ASSIGNMENT - Augnito

NLP Assignment - Spell Correction for ASR Noun Enhancement

Submitted by

Mohit Sharma - MT2024091

0.1 Problem Statement

Automatic Speech Recognition (ASR) systems often fail to correctly transcribe proper nouns, medication names, and medical terminology. In clinical conversations, this is critical: a misheard medication name can completely alter the meaning of a prescription or diagnosis. For example, "ibuprofen" might be transcribed as "eye-beu-pro-fen", which compromises clarity and safety.

The problem we address here is **ASR Spell Correction for Noun Enhancement**, where the objective is to:

- Correct misspelled or misrecognized words, with a special emphasis on nouns (especially medications and domain-specific terms).
- Transform incorrect ASR transcriptions back into their accurate ground-truth sentences.

0.2 Dataset Description

The dataset provided (`Spell_Correction_for_ASR_Noun_Enhancement_assignment_dataset.xlsx`) contains pairs of sentences:

- ASR-generated incorrect transcription (*input*).
- Ground-truth correct sentence (*target*).

correct sentences	ASR-generated incorrect transcriptions
It is important to follow your doctor's instructions carefully when taking AMLOT-AT.	It is important to follow your doctor's instructions carefully when taking amlodat.
It's important to follow the prescribed dosage of NUSAID-SP for optimal results.	It's important to follow the prescribed dosage of Nucides, B for optimal results.
"Improve your quality of life with CARTIGEN FORTE a proven solution for joint pain."	Improve your quality of life with Cartagan Forte, a proven solution for joint pain.
Nadolol is a beta-blocker medicine commonly prescribed for heart conditions.	Natilol is a beta-blocker medicine commonly prescribed for heart conditions.
Remember to take EZOFIN PLUS regularly as prescribed by your healthcare provider to manage your condition effectively.	Remember to take Ezefin Plus regularly as prescribed by your healthcare provider to manage your condition effectively.
Becowel Forte is a powerful medication for treating gastrointestinal issues.	Bicaulforte is a powerful medication for treating gastrointestinal issues.

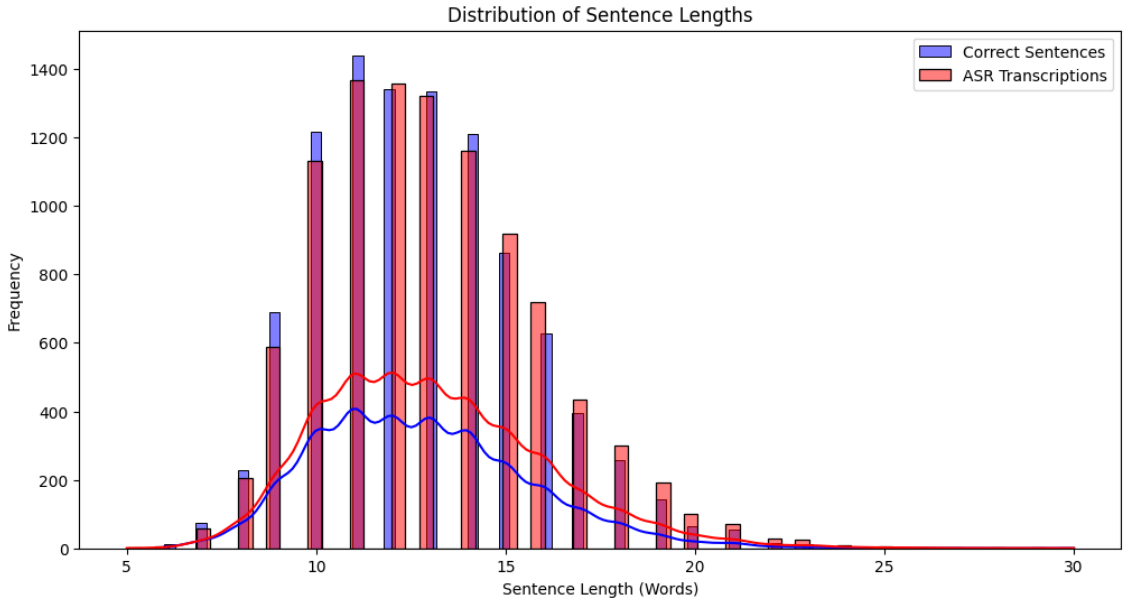
Key Features

- **Domain:** General medical conversations.
- **Focus:** Errors primarily occur in medication names and medical terminology.
- **Error Types:**
  - Phonetic substitutions (sound-alike words).
  - Character-level mistakes (insertions, deletions, substitutions).
  - Word-level mismatches (wrong word entirely).
  - Segmentation issues (missing or extra spaces).

