**CSCE 5290: Natural Language Processing**

**Project Proposal**

**Title: Sentence to Sentence Semantic Similarity for Identifying Similar Questions on Quora**

**1. Motivation**

In real world, we know different people have different writing styles and words usage to express a feeling or to ask a question. It is difficult for NLP systems to accurately identify underlying intent of sentences regardless of the writing style and the way it phrased. It essential to build some semantic similarity checker which can help for the below usecases notably.

Question Answering Systems: Even when a user's inquiry doesn't exactly match one that already exists, a system can nevertheless find the most pertinent response by recognising similar inquiries.

Chatbots and Virtual Assistants: Chatbots can identify rephrased versions of frequently requested queries and point users towards the right answers with the use of sentence similarity.

Search Engines: Grouping the questions with similar intent can improve the search result of a search engine though different users queries with different wordings.

This project aims to develop a machine learning model that can identify question pairs with the same intent, even if the wording differs. This will improve the user experience on Q&A platforms by:

* Reducing duplicate questions
* Helping users find existing answers to their questions faster
* Leading to more efficient knowledge sharing

**2. Significance**

Semantic similarity detection has a wide range of applications in the field of Natural Language Processing. This project specifically focuses on the domain of Question & Answer platforms like Quora. A successful model will significantly improve the efficiency of these platforms by:

* Improved User Experience: Regardless of the way sentence/question phrased their questions, users receive the answers best answer they are looking for.
* Enhanced System Efficiency: Reduces the redundancy and wasted effort, so that systems can answer questions more effectively.
* Increased Scalability: As the volume of user queries grows, the ability to identify similar questions becomes even more critical for managing large datasets.
* Improved knowledge discovery: By getting to know similar questions, users can explore related topics and gain in-depth understanding.

**3. Objectives**

The primary objective of this project is to develop a machine learning model that can accurately identify question pairs on Quora with the same intent, even if the wording is different. Success will be measured by:

* **Accuracy:** Achieve an accuracy of at least 70% in identifying question pairs with the same intent on the Quora question-pair dataset. This metric emphasizes correctly identifying both similar and dissimilar pairs.
* **Focus on Semantic Similarity:** The model should prioritize identifying questions with similar meaning over exact wording matches. This can be measured by evaluating performance on datasets containing paraphrased questions.
* **Generalizability to Unseen Questions:** The model should perform well on unseen Quora questions, not just those included in the training dataset. This can be assessed by evaluating the model on a hold-out test set from Quora.

**4. Features**

This project follows sentence embedding techniques to represent questions as numerical vectors. These vectors will capture the semantic meaning of the question, allowing for comparison based on meaning rather than just keywords.

Here are some distinctive features:

* **Sentence Embedding with Attention:** Using sentence embedding techniques that incorporate attention mechanisms. Attention allows the model to focus on the most important words in each question for semantic comparison, improving accuracy for questions with slightly different phrasings.
* **Domain-Specific Preprocessing:** Perform preprocessing specific to Quora questions. This might include techniques like identifying and handling named entities frequently mentioned on Quora.
* **Incorporate Question Context:** Explore methods to perform additional context from the question thread to improve semantic understanding. This could involve techniques like named entity linking or topic modeling to capture the similarity to question.

**5. Dataset.**

For this project we are going to use a publicly available dataset of question pairs from Quora. This dataset will contain questions labelled as duplicate or not. Here are some details about the dataset:

* Source: <https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs>
* Type: Labeled text data
* Size: 404290 rows, 6 columns
* Preprocessing: The data will be preprocessed by cleaning (removing noise) and tokenization (splitting text into words).

**6. Visualization**

Workflow Diagram: Below is our flowchart which explains the steps involved in the project, including data collection, preprocessing, model training, and evaluation.

A screenshot of a computer

Description automatically generated