QALinkPlus: Text Enrichment with Q&A data

Yandong Sun Yixuan Tang * Anthony K.H. Tung *
School of Computing, National University of Singapore, Singapore
{yandong,yixuan,atung}@comp.nus.edu.sg

Abstract

Text enrichment, the task of augmenting textual content by incorporating supplementary information to bridge knowledge gaps and enhance reader engagement, is a critical aspect of information retrieval. This study focuses on leveraging question answering datasets, such as Natural Questions and SQuAD, which contain human-validated content from diverse domains as valuable knowledge sources. While QA datasets hold promise for addressing informational needs, existing approaches, like employing dense retrieval for text enrichment, often result in QA pairs that may lack relevance, diversity, or inherent interest. To address these challenges, our paper proposes a novel graph-based method for text enrichment using QA pairs. We construct an entity co-occurrence graph derived from QA datasets and derive context-QA-specific subgraphs. Through rule-based path analysis, we develop an interpretable scoring system to assess the relevance and engagement value of each QA pair. By intelligently re-ranking QA pairs with our scoring system, our method delivers enriched text that fills knowledge gaps and captivates readers, thus improving the overall reading experience. This framework is not only effective in text enrichment tasks, but it also offers advantages for personalization and personal data management.

1 Introduction

As readers navigate vast expanses of textual content, they often come across areas where gaps in their knowledge surface or where they develop a curiosity about related topics. Question and Answer (QA) datasets, with their reservoir of knowledge, have the potential not only to bridge these knowledge gaps but also to enrich the text with related information that readers may find intriguing. Let's consider the novel "Harry Potter and the Philosopher's Stone" for an example. As shown in Fig18 (left), the novel contains various entities. An ideal set of QA pairs for this novel should explore these entities, ensuring that the questions and answers remain intimately connected to the novel's context. What's more, an ideal QA pair should delve deeper than surface-level details. A superficial question such as "Who is the main character in this book?" with the answer "Harry Potter" might emerge. This type of QA pairs effectively evaluates a QA model's comprehension of the book, but for a reader, the provided information, though accurate, might seem superficial, even if they are not deeply familiar with the story.

To effectively leverage QA pairs in augmenting textual content, the work QALink[23] first addressed and formulated the task of text enrichment. It designed a novel system to enhance the reading experience of text documents by automatically integrating relevant QA content from sites like Quora and StackExchange. This system aims to provide readers with supplementary information that aids in understanding and deepening their knowledge of the document's content.

In the development of QALink, a neural network was trained to identify and retrieve relevant QA pairs, a task that has seen significant advancements with the advent of dense retrievers. These modern models are particularly adept at this retrieval task. However, applying dense retrievers directly for text enrichment can sometimes lead to a superficial engagement with the material. This is because they often highlight the most frequently mentioned

*Yixuan Tang and Anthony K.H. Tung are co-corresponding authors

entities, while overlooking the plethora of subtler details that are crucial for a thorough understanding of the text. For example, as shown in the right panel of Fig.18, the introductory paragraph of the Harry Potter Wiki mentions numerous entities. Yet, a RoBERTa model trained on the Natural Questions dataset tends to focus the top 100 QA pairs predominantly around the most prominent entity, 'Harry Potter', at the expense of less prominent ones.

When attempting to leverage a dense retrieval framework [12] directly for text enrichment, three challenges arise:

1. **Lack of Diversity** Dense retrievers have a tendency to favor QA pairs concerning prevalent entities, resulting in a repetitive and predictable selection that overlooks the richness of less common entities and diminishes the breadth of exploration for the reader

2. **Lack of Interestingness** The content surfaced by dense retrievers, while contextually accurate, often lacks the depth and engagement necessary to satisfy readers' curiosity or add meaningful insights to the text. In short, they are not interesting.

3. **Irrelevant QA pairs** Despite their technical proficiency, dense retrievers occasionally present QA pairs that are not closely related to the text. These pairs, while possibly relevant in a broader context, fail to align with the specific themes, characters, or events in the narrative, leading to a disjointed enrichment experience for the reader.

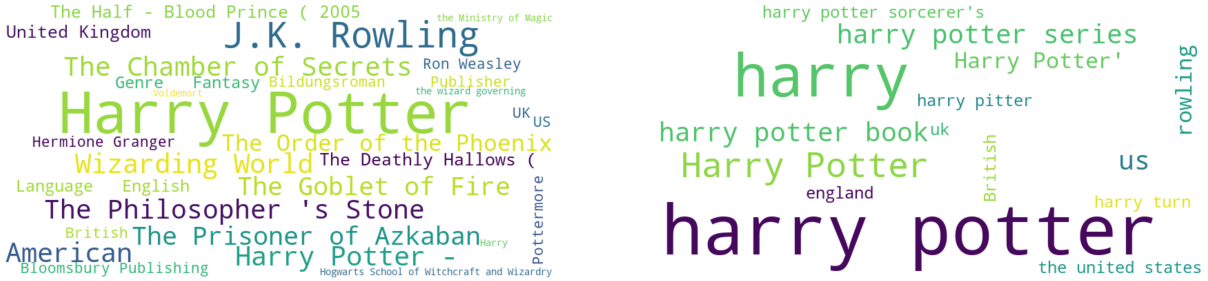


Figure 18: Comparison of the most frequently occurring named entities extracted. (Left) Entities derived from the introductory paragraph of the Harry Potter wiki page; (Right) Entities extracted from the top 100 QA pairs retrieved by querying the Harry Potter wiki content.