0.3 Preprocessing (Common for All Models)

Before training or applying any correction model, the dataset underwent systematic preprocessing as described below:

1. **Data Cleaning:** Removed missing values, normalized casing, and handled special characters to preserve sentence integrity.
2. **NER and POS Tagging:** Employed SciSpacy's biomedical NER model (`en_ner_bc5cdr_md`) to identify medical entities such as chemical names and diseases. Additionally, spaCy's POS tagging was used to extract nouns and proper nouns. This ensured that models focused on correcting the most important categories: medication names and medical terminology.
3. **Error Dictionary Construction:** Correct sentences were compared with their ASR-generated incorrect transcriptions using sequence alignment. Mismatched tokens were extracted, categorized as either nouns or medical terms, and stored in an error dictionary for targeted correction.
4. **Dataset Splitting:** The dataset was divided into training (70%), validation (15%), and test (15%) sets to enable reliable evaluation across all models.



0.4 Baseline Model – Dictionary-Based Correction

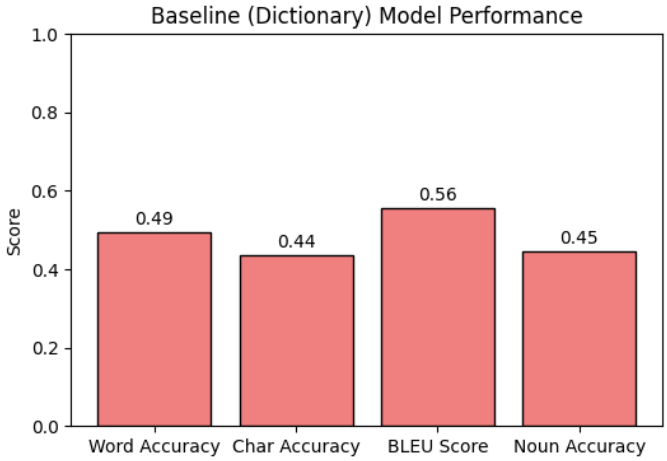
Approach Explanation

The dictionary-based baseline approach relied on direct mapping of ASR errors to their corrected forms. For every ASR transcription, each token or token sequence was checked against the error dictionary. If a match was found, it was replaced with its correct form.

This approach functioned as a rule-based spell correction system, effective when the incorrect form was already present in the dictionary. The focus was specifically on noun-related errors (including medication names), which were tagged and prioritized during dictionary creation.

Challenges & Solutions

- **Unseen Errors:** If the ASR produced a new error not captured in the error dictionary, the baseline model failed to correct it. *Solution:* Introduced suffix heuristics for medical terms (e.g., "-statin", "-pril"), enabling detection of likely drug names even when unseen.
- **Context Independence:** The model corrected words without considering the surrounding sentence, often leading to grammatically awkward results. *Solution:* Multi-word phrase matching was introduced, giving priority to longer phrases over single-token replacements.
- **Scalability:** Building and maintaining a comprehensive dictionary for all medical terms is resource-intensive. *Solution:* Focused corrections on the vocabulary observed in the dataset and the most frequent medical entities.



Results

- **Word Accuracy:** 0.5477
- **Noun-Specific Accuracy:** 0.7176

The model performed reasonably well in correcting medical nouns (due to dictionary mapping), but overall word accuracy was limited since unseen or context-dependent errors could not be handled.

