**DSAN 6600 Final Project Proposal**

*Group: Hongxin Wu, Yanyan Li, Xinran Zhang*

1. **Project Summary**

Our project aims to address the limitations of current autocorrect systems by developing a neural network-based model that can detect and correct homophone errors within text. Our goal is to analyze the context surrounding homophones and employ advanced language models to discern and rectify incorrect usage.

* We have noticed that many auto-correct tools, such as Apple's in text messages, lack the ability to correct when an incorrect homophone is used. For example, all of the words in "We one the game!" are spelled correctly, but the wrong "won" was used.
* Some APPS use the wrong their/there/they're, petal/pedal, son/sun, knight/night, etc.

The context around these words should make which spelling is correct apparent, enabling context-based language models to predict the correct spelling. We envision this tool to process significant text inputs, identify homophone errors, suggest corrections, or even provide the corrected version directly. This tool, hosted on a Shiny dashboard or similar platform, will offer assistance and potentially integrate predictive text suggestions. This project is exciting due to its practical application in everyday communication, where such errors are common yet often overlooked due to cognitive load or oversight.

1. **Modeling Approach**

* We will utilize a pre-trained causal language model from Hugging Face, focusing on generative models for their prowess in predictions. These models excel in contextual understanding, which is crucial for homophone correction.
* Fine-tuning the model on a specialized dataset for homophone correction will enhance its specificity to our task.
* We plan to perform hyperparameter tuning and regularization to optimize the model's performance while preventing overfitting.
* Beam Search will be implemented during inference to explore multiple correction possibilities, ensuring both grammatical accuracy and effective homophone correction.
* Rule-based algorithms may supplement the model for handling straightforward homophone confusions.

1. **Data Selection**

* Our primary resource will be a pre-trained language model from HuggingFace, selected for its size (≤100 million parameters) and generative nature.
* We intend to generate a labeled dataset using large text corpora, such as Project Gutenberg or Reuters Corpus, by inserting homophone errors at a controlled rate. This will create a balanced dataset of error-laden and error-free texts.
* We might have to filter out non-printable characters, normalize text, and tokenize into words during preprocessing.

1. **Model Evaluation**

* Sample Text Evaluation: Testing the model with texts containing intentional homonym mistakes.
* Performance Metrics: Assessing the model using accuracy, precision, recall, speed, coverage, Word Error Rate (WER), Error Detection Rate, and Error Correction Rate.

1. **Optimization and Others**

* Web Application: Development of a user-friendly web interface for model interaction, potentially via Streamlit.
* Model Selection: Identifying an effective yet compact pre-trained language model.
* Beam Search Algorithm: Crafting a custom algorithm tailored for grammar correction.

1. **Project Challenges**

* Choosing a suitably robust yet compact pre-trained model is critical for the project's success.
* Implementing Beam Search for grammatical correction poses a significant challenge and will require innovative coding and algorithm development.

1. **Conclusion**

The successful implementation of this project has the potential to significantly enhance the accuracy and reliability of autocorrect systems. By leveraging deep learning techniques and advanced models, we aim to create a sophisticated tool that not only corrects homophonic errors but also enhances the overall quality of written communication.