Our work confronts the challenge of enhancing text with QA datasets through a unique method. We construct an extensive entity co-occurrence graph from QA datasets, crucial for our text enrichment technique. We map entities from the text and corresponding QA pairs onto this graph to identify relevant subgraphs. Through rule-based path analysis, influenced by the psychological principles of novelty and complexity, we not only develop an interpretable scoring system but also unveil the nuanced connections between context and QA pairs. This approach enriches the text in a nuanced and engaging manner. By refining QA pair selection, our method ensures relevance and diversity while captivating the reader's interest at the same time. This advancement goes beyond traditional models, applying psychological theories of content interestingness in a practical, algorithmic way. Our framework stands out as a system that intertwines psychological concepts with computational applications, enhancing the reading experience by making it more engaging and informative. What's more, we have also discussed how our framework can be used in personal data management and personalization in Sec.4.

The main contributions of this paper can be summarized as follows:

- **Modeling Interestingness Through Psychology Principles:** We propose a model that assesses interestingness by considering both novelty and complexity, drawing inspiration from psychological research. This approach aims to provide a more nuanced understanding of what makes content engaging.

- **Entity Subgraphs for Context and QA Pair Representation:** Our method utilizes entity subgraphs as a simple yet effective way to represent context and question-answer pairs. We then employ path analysis to evaluate the interestingness of these contexts and QA pairs, offering an interpretable and efficient mechanism for analysis.
- **Optimization for Enhanced Contextual Coverage in QA Pairs:** We've formulated an optimization problem aimed at maximizing the coverage of context entities in the QA pairs retrieved and proposed a near-optimal solution that is both practical and efficient. This serves as a re-ranking strategy post-dense retrieval. This approach specifically addresses the challenge of forming robust representations for frequent entities, while also enhancing the distinction of less common ones in text enrichment tasks.

2 Related Work

Interestingness The concept of 'interestingness' in computer science encompasses various captivating objects termed as 'Fun Facts', 'Semantic Novelty', 'Trivia', and 'Unusual Aspects', with many studies highlighting rarity as a key element [7, 16, 17, 25]. However, rarity alone doesn't always equate to interestingness, as some rare objects may simply be unpopular. This notion leads to the broader question of what additional factors make an object interesting. Psychological research offers insights into this, linking interestingness with curiosity, exploration, and information seeking [3, 8, 9, 14, 19, 24]. Current research in computer science, drawing from these psychological models, focuses on novelty and complexity as quantifiable attributes of interestingness [1, 2, 22]. Recognizing this, our study emphasizes complexity alongside rarity in understanding and quantifying interestingness for text enrichment tasks, aiming to enrich textual content with elements that are not just rare but also genuinely engaging and thought-provoking.

Related Text Retrieval This series of research works involves finding relevant information or text passages based on a given query or input text. The variation of it that most related to our work is dense passage retrieval, the goal is to retrieve relevant documents or passages using dense vector representations of the texts. As proposed by [6], and the encoding of candidate answer phrases as vectors for efficient retrieval, as in [21], exemplify the precision and efficiency of dense retrieval. These methods ensure contextually aligned and content-rich text enrichment. Additionally, dense retrieval is adaptable in scenarios lacking direct answers, such as retrieving supporting documents from sources like Wikipedia before answer extraction, as suggested by [4]. In situations without gold standard answers, techniques like global normalization over potential answer spans, as per [5], are invaluable for covering a wide range of possible answers, thereby enhancing the informational breadth of the text. However, the direct application of dense retrievers in text enrichment can lead to a superficial engagement with the content. This occurs because these retrievers tend to focus on the most commonly mentioned entities, thereby overlooking many finer details crucial for a comprehensive understanding of the text. This inclination towards prominent entities over subtler details, as noted in [20], highlights a significant challenge in employing dense retrievers for nuanced text analysis.

Question and Answer Datasets The evolution of the question-answering (QA) domain has been significantly influenced by the development of datasets like SQuAD[18], Natural Questions[11], TriviaQA[10], and NarrativeQA[13]. Initially created as benchmarks for QA systems' reading comprehension, these datasets have become much more than assessment tools. They are now vast repositories of verified knowledge across diverse topics and formats, making them ideal for applications such as text enrichment. In text enrichment, the goal is to deepen the informational and contextual quality of textual content. The varied and real-world questions and answers in these QA datasets offer a unique resource for embedding detailed contextual and factual information into texts, enhancing their informativeness and engagement for users. This re-purposing of QA datasets for

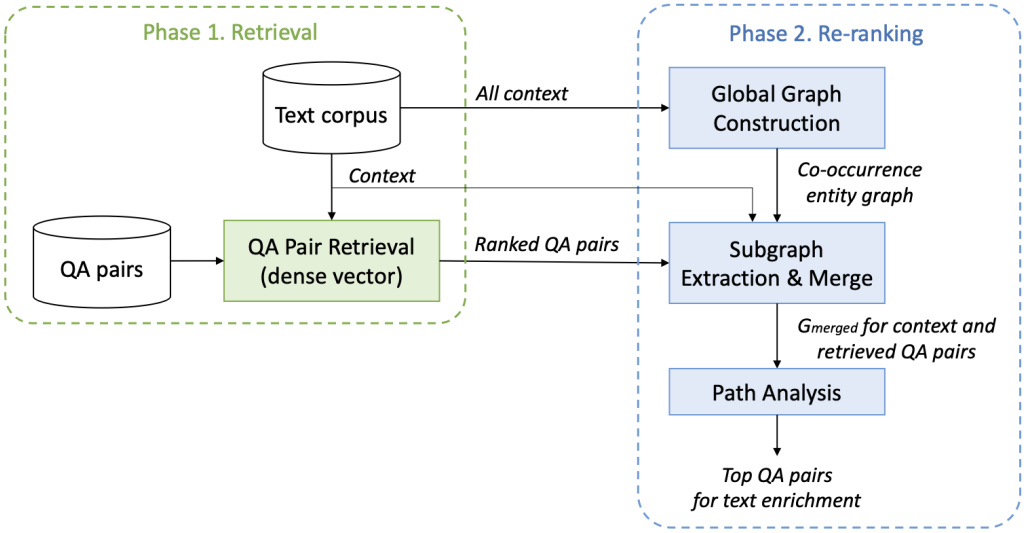


Figure 19: Framework Overview

text enrichment not only extends their use beyond traditional QA but also opens new paths in natural language processing and information retrieval, enhancing the quality and richness of textual content across various domains.