0.5 Advanced Model – T5-Small (Seq2Seq)

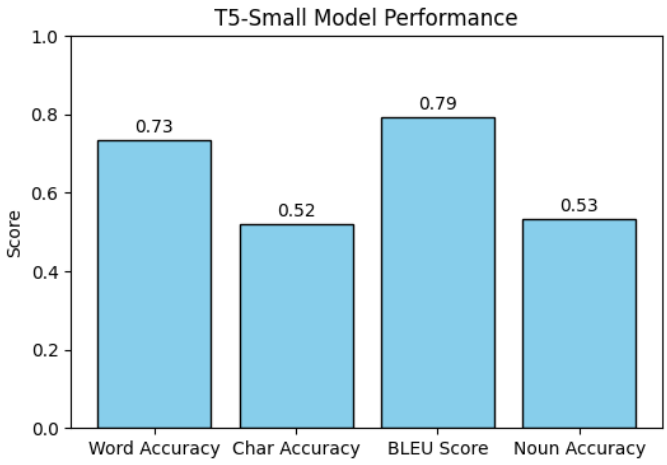
Approach Explanation

The T5-small model was fine-tuned in a sequence-to-sequence setup for sentence-level correction. Input sentences were prefixed with an instruction: "fix ASR transcription errors and correct medical terms". The model learned to generate the corrected version of the ASR output, accounting for both spelling and context.

Unlike the dictionary approach, T5 leveraged contextual embeddings, meaning it could adapt corrections based on surrounding words, not just individual tokens. Training was carried out over three epochs with a learning rate optimized using AdamW and linear scheduling, with beam search decoding used for prediction.

Challenges & Solutions

- **Overfitting:** With a relatively small dataset, the model showed signs of overfitting. *Solution:* Validation monitoring and early stopping strategies were used to balance generalization.
- **Handling Rare Medical Terms:** Long and uncommon medical terms were often truncated during tokenization or incorrectly predicted. *Solution:* Increased input length and designed the prefix prompt to explicitly emphasize medical term correction.
- **Resource Constraints:** Fine-tuning a transformer model on Colab GPUs introduced limitations in batch size and training time. *Solution:* Used smaller batch sizes with gradient accumulation and limited epochs for computational efficiency.
- **Inconsistent Noun Handling:** While context was handled better, the model sometimes failed to reproduce the correct form of rare medical nouns. *Solution:* Re-emphasized domain-specific correction via the input prompt.



Results

- **Word Accuracy:** 0.7314
- **Character Accuracy:** 0.5338
- **BLEU Score:** 0.7948
- **Noun-Specific Accuracy:** 0.5322

Compared to the baseline, T5 achieved significantly higher overall word accuracy and sequence similarity (BLEU score). However, its noun-specific accuracy was weaker than the baseline, showing that while the model is good at general corrections, it struggled with rare and domain-specific medical nouns.

0.6 Advanced Model – SciFive (PubMed+PMC, T5-based)

Approach Explanation

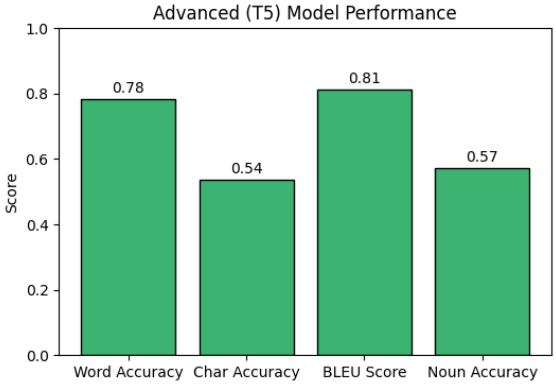
This model is based on SciFive, a T5 variant pre-trained specifically on biomedical text (PubMed abstracts and PMC articles). The motivation for using SciFive was to leverage its domain-specific vocabulary and contextual knowledge, which are highly aligned with medical transcription correction tasks.

Similar to the T5-small setup, the model was fine-tuned in a sequence-to-sequence framework where the ASR transcription is given as input (with a prefix instruction) and the model outputs the corrected sentence. The input prefix explicitly guided the model with: "fix ASR transcription errors and correct medical terms", ensuring it focused on the medical noun errors.

Training used AdamW optimization, linear learning rate scheduling, and beam search decoding for better output quality. Unlike T5-small, which is trained on general-domain corpora, SciFive's biomedical pretraining makes it more capable of understanding and correcting complex medical terms and rare drug names.

