium)* It's just *(vocal_noises)* too long.
Corpus sentence with insertions: *(pause_short)* It's too *(pause_short)* bad.

Test sentence	***(pause_medium, J') , (it's, NN)***	*(it's, NN), (just, RB)*	*(just, RB), (vocal_noises, VB)*	*(vocal_noises, VB), (too, RB)*	*(too, RB), (bad, JJ)*
Corpus sentence	***(pause_short, JJ), (it's, NN)***	*(it's, NN), (too, RB)*	*(too, RB), (pause_short, RB)*	*(pause_short, RB), (long, RB)*	-



Scan the QR Code to try out a working demo of our project!