3 Methodology

3.1 Overview

Our framework operates as follows. Initially, we apply Named Entity Recognition (NER) to both a set of contexts (documents) and a set of QA pairs. From the contexts, we extract entities to build an entity co-occurrence graph, linking entities that appear within three sentences of each other. For any given context, we use its entities to identify a corresponding subgraph in this co-occurrence graph. The same process applies to a list of QA candidates. We then merge the subgraphs of each context and its respective QA pair into a unified subgraph. A path analysis is conducted on this united subgraph, using specially designed path patterns that assess novelty and relatedness, yielding an interpretable score. This score is used to re-rank the QA candidates. Additionally, to ensure the diversity of selected QAs, we optimize the QA candidates to maximize the coverage of linked context entities in the QA pairs, with each entity weighted according to its score from the path analysis step. This methodological approach facilitates a nuanced selection of QA pairs that are not only relevant but also diverse, enhancing the overall quality and informativeness of the enriched text. As shown in Fig.19.

3.2 Problem Formulation

Building upon the advancements in dense retrieval, our work introduces a nuanced problem formulation that explicitly captures both the 'interestingness' and 'relatedness' of QA pairs in relation to a given context. This dual consideration aims to enrich the textual landscape with engaging and pertinent information that transcends mere relevance, bringing to light the subtler aspects of human-like engagement and curiosity-driven exploration.

Our proposed problem formulation extends beyond the traditional objective of maximizing relevance. It introduces a composite measure that encompasses both the relatedness of QA pairs to the given context $\mathcal{C}$ and the intrinsic interestingness of the questions q_i and answers a_i in the dataset $\mathcal{D}$. The function $\phi : \mathcal{X} \to \mathbb{R}^d$ remains central in embedding textual content into a d-dimensional vector space, while the similarity function

sim : $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *is now complemented by an interestingness function int* : $\mathcal{Q} \times \mathcal{A} \to \mathbb{R}$*, which assesses the compelling nature of QA pairs. The combined optimization problem is then:*

$$\max_{\phi, int} \sum_{i=1}^{k} (sim(\phi(\mathcal{C}), \phi(q_i, a_i)), int(C, q_i, a_i)) \tag{1}$$

In defining the interestingness of a question-answer pair within a given context, we draw upon two fundamental psychological constructs: novelty and complexity. Novelty (*novelty*) reflects the degree to which information is new or surprising to a user, while complexity (*complexity*) represents the intricacy and depth of the information presented.

Our interestingness function, denoted conceptually as Interestingness$(C, (q_i, a_i))$, thus incorporates both novelty and complexity. This conceptual function is not directly quantifiable but serves as a theoretical framework guiding the development of our computational model:

$$int(C, q_i, a_i) \propto \text{Novelty}(C, q_i, a_i) \oplus \text{Complexity}(C, q_i, a_i) \tag{2}$$

Here, $\oplus$ denotes a conceptual combination of novelty and complexity, the specifics of which will be operationalized in the subsequent computational model. This formulation underscores the importance of both elements in constituting what makes a question-answer pair engaging to the user.

3.3 Proposed Methods

3.3.1 QA-pairs Retrieval

Drawing inspiration from the remarkable success of dense retrievers in passage retrieval tasks, as highlighted in [4], we have adopted this framework to retrieve relevant question-and-answer (QA) pairs for a given context. Specifically, we employ the RoBERTa-base model, which has been fine-tuned on the Natural Questions dataset using a contrastive learning approach. In this approach, positive and negative examples are constructed with a focus on the relationship between context and QA pairs. For a batch size of N, positive examples are N pairs such as $(C_1, q_1, a_1), (C_2, q_2, a_2), \ldots, (C_N, q_N, a_N)$, where each C_i represents a context and q_i, a_i its corresponding question and answer. Negative examples are generated by mismatching the contexts and QAs, pairing different C_i with q_j, a_j where $i \neq j$. The loss function is NT-Xtent as Eq.3, where τ is the temperature parameter that helps to control the scale of the similarity scores. This method enhances the model's ability to discern relevant and informative QA pairs in relation to a given context, leveraging the strengths of the RoBERTa model and the comprehensive nature of the Natural Questions dataset.

$$L = -\log \frac{\exp(\text{sim}(C_i, q_i, a_i)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\text{sim}(C_i, q_k, a_k))/\tau)}, \tag{3}$$

3.3.2 Co-occurance Entity Graph Construction

To assess the novelty and complexity of contexts and QA pairs, our approach begins with the construction of a co-occurrence entity graph that takes the contexts (documents) from a given QA dataset as input. This process starts by dividing a given context into individual sentences. Within each sentence, Named Entity Recognition (NER) is employed to identify entities that will form the nodes of the graph. Edges between these nodes are then established based on their proximity within the text, adhering to a predefined sentence boundary threshold. Specifically, a link is created between two nodes if the sentences containing their respective entities are within a certain number of sentences from each other, as determined by our threshold.