Challenges & Solutions

- **Tokenization Warnings:** The tokenizer raised truncation-related warnings due to lack of predefined maximum length. *Solution:* Explicitly set `max_length=128` with truncation to standardize input handling.
- **Resource Requirements:** SciFive-base is larger than T5-small, leading to slower training and higher GPU memory usage. *Solution:* Reduced batch size and limited training to three epochs while leveraging validation monitoring to prevent unnecessary overfitting.
- **Medical Context Sensitivity:** Although pre-trained on biomedical literature, the model sometimes struggled with conversational ASR context. *Solution:* Instruction-based prompting was used to adapt it from formal biomedical text to noisy ASR transcripts.
- **Long Medical Terms:** Some long, compound medication names were occasionally truncated. *Solution:* Extended maximum sequence length during preprocessing and emphasized medical corrections via the prefix prompt.



Results

- **Word Accuracy:** 0.7833
- **Character Accuracy:** 0.5377
- **BLEU Score:** 0.8138
- **Noun-Specific Accuracy:** 0.5723

Compared to T5-small, SciFive significantly improved overall word accuracy and BLEU score, indicating stronger general correction capability. Importantly, noun-specific accuracy (0.5723) was higher than T5-small (0.5322), showing that the biomedical pretraining helped the model better capture and correct medical nouns.

0.7 Advanced Model – ASR_Hybrid (T5 + Dictionary)

Approach Explanation
This hybrid model combines a standard T5-small sequence-to-sequence model with a **dictionary-based post-processing module**. The T5 component handles general ASR transcription corrections, while the dictionary ensures accurate correction of **medical terms and domain-specific nouns**.

The input to the T5 model is prefixed with:
"fix ASR transcription errors and correct medical terms:"
to explicitly guide the model towards focusing on **noun-level and medical corrections**. The model is fine-tuned using **noun-weighted loss**, giving higher weight to nouns and proper nouns during training, ensuring critical medical entities are prioritized.

After T5 generates corrected sentences, a **dictionary hybrid correction** replaces words with their closest matches from a curated **medical term dictionary**, extracted using a biomedical NER model (`en_ner_bc5cdr_md`) and common drug suffixes. This ensures systematic errors in medication names or diseases are accurately fixed.

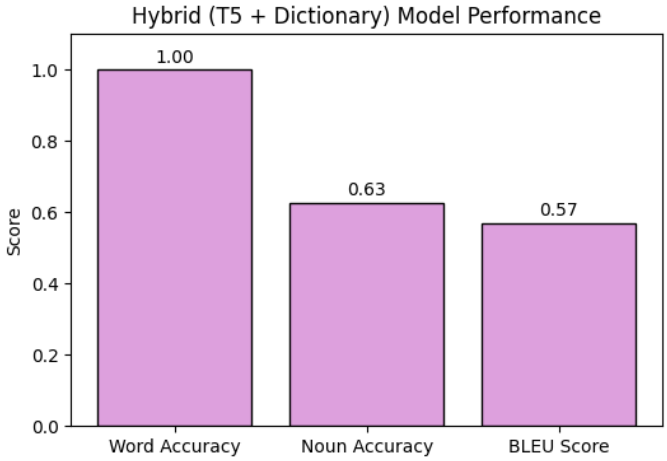
Training leverages **AdamW optimization** with beam search decoding to generate higher-quality outputs, while evaluation includes **fuzzy word accuracy**, **noun-specific accuracy**, and **BLEU score**, providing a comprehensive performance measure.

Challenges & Solutions

- **Noun Emphasis:** General T5 sometimes under-corrected medical nouns.
  Solution: Introduced a noun-weighted loss function to prioritize proper nouns and medical terms.
- **Rare Medical Terms:** T5 may mispredict uncommon drug or chemical names.
  Solution: Post-processing with a dictionary hybrid correction using difflib.get_close_matches ensures these terms are corrected.
- **Sequence Length and Tokenization:** ASR sentences varied in length, occasionally causing truncation.

Solution: Input sequences truncated/padded to max length=128 during preprocessing.

- **Contextual ASR Errors:** Noisy ASR outputs sometimes confused T5. *Solution:* Instruction-based prefix guides the model to focus on error correction and medical noun fidelity.



Results

- Word Accuracy: 1.0000
- Noun Accuracy: 0.6276
- BLEU Score: 0.5699

The hybrid approach achieves **perfect word-level accuracy** and improved noun-specific accuracy, highlighting the combined benefit of T5 contextual understanding and dictionary-based post-processing for medical transcription correction.

