SPELL-CHECKER FOR CANCER RESEARCH

MOHAMMAD ADNAN

SYSTEM OVERVIEW

Model: The back-end uses DistilBERT, a distilled version of BERT, fine-tuned for Masked Language Modeling (MLM) on a cancer-specific corpus.

Backend: The spell checker logic, including model predictions and text processing, is handled in Python.

Libraries:

- PyTorch and Hugging Face's Transformers: For model implementation.
- NLTK: For linguistic processing like tokenization and bigram modeling.
- Tkinter: Used for creating the simple, interactive GUI.

Frontend: The user interface allows real-time interaction, displaying suggestions for errors detected in the input.

99.8%

MODEL ACCURACY

The model achieved an impressive accuracy indicating strong performance in predicting masked tokens within the domain-specific corpus. This high accuracy suggests that the fine-tuned DistilBERT model effectively captures context and domain knowledge. However, further evaluation on a broader set of tasks is needed to ensure it generalizes well beyond this dataset.

SPELL CHECKING WEAKNESS

KEY ISSUES IN CURRENT SYSTEM

01.

FALSE POSITIVES

Many domain-specific terms (like complex medical terms) are flagged as misspelled even though they are correct.
For eg, new cancer terms might not be recognized by the model.

02.

OVER-RELIANCE ON EDIT DISTANCE

Edit distance focuses only on word similarity (i.e., the number of edits between words), which may overlook context or frequency of domain-specific terms.

03.

LACK OF CONTEXTUAL AWARENESS

The system checks each word in isolation, leading to inappropriate suggestions.
For eg, it might suggest common words rather than the medical context of surrounding words.

04.

NO GRAMMAR OR SYNTAX CHECKS

The system does not understand the grammatical or syntactical role of the word in a sentence, which leads to inappropriate word replacements that grammatically do not fit.

CANDIDATE WORD GENERATION

USING HIGHER-ORDER N-GRAMS

CURRENT LIMITATION:

The system uses word pairs, limiting understanding of longer medical terms.

Bigrams may fail when complex medical terms require understanding beyond two-word relationships, such as in cancer research ("tumor necrosis factor").

Example:
Input phrase: "tumor necrosis"
Bigram-based correction: Might suggest "tumor neck" due to bi-word proximity.

PROPOSED IMPROVEMENT:

TRIGRAMS AND QUADGRAMS

Expand the model to trigrams and quadgrams to capture longer dependencies between words.

Benefit: This improves handling of complex phrases and word connections, vital for fields like cancer research.

BACKOFF MODELS

Implement backoff models where the system defaults to bigrams if a trigram or quadgram isn’t available.

Benefit: This maintains relevant suggestions when longer word patterns aren't available.

CONTEXT-AWARE CANDIDATE RANKING

USING SEMANTIC AND POS TAGGING

CURRENT LIMITATION:

Ranking relies on a combination of edit distance and bigrams, which leads to incorrect suggestions when:

- The model doesn't capture the semantic meaning of words.
- The system doesn’t differentiate between parts of speech (e.g., suggesting a noun in place of a verb).

Example:
Sentence: “The treatment was affective.”
With POS tagging and semantic analysis: System would suggest "effective".

PROPOSED IMPROVEMENT:

PART-OF-SPEECH (POS) TAGGING INTEGRATION

Use POS tagging for better sentence understanding.

Match suggestions to correct word types (nouns or verbs)

Benefit: More grammatically consistent suggestions that align with the sentence structure.

SEMANTIC ANALYSIS USING WORD EMBEDDINGS

Use medical word embeddings for term similarity matching.(e.g., BioWordVec)

Benefit: Instead of focusing purely on spelling similarity, the system suggests words that are semantically relevant, leading to better, domain-specific corrections.

INFORMATION RETRIEVAL TECHNIQUES

BETTER CANDIDATE RANKING

CURRENT LIMITATION:

System only checks single words, missing multi-word medical terms

Can't handle compound terms, may wrongly flag parts of valid phrases

Example:
Current system:

A search for "heart" only returns results for the word "heart," missing phrases like "heart disease" or "heart failure."

PROPOSED IMPROVEMENT:

TF-IDF INTEGRATION

TF-IDF ranks words by importance in specialized texts.
Benefit:

- TF-IDF prioritizes domain-specific terms, reducing false positives in candidate generation.
  (eg. Neoadjuvant Chemotherapy)
- It enhances search by better recognizing multi-word terms, especially in technical or medical texts.

MULTI-WORD PHRASE RECOGNITION

Upgrade the search functionality to handle multi-word phrases by implementing a phrase-based IR system.

Benefit: Allows the system to search for the whole phrases instead of individual words.

CONCLUSION

FINAL THOUGHTS

The current spell-checking system effectively combines edit distance, bigrams, and a fine-tuned BERT model but faces limitations in handling domain-specific terms, multi-word phrases, and contextual accuracy. By enhancing the system with improved candidate generation, multi-word search capabilities, and advanced IR techniques, we can significantly boost its precision and relevance for cancer research.


THANK YOU
VERY MUCH!