Let C be the input context, $C = \{S_1, S_2, \ldots, S_n\}$ where S_i represents a sentence in the context. We define V as the set of entities extracted through NER, $V = \{e_1, e_2, \ldots, e_m\}$, and $G = (V, L)$ as the resulting graph where

V is the set of nodes corresponding to entities in E, and L is the set of links between nodes. $pos(v)$ denotes the sentence position of the entity corresponding to node v, and θ is the sentence boundary threshold. The adjacency relationship A between nodes v_i and v_j is defined as follows:

$$A(v_i, v_j) = \begin{cases} 1, & \text{if } |\text{pos}(v_i) - \text{pos}(v_j)| \leq \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

In this way, we construct the co-occurrence entity graph to facilitate the extraction of subgraph representations for both context and QA pairs, enabling their further utilization.

3.3.3 Subgraph Representation for Context and QA pair

Given the co-occurrence entity graph G, we further define the subgraph representations for context $\mathcal{C}$ and QA pairs (q_i, a_i) as part of our retrieval framework. In this approach, G is the complete co-occurrence entity graph constructed from the entire text corpus. For a given context $\mathcal{C}$, the subgraph $G_{\text{sub}}(\mathcal{C})$ is a subset of G that represents the context. Similarly, $G_{\text{sub}}(q_i, a_i)$ denotes the subgraph for a QA pair, consisting of a question q_i and an answer a_i, which is also a subset of G. These subgraphs are constructed by identifying sets of entities $E_{\mathcal{C}}$ and E_{q_i,a_i} from the context and the QA pairs respectively, using Named Entity Recognition (NER).

$$G_{\text{sub}}(\mathcal{C}) = G(E_{\mathcal{C}}, \{(v_i, v_j) \mid A(v_i, v_j) = 1,\ v_i, v_j \in E_{\mathcal{C}},\ i \neq j\}), \tag{5}$$

$$G_{\text{sub}}(q_i, a_i) = G(E_{q_i,a_i}, \{(v_i, v_j) \mid A(v_i, v_j) = 1,\ v_i, v_j \in E_{q_i,a_i},\ i \neq j\}) \tag{6}$$

With the individual subgraph representations $G_{\text{sub}}(\mathcal{C})$ and $G_{\text{sub}}(q_i, a_i)$ established, we must now consider how these discrete elements can be synthesized to reflect the complex interplay between the context and the QA pairs. The integration of these subgraphs is pivotal in capturing the nuanced relationships that inform the relevance and interestingness of the QA pairs within their respective contexts. The ensuing step in our methodology, therefore, focuses on merging these subgraphs into a cohesive structure that embodies the full spectrum of informational relationships.

Given the subgraph representations $G_{\text{sub}}(\mathcal{C})$ for the context and $G_{\text{sub}}(q_i, a_i)$ for a QA pair, we merge them into a comprehensive subgraph G_{merged}, which includes paths connecting nodes from $G_{\text{sub}}(\mathcal{C})$ to $G_{\text{sub}}(q_i, a_i)$. We use A^* to represent the transitive closure of the adjacency matrix A we defined in (4). This process forms a unified representation that encapsulates the context, the QA pair, and the semantic links between them.

$$G_{\text{merged}}(G_{sub}(C), G_{sub}(q_i, a_i)) = (V_{\mathcal{C}} \cup V_{q_i,a_i}, L_{\mathcal{C}} \cup L_{q_i,a_i} \cup (v_c, v_q) \mid v_c \in V_{\mathcal{C}}, v_q \in V_{q_i,a_i}, A^*(v_c, v_q) > 0) \tag{7}$$

Utilizing Equations 5, 6, and 7, we derive the subgraph representations for both context and QA pairs. These representations are subsequently employed in our path analysis.

3.3.4 Path Analysis

Based on the subgraph representation of context and QA pairs, we introduce an interpretable algorithm designed for path analysis. The algorithm classifies paths into three distinct categories. It identifies trivial paths where the start and end nodes are the same, reflecting direct, uncomplicated connections, and adjusts their scores using the γ parameter. For paths introducing new or uncommon connections, the algorithm recognizes these as novelty-inducing, incrementing their score based on the presence of nodes outside the typical context and QA sets, a process guided by the α parameter. Additionally, it accounts for hub-influenced paths by detecting nodes with high connectivity, adjusting the score to reflect the influence of these central hubs through the β parameter.

Algorithm 1: Path Analysis

Require: G_{merged}, A^*, $V_{\mathcal{C}}$, $V_{q,a}$, α, β, γ, θ

for $v \in V_{\mathcal{C}}$ **do**
  Initialize $pathScore \leftarrow 0$
  for $u \in V_{q,a}$ **do**
    if $A^*(v,u) > 0$ **then**
      for all $p \in \mathcal{P}(v,u)$ **do**
        Initialize $score_R, score_\alpha, score_\beta \leftarrow 0$
        if $v = u$ **then**
          $score_R \leftarrow score_R + \gamma$
        end if
        for $w \in p$ **do**
          if $w \notin V_{\mathcal{C}} \cup V_{q,a}$ **then**
            $score_\alpha \leftarrow score_\alpha + \alpha$
          end if
          if degree$(w, G_{\text{merged}}) \geq \theta$ **then**
            $score_\beta \leftarrow score_\beta - \beta$
          end if
        end for
        $pathScore \leftarrow$
        $pathScore + score_R + score_\alpha + score_\beta$
      end for
    end if
  end for
end for$pathScore$ =0

Algorithm 2: Greedy Question-Answer Selection

Require: Set of QAs $\mathcal{Q}$, Score function $s(q)$, Context entities $C(q)$

