NLP Assignment - Spell Correction for ASR Noun Enhancement

Background:

Automatic Speech Recognition (ASR) systems often struggle with accurately transcribing proper nouns, technical terms, and domain-specific vocabulary, leading to misspelled or incorrectly transcribed words in the output text. This is particularly problematic in medical conversations where accurate terminology, especially medication names and medical terms, is crucial for understanding and analysis.

Spell correction is a fundamental task in Natural Language Processing (NLP) that involves identifying and correcting misspelled words in text. The goal of this assignment is to create a specialized spell correction system that focuses specifically on correcting noun-related errors commonly produced by ASR systems, particularly medication names and medical terminology in general medical conversations.

Note: For this assignment, you will work with a pre-prepared dataset containing pairs of correct sentences and their corresponding ASR-generated incorrect transcriptions from general medical conversations. Your task is to build a spell correction system that can transform the ASR output back to the original correct sentences, with particular focus on correcting medication names and other medical nouns.

Assignment Tasks:

1. Dataset Understanding and Analysis

- **Dataset Structure Analysis:**
  - Examine the provided dataset containing pairs of correct sentences and ASR-generated incorrect sentences
  - Identify the input features (ASR transcriptions) and target labels (ground truth sentences)
  - Analyze sentence length distributions, vocabulary size, and medical terminology coverage
- **Error Word Extraction and Analysis:**
  - **Identify Error Words:** Compare correct and incorrect sentence pairs to extract specific words that were transcribed incorrectly
  - **Categorize Error Words:** Focus on medication names and other medical nouns that are commonly mistranscribed by ASR systems

    - **Create Error Word Database:** Build a comprehensive list of error patterns specifically for:
        - Medication names (e.g., "ibuprofen" → "eye-beu-pro-fen")
        - Medical terminology
        - General nouns in medical context
- **Error Type Classification:** Categorize errors into basic types such as:
    - Phonetic substitutions (sound-alike errors)
    - Character-level errors (insertions, deletions, substitutions)
    - Word-level errors (completely wrong words)
    - Segmentation errors (word boundary issues)

2. Data Preprocessing

- Text cleaning: Handle special characters while preserving context
- NER: Use Named Entity Recognition to extract the noun words.
- POS tagging: Use Part-of-Speech tagging to specifically target nouns
- Data splitting: Create training, validation, and test sets (70-15-15 split)

3. Model Development

- Baseline Model: Implement traditional spell correction approaches:
    - Edit distance algorithms (Levenshtein distance)
    - N-gram language models
    - Dictionary-based correction
- Advanced Model: Fine-tune transformer-based models:
    - BERT-based approaches for contextual correction
    - T5 or similar seq2seq models for text correction
    - Custom architectures focusing on noun correction

3. Data Preprocessing

- **Text cleaning:** Handle special characters while preserving context
- **Tokenization:** Implement noun-aware tokenization to identify target words
- **POS tagging:** Use Part-of-Speech tagging to specifically target nouns
- **Handling missing values and duplicates**

- **Data splitting:** Create training, validation, and test sets (70-15-15 split)

4. Model Development

- **Baseline Model:** Implement traditional spell correction approaches:
  - Edit distance algorithms (Levenshtein distance)
  - N-gram language models
  - Dictionary-based correction
- **Advanced Model:** Fine-tune transformer-based models:
  - BERT-based approaches for contextual correction
  - T5 or similar seq2seq models for text correction
  - Custom architectures focusing on noun correction

5. Evaluation

- Metrics: Use appropriate evaluation metrics:
  - Word-level accuracy
  - Character-level accuracy
  - BLEU score for sequence comparison
  - Noun-specific accuracy rates
- Error Analysis: Analyze remaining errors and categorize failure modes

6. Exploratory Data Analysis (EDA)

- **Training Data Analysis:**
  - Frequency distribution of corrupted noun types
  - Error pattern analysis
  - Vocabulary coverage statistics
- **Results Analysis:**
  - Performance across different noun categories
  - Impact of context length on correction accuracy
  - Comparison of correction confidence scores

Deliverables:

Technical Implementation:

- Working code with clear documentation
- Jupyter notebooks with step-by-step implementation
- Modular code structure with reusable components
- Requirements.txt with all dependencies

Analysis and Documentation:

- **Approach Documentation:** Detailed explanation of methodologies used
- **Challenge Analysis:** Document specific challenges faced and solutions implemented
- **Results Presentation:** Clear visualization of findings and performance metrics

Dataset:

Excel file: Spell_Correction_for_ASR_Noun_Enhancement_assignment_dataset.xlsx

Pre-prepared ASR Spell Correction Dataset

- Contains pairs of correct sentences and ASR-generated incorrect transcriptions
- Sourced from general medical conversation domains
- Focus on analyzing and correcting the provided ASR errors, particularly medication names and medical terminology

Platform:

Google Colab (Google Colab)

- Utilize GPU acceleration for transformer model training
- Implement efficient memory management for large datasets

Time Duration:

3 days