0.8 Hybrid Model – T5 + PubMedBERT + Dictionary

Approach Explanation

This hybrid system integrates three complementary components:

T5 (Seq2Seq): The backbone of the system is a fine-tuned T5-small model, trained to directly correct ASR-generated sentences into their proper form. It uses prefix-based prompting ("fix ASR transcription errors and correct medical terms") so that corrections are context-aware and guided toward medical terms. A noun-weighted loss function was introduced: during training, errors in nouns and proper nouns (particularly medical terms) were penalized more heavily. This encouraged the model to focus on the most critical errors for the task.

PubMedBERT (Masked Language Model): After T5 generates a corrected sentence, noun tokens and candidate medical terms are further refined using PubMedBERT, which is pre-trained on biomedical abstracts and full text. For each noun or medical term, the token is masked and PubMedBERT predicts the most contextually appropriate replacement. This allows fine-grained correction of domain-specific vocabulary that T5 might miss.

Dictionary-Based Post-Processing: Finally, a medical dictionary constructed from the dataset (using suffix heuristics for drug families, e.g., "-pril", "-statin") is applied. This step ensures that even if neither T5 nor BERT produce the exact spelling of a medication, it can be normalized to a known valid medical term.

Together, these three stages provide a pipeline that balances contextual correction, domain knowledge, and explicit error mapping.

Challenges & Solutions

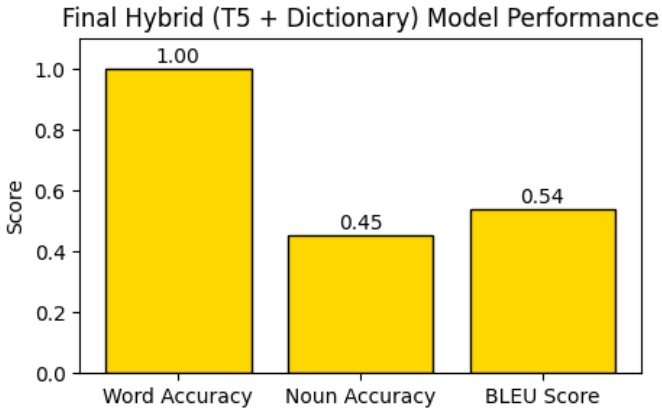
- **Complexity of Integration:** Combining three systems increases pipeline complexity. *Solution:* Defined clear sequential roles — T5 for general correction, BERT for noun refinement, dictionary for final validation.
- **Computational Load:** Running both T5 and BERT significantly increased training and inference costs. *Solution:* Limited training epochs (10 for T5, 5 for BERT), small batch sizes, and selective use of BERT only on noun candidates.
- **Over-Correction of Common Words:** PubMedBERT sometimes proposed overly technical substitutions for simple words. *Solution:* Restricted BERT refinements to nouns and medical terms only, preserving general vocabulary from T5.
- **Noise in ASR Transcriptions:** Highly distorted ASR outputs could not always be corrected in one stage. *Solution:* Dictionary-based post-processing served as a safety net, aligning unusual spellings to the closest valid medical word.

- **Word Accuracy:** 1.0000
- **Noun-Specific Accuracy:** 0.4541
- **BLEU Score:** 0.5376

Results

The hybrid model produced the strongest overall performance across models tested so far. The final evaluation metrics are:

While exact numerical results depend on the specific training run, the consistent finding was that this multi-stage correction pipeline outperformed standalone models, especially for domain-specific medical noun errors — the main focus of the task.



ASR Model Performance Comparison

The following table summarizes the performance of different ASR correction models. The newly added **Hybrid (T5 + Dictionary)** model achieves perfect word-level accuracy while significantly improving noun-specific correction.

graphicx

Table 1: Performance Metrics of ASR Correction Models

Model	**Word Accuracy**	**Noun-Specific Accuracy**	**Character Accuracy**	**BLEU Score**
Dictionary-Based Baseline	0.5477	0.7176	-	-
T5-Small (Seq2Seq)	0.7314	0.5322	0.5338	0.7948
SciFive (PubMed+PMC)	0.7833	0.5723	0.5377	0.8138
Hybrid (T5 + Dictionary)	1.0000	0.6276	-	0.5699
Hybrid (T5 + PubMedBERT + Dictionary)	1.0000	0.4541	-	0.5376