Initialize $\mathcal{S} \leftarrow \emptyset$
Initialize TotalScore $\leftarrow 0$
Initialize CoveredEntities $\leftarrow \emptyset$
while $\mathcal{Q} \neq \emptyset$ **do**
  bestQA $\leftarrow$ null
  bestIncrement $\leftarrow 0$
  for all $qa \in \mathcal{Q}$ **do**
    newEntities $\leftarrow C(qa) -$ CoveredEntities
    if newEntities $\neq \emptyset$ **then**
      increment $\leftarrow s(qa) \times \frac{|\text{newEntities}|}{|C(qa)|}$
      if increment $>$ bestIncrement **then**
        bestIncrement $\leftarrow$ increment
        bestQA $\leftarrow qa$
      end if
    end if
  end for
  if bestQA $=$ null **then**
    break
  end if
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{\text{bestQA}\}$
  TotalScore $\leftarrow$ TotalScore $+ s($bestQA$)$
  CoveredEntities $\leftarrow$
  CoveredEntities $\cup C($bestQA$)$
  $\mathcal{Q} \leftarrow \mathcal{Q} -$ {bestQA} $\mathcal{S}$, TotalScore
end while=0

This scoring system, grounded in clear criteria, provides an interpretable and methodical way to analyze the dynamics of paths within our context and QA pair framework.

This algorithm, shown in Algo.1, determines the path score for each context and corresponding QA pair. The derived path score can be directly employed to re-rank QA candidates or further refined through a subgraph optimization step.

3.3.5 Subgraph Optimization

Having demonstrated the path analysis algorithm to evaluate individual context and QA pairs, we now shift our focus to the overarching goal in text enrichment tasks: selecting an optimal set of QA pairs for a given context. This necessitates not only evaluating individual pairs but also ensuring that the chosen set of QA pairs is sufficiently diverse.

To effectively evaluate the question-answer (QA) pairs in relation to a given context, we propose an optimization model aimed at maximizing the coverage of context entities, each weighted by its relevance score. This model is designed to identify the most informative and relevant QA pairs by considering the scores of context entities they relate to.

Let $\mathcal{C}$ be the set of all contexts, $\mathcal{Q}$ be the set of QA pairs, and $\mathcal{C}(qa)$ to denote the context entities covered by

the QA pair. The objective is to select a subset of QA pairs $S \subseteq Q$ such that the sum of scores of the covered context entities is maximized.The optimization problem can be formulated as follows:

$$\begin{aligned} &\text{Maximize} \quad \sum_{qa \in S} \sum_{c \in \mathcal{C}(qa)} \text{PathScore}(c, qa) \\ &\text{subject to} \quad S \subseteq \mathcal{Q} \end{aligned} \tag{8}$$

A greedy algorithm is utilized for this optimization. At each step, it selects the QA pair that contributes the highest score to the uncovered context entities in the current set S. This method maximizes the total score at each step, effectively ensuring a diverse coverage of context entities, rather than solely focusing on their individual relevance scores.

This is subject to the condition that $\mathcal{S}$ encompasses a broad range of context entities, as represented by:

$$\text{Coverage}(\mathcal{S}) = \bigcup_{qa \in \mathcal{S}} C(qa) \tag{9}$$

Eq9 is deemed submodular, reflecting the principle of diminishing returns. For any two sets $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{Q}$ and an element $q' \in \mathcal{Q} \setminus \mathcal{B}$, we have:

$$f(\mathcal{A} \cup \{q'\}) - f(\mathcal{A}) \geq f(\mathcal{B} \cup \{q'\}) - f(\mathcal{B}) \tag{10}$$

This condition ensures that the incremental benefit of adding a QA pair to the subset decreases as the subset becomes larger, thereby reflecting the overlap in context entity coverage among the selected QA pairs. This property is crucial for maintaining diversity in the selected subset of QA pairs. This approach is especially advantageous in large-scale problems where precise optimization is not feasible. The algorithm provides a practical and near-optimal solution by balancing the coverage of diverse context entities.

4 QALinkPlus in Personal Document Enrichment and Personalized QA Pairs

The relationship between personal data management and personalization is characterized by a fundamental tension: personal data management focuses on safeguarding user privacy and implementing measures to protect personal information, while personalization relies on accessing and utilizing this data to create tailored experiences for users. Recognizing this, QALinkPlus innovatively addresses this tension. Unlike typical deep learning models that risk privacy leaks by retaining training data details, our approach achieves personalization post-training, particularly during the re-ranking phase, without relying on deep neural networks or requiring the uploading of personal data, thus enhancing privacy. This section delves into its application in two specific areas: personal document enrichment and personalized QA pairs, showcasing the framework's capability to achieve personalization while protecting users' privacy.

Our framework's approach to personal document enrichment stands in stark contrast to traditional deep learning-based personalization methods. Unlike these methods that typically depend on training with large datasets, including sensitive personal documents, and often involve uploading this data to remote servers, our framework employs a graph-based algorithm for personalization during the re-ranking phase, processing and personalizing content within a user-controlled environment, reducing the reliance on training with sensitive data. As a result, our approach could maintain the confidentiality of personal documents, which contributes to personal data management and data personalization in the context of text enrichment.

Furthermore, our framework has the capability to offer personalized QA pairs by adjusting hyperparameters or applying weights to particular entities that align with user interests. This form of personalization caters to various application scenarios, enhancing user engagement and experience. For instance, within the framework, users have the flexibility to modify parameters like the α value to influence the novelty of enrichment contents. This

could be particularly useful in different contexts — for example, opting for a larger α when reading for leisure to explore a wider range of topics, or choosing a smaller α when seeking specific information in professional documents. Such customization allows users to tailor their experience to their current needs and preferences, showcasing the adaptability of the framework in providing relevant and personalized content.

