FAQ Retrieval System Using Weighted Edit Distance For Noisy Queries

Dhruvi Mehta, Prachi Mehta, Riya Dave and Satyam Maravaniya

Abstract

Develop a system to match user queries with FAQs effectively, even in the presence of errors like typos or variations in phrasing. The system employs weighted edit distance for improved accuracy in handling noisy queries and includes a mechanism to detect out-of-domain queries using a classifier. This approach demonstrates robust performance in accurately identifying both relevant and out-of-domain queries.

Keywords

User Queries, FAQs, Weighted Edit Distance, Out-of-Domain, Classifier

1. Problem Statement

The goal is to develop and implement a system that accurately matches user queries to Frequently Asked Questions (FAQs), even if the queries contain errors such as typos, misspellings, or other phrasing. The system accomplishes this by using weighted edit distance, which allocates varying costs to string operations (such as insertions, deletions, and substitutions) in order to better handle ”noisy” searches.

Furthermore, the system includes a classifier to recognize out-of-domain inquiries, which are ones that do not fit within the accessible FAQs. By recognizing such inquiries, the system guarantees that it only gives answers that are relevant to its scope, hence boosting dependability and satisfaction.

This dual approach enables the system to provide robust performance by effectively recognizing relevant matches for in-domain inquiries while separating those that cannot be answered.

2. Dataset Description

The system uses the **Yahoo! Answers Dataset**, a large and diverse dataset ideal for FAQ retrieval tasks.

- **Size:**
  - **Training Data:** 1.4 million samples
  - **Testing Data:** 60,000 samples
- **Categories:** The dataset is divided into 10 distinct categories, ensuring a variety of question types and domains for analysis.
- **Content:** Each sample in the dataset contains:
  - **Question Titles:** Short phrases summarizing the query.
  - **Question Content:** Detailed descriptions of the query.
  - **Best Answers:** High-quality responses associated with each question.

3. Evaluation Metrics

To measure the system’s effectiveness, the following metrics are used:

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India

202318003@daiict.ac.in (D. Mehta); 202318008@daiict.ac.in (P. Mehta); 202318011@daiict.ac.in (R. Dave); 202318026@daiict.ac.in (S. Maravaniya)

202318003 (D. Mehta); 202318008 (P. Mehta); 202318011 (R. Dave); 202318026 (S. Maravaniya)

1. **Accuracy:**
   - **Definition:** The proportion of queries correctly matched to relevant FAQs or classified as out-of-domain.
   - **Importance:** Reflects the overall performance of the system in correctly handling both in-domain and out-of-domain queries.
2. **Mean Reciprocal Rank (MRR):**
   - **Definition:** A metric that evaluates the ranking of results, focusing on the position of the first relevant FAQ.
   - **Formula:**

     $$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{Rank}_i}$$

     Where Rank_i is the position of the first relevant FAQ for query i, and $|Q|$ is the total number of queries.
   - **Importance:** Ensures that the most relevant FAQs are ranked higher, improving user satisfaction.

4. Results

The system achieved **strong performance**, excelling in both key tasks:

- **In-Domain Query Matching:** Successfully matched the majority of in-domain queries to their correct FAQs.
- **Out-of-Domain Detection:** Accurately identified queries that fall outside the scope of available FAQs.

The ranking mechanism consistently returned the most relevant FAQs for user queries, with the top results aligning well with user expectations. This demonstrates the system's ability to handle both noisy inputs and domain-specific classifications effectively.

5. Key Challenges

This section highlights the primary challenges encountered and the strategies adopted to address them.

5.1. Handling Noisy Queries

Queries often contain typographical errors, unusual phrasing, or missing information. The weighted edit distance algorithm effectively addresses these variations by assigning appropriate costs to string operations. Key strategies include:

- Allocating lower costs for substitutions involving similar characters (e.g., vowels).
- Increasing costs for deletions that significantly alter the meaning of a query.

5.2. Weighted Edit Distance Calibration

The weighted edit distance formula, used for handling noisy queries, requires careful calibration of operation weights. It is designed to minimize the impact of minor errors while still distinguishing between major mistakes that change the query's meaning.

$$\text{Weighted Edit Distance}(A, B) = \sum_{i=1}^{n} w(o_i) \cdot \text{cost}(o_i)$$

Where: - $w(o_i)$ is the weight of the i-th operation. - $\text{cost}(o_i)$ is the cost of the i-th operation (insertion, deletion, or substitution). - n is the total number of operations performed to convert string A to string B.

5.3. Out-of-Domain Query Detection

The system uses **Inverse Document Frequency (IDF)** to detect out-of-domain queries. This metric helps prioritize relevant terms in the dataset and classify queries based on their relevance to available FAQs.

The formula for **IDF** is:

$$\text{IDF}(t) = \log\left(\frac{N}{df(t)}\right)$$

Where: - N is the total number of documents in the corpus. - $df(t)$ is the number of documents containing the term t.

For **IDF-based scoring**, the formula is:

$$\text{Score}(q, D) = \sum_{t \in q} \text{IDF}(t) \cdot \text{TF}(t, D)$$

Where: - q is the query. - D is the document being scored. - $\text{TF}(t, D)$ is the term frequency of term t in document D.

5.4. Efficient Ranking

To rank queries efficiently, the system uses similarity measures such as the **Longest Common Subsequence (LCS) Ratio**, **Weighted Edit Distance**, and **IDF-based scoring**. These metrics help in determining which FAQ most closely matches a given query.

The LCS Ratio formula is:

$$\text{LCS Ratio} = \frac{|LCS(A, B)|}{\max(|A|, |B|)}$$

Where: - $LCS(A, B)$ is the longest common subsequence of strings A and B. - $|A|$ and $|B|$ are the lengths of strings A and B, respectively.

6. Learnings

- **Effectiveness of Weighted Edit Distance:** Weighted edit distance proved effective in handling noisy queries, with varying operation costs improving resilience to typos and phrasing errors.
- **Importance of Text Normalization:** Preprocessing techniques like lowercasing, stemming, and stopword removal helped simplify text, boosting system performance.
- **IDF as a Relevance Metric:** IDF helped prioritize rare, meaningful terms, improving query matching and out-of-domain detection.
- **Classifier Simplicity and Effectiveness:** A simple classifier using IDF scores and a threshold effectively detected out-of-domain queries, highlighting the importance of well-defined thresholds.
- **Balancing Accuracy and Efficiency:** By combining similarity metrics and precomputed weights, the system maintained high accuracy without excessive computational costs.

References

[1] *Weighted Edit Distance based FAQ Retrieval using Noisy Queries.*