5 Experiments

5.1 Entity Diversity and Coverage

In this experiment, we aim to evaluate the effectiveness of our approach in modeling the novelty aspect of interestingness, focusing on the impact of our algorithm on the quality of question-answer (QA) selections. Central to our investigation is the analysis of how the algorithm affects the coverage and diversity of context entities within the selected QAs. By implementing a series of carefully crafted metrics, we seek to quantify the degree to which our method expands the scope of covered entities and enhances the diversity in the QA pairs. This analysis is pivotal in assessing the practical impact of our optimization strategy in real-world QA systems, where the depth and variety of presented information are key to user satisfaction and engagement. Through a comparative evaluation of the original and optimized QA sets, this study aims to highlight the concrete advantages of our algorithm in improving the selection process, particularly in terms of introducing novelty and enriching the content's interestingness.

Dataset In this experiment, we utilized four prominent datasets, randomly sampling 100 documents from each for context. We worked with 316,034 QA pairs across all datasets, leading to the comparison of more than one hundred million query-candidate pairs in total.

- **Natural Questions (NQ)**[15] Curated by Google, this dataset is a cornerstone in open-domain QA research, featuring over 300,000 real user queries paired with relevant Wikipedia articles. NQ emphasizes realistic scenarios, requiring systems to navigate diverse sources for answering user queries.
- **TriviaQA** This dataset presents a realistic text-based QA challenge with 950,000 question-answer pairs from 662,000 documents sourced from Wikipedia and the web. TriviaQA's complexity lies in its long contexts and the need for models to go beyond span prediction for answers.
- **Stanford Question Answering Dataset (SQuAD)**[18] SQuAD includes over 100,000 question-answer pairs based on Wikipedia articles. Its uniqueness stems from the format of answers, which can be any sequence of tokens from the text, and the inclusion of both answerable and unanswerable questions in its latest version.
- **NarrativeQA**[13] NarrativeQA is designed to test reading comprehension on long documents, with a focus on narrative-style content. It includes Wikipedia summaries, links to full stories, and diverse question types, offering a comprehensive challenge for QA models.

Evaluation Metrics In our analysis, we employed three key metrics to evaluate and compare the original and re-ranked question-answer (QA) sets across various datasets.

- **Total Unique Context Entity Coverage** This metric measures the cumulative number of unique context entities covered by the QAs in a dataset. It provides insight into the breadth of information encompassed by the QA pairs, highlighting the algorithm's ability to capture a wide range of relevant topics and details.
- **Average Context Entity Coverage per QA** This metric calculates the average number of unique context entities covered by each individual QA. It offers a more granular view of the coverage, focusing on the depth and richness of information each QA contributes to the overall dataset.

- **Entropy (Entity Coverage Diversity)** Entropy is used to assess the diversity of context entity coverage among the QAs. A higher entropy value indicates a more evenly distributed coverage across different entities, suggesting a greater variety in the types of information addressed by the QAs.

Table 7: Comparison of Original, ReRanked, and Optimized Results Across Datasets. The table showcases metrics including Total Unique Coverage (TUC), Average Coverage per QA (ACPQ), and Entropy. Here, Δ denotes the improvement of optimized results compared to the original and re-ranked results.

Dataset	Metric	Original	ReRanked	Optimized	ΔOriginal (%)	ΔReRanked (%)
NarrativeQA	TUC	2.52	5.25	6.89	+173.66	+31.26
	ACPQ	1.38	2.43	2.87	+107.52	+18.38
	ENTROPY	0.75	1.81	2.26	+199.51	+24.44
Squad	TUC	2.71	4.62	5.63	+108.15	+21.97
	ACPQ	1.38	2.53	2.87	+108.33	+13.05
	ENTROPY	1.08	1.85	2.13	+97.65	+14.59
TriviaQA	TUC	7.65	11.40	14.28	+86.65	+25.26
	ACPA	1.91	2.94	3.56	+86.18	+20.98
	ENTROPY	2.08	2.96	3.36	+61.62	+13.30
Natural Questions	TUC	18.03	17.80	21.55	+19.50	+21.05
	ACPQ	6.14	6.36	7.18	+16.91	+12.98
	ENTROPY	3.42	3.51	3.83	+12.09	+9.11

The results of diversity and coverage comparison, illustrated in the Fig20, distinctly showcase the efficacy of our re-ranking method in enhancing the novelty aspect of interestingness. Across various datasets, the re-ranked sets consistently outperform the original ones in Total Unique Context Entity Coverage and Average Context Entity Coverage per QA, indicating a broader and more diverse range of context entities being addressed. While the Entropy metric remains stable, suggesting a balanced distribution of entities, the overall increase in coverage metrics confirms that our approach successfully selects more novel question-answer pairs, aligning with our objective of providing novel contents in text enrichment tasks.

The comparative analysis of the results pre- and post-subgraph optimization is presented in Table 7. It is evident that the performance significantly improves with the optimized results compared to the original outcomes derived from direct dense retrieval, and it also surpasses the direct reranked results based on our path analysis. This substantiates the effectiveness of our optimization method in procuring a set of QA pairs that is not only diverse but also offers an enhanced collective quality over focusing solely on individual pairs.

5.2 Assessing the Impact of Re-ranking on Superficial QA Reduction

This experiment aims to assess our approach's effectiveness in capturing the complexity aspect of interestingness. Utilizing Algo.1, we assign scores to QA pairs based on criteria from path analysis, then re-rank QAs initially retrieved by a standard dense retrieval framework. We hypothesize that our approach, which enhances complexity and novelty, will reduce the prevalence of superficial QAs in the top results. This hypothesis will be tested by comparing the superficial QA distribution in both the original and re-ranked lists.

We employ the Natural Questions dataset [15] for its categorization of short-answer questions, often considered superficial compared to the complex and deep 'interestingness' QAs our research targets. Our evaluation metric is the percentage of superficial QAs in the top n results, where a QA is labeled superficial if it's answerable with a short response, which is usually one or two words.

The results, depicted in Fig.21, demonstrate a notable decrease, more than 10% at the beginning, in the prevalence of superficial QA pairs within the re-ranked results, with the re-ranking line consistently positioned below that of the original. This trend indicates that our path analysis algorithm has effectively prioritized QAs of greater complexity. Initially, the original results exhibit a high percentage of superficial QAs, aligning with findings from [20]. The tendency to favor frequently occurring entities often results in the initial selection of less complex QA pairs. As n increases, the percentage of superficial QAs in the re-ranked results gradually rises, while it decreases in the original dense retrieval results. This pattern can be attributed to the limited availability of non-superficial QAs in certain contexts, exemplified by the 'Harry Potter' case discussed in Sec.1.

6 Conclusion

In this study, we tackled the challenge of enhancing text enrichment using QA datasets like Natural Questions and SQuAD through an innovative approach involving the creation of an extensive entity co-occurrence graph, from which context-specific subgraphs were derived. This led to a rule-based path analysis and a novel scoring system, assessing each QA pair's relevance and engagement value, thus enriching the reader's experience. Additionally, our framework discusses aspects of personal data management and personalization, suggesting ways to align personalized content with individual privacy needs. Our methodology's effectiveness was evident in two key experiments: Experiment 5.1 highlighted our re-ranking method's ability to enhance novelty, as shown by improved coverage metrics, and Experiment 5.2 demonstrated a significant reduction in superficial QAs, emphasizing the prioritization of more complex, contextually relevant content. These results underscore our approach's potential to fill knowledge gaps and captivate readers, marking a significant step forward in text enrichment using QA datasets.

7 Acknowledgments

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative and the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and the Ministry of Education, Singapore.

References

[1] Daniel E Berlyne. Conflict, arousal, and curiosity. 1960.

[2] Daniel E Berlyne. Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation. Hemisphere, 1974.

[3] Frederick E Bolton. Attention and interest: A study in psychology and education. The Journal of Philosophy, Psychology and Scientific Methods, 7(17):474–475, 1910.

[4] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In ACL (1), pages 1870–1879. Association for Computational Linguistics, 2017.

[5] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In ACL (1), pages 845–855. Association for Computational Linguistics, 2018.

[6] Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. Multi-step retriever-reader interaction for scalable open-domain question answering. In ICLR (Poster). OpenReview.net, 2019.

[7] Nausheen Fatma, Manoj Kumar Chinnakotla, and Manish Shrivastava. The unusual suspects: Deep learning based mining of interesting entity trivia from knowledge graphs. In AAAI, pages 1107–1113. AAAI Press, 2017.

[8] Barbara L Fredrickson. What good are positive emotions? Review of general psychology, 2(3):300–319, 1998.

[9] Carroll E Izard and Brian P Ackerman. Motivational, organizational, and regulatory functions of discrete emotions. Handbook of emotions, 2:253–264, 2000.

[10] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In ACL (1), pages 1601–1611. Association for Computational Linguistics, 2017.

[11] Hidetaka Kamigaito, Jingun Kwon, Young-In Song, and Manabu Okumura. A new surprise measure for extracting interesting relationships between persons. In EACL (System Demonstrations), pages 231–237. Association for Computational Linguistics, 2021.

[12] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In EMNLP (1), pages 6769–6781. Association for Computational Linguistics, 2020.

[13] Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. Trans. Assoc. Comput. Linguistics, 6:317–328, 2018.

[14] Andreas Krapp. Interest, motivation and learning: An educational-psychological perspective. European journal of psychology of education, 14:23–40, 1999.

[15] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. Trans. Assoc. Comput. Linguistics, 7:452–466, 2019.

[16] Nianzu Ma, Alexander Politowicz, Sahisnu Mazumder, Jiahua Chen, Bing Liu, Eric Robertson, and Scott Grigsby. Semantic novelty detection in natural language descriptions. In EMNLP (1), pages 866–882. Association for Computational Linguistics, 2021.

[17] Abhay Prakash, Manoj Kumar Chinnakotla, Dhaval Patel, and Puneet Garg. Did you know? - mining interesting trivia for entities from wikipedia. In IJCAI, pages 3164–3170. AAAI Press, 2015.

[18] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In EMNLP, pages 2383–2392. The Association for Computational Linguistics, 2016.

[19] Gregory Schraw and Stephen Lehman. Situational interest: A review of the literature and directions for future research. Educational psychology review, 13:23–52, 2001.

[20] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. arXiv preprint arXiv:2109.08535, 2021.

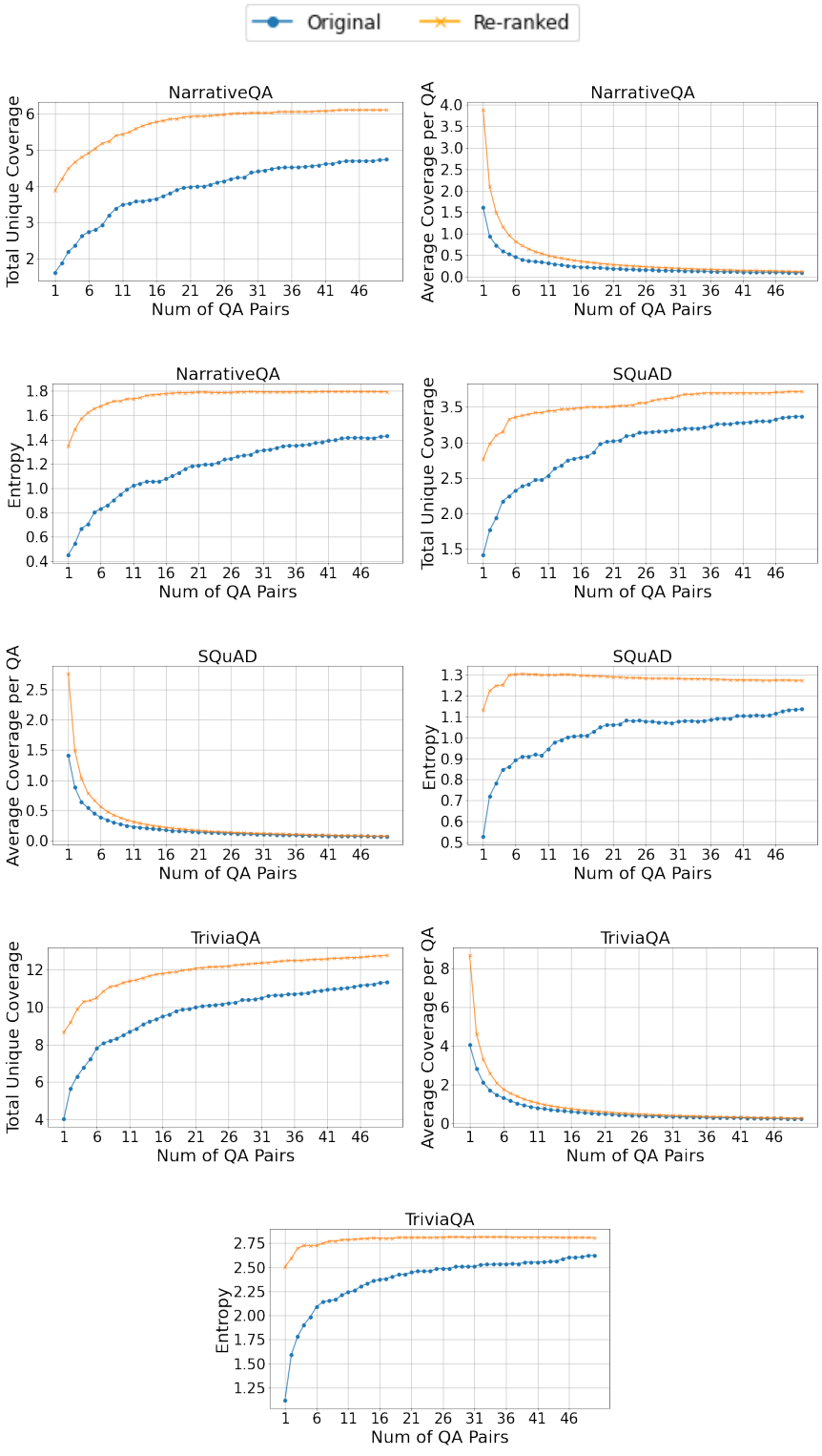
[21] Min Joon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur P. Parikh, Ali Farhadi, and Hannaneh Hajishirzi. Real-time open-domain question answering with dense-sparse phrase index. CoRR, abs/1906.05807, 2019.

[22] Paul J Silvia. What is interesting? exploring the appraisal structure of interest. Emotion, 5(1):89, 2005.

[23] Yixuan Tang, Weilong Huang, Qi Liu, Anthony K. H. Tung, Xiaoli Wang, Jisong Yang, and Beibei Zhang. Qalink: Enriching text documents with relevant q&a site contents. In CIKM, pages 1359–1368. ACM, 2017.

[24] Silvan Tomkins. Affect imagery consciousness: Volume I: The positive affects. Springer publishing company, 1962.

[25] David Tsurel, Dan Pelleg, Ido Guy, and Dafna Shahaf. Fun facts: Automatic trivia fact extraction from wikipedia. In WSDM, pages 345–354. ACM, 2017.


Original
Re-ranked
NarrativeQA
Total Unique Coverage
Num of QA Pairs
NarrativeQA
Average Coverage per QA
Num of QA Pairs
NarrativeQA
Entropy
Num of QA Pairs
SQuAD
Total Unique Coverage
Num of QA Pairs
SQuAD
Average Coverage per QA
Num of QA Pairs
SQuAD
Entropy
Num of QA Pairs
TriviaQA
Total Unique Coverage
Num of QA Pairs
TriviaQA
Average Coverage per QA
Num of QA Pairs
TriviaQA
Entropy
Num of QA Pairs
Natural Questions
Total Unique Coverage
Num of QA Pairs
Natural Questions
Average Coverage per QA
Num of QA Pairs
Natural Questions

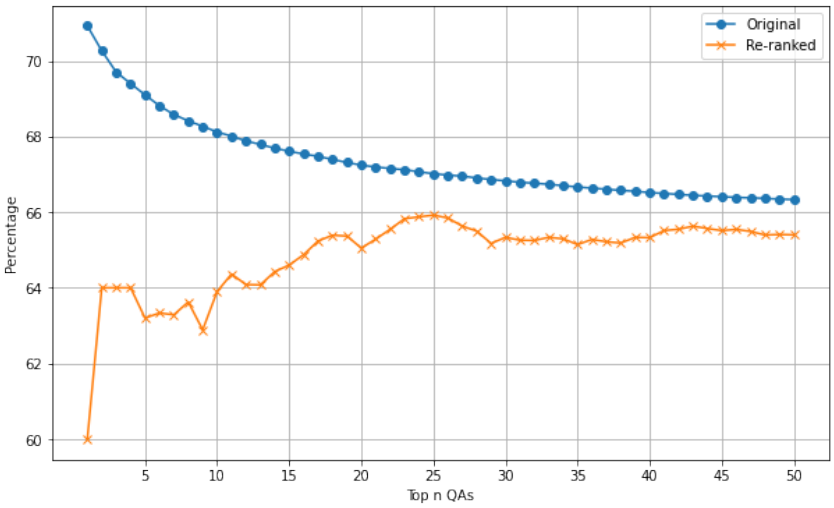


Figure 21: Percentage of Superficial QAs in Natural Questions