THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators: A Reproducibility Study

by
Krista Bradshaw

School of Information Technology and Electrical Engineering,
University of Queensland.

Submitted for the degree of Bachelor of Engineering (Honours)
in the division of Software Engineering.

October 25, 2023

Krista Bradshaw
krista.bradshaw@uqconnect.edu.au

October 25, 2023
Prof Michael Bruenig
Head of School
School of Information Technology and Electrical Engineering
The University of Queensland
St Lucia, Q 4072
Dear Professor Bruenig,

In accordance with the requirements of the degree of Bachelor of Engineering (Honours) in the School of Information Technology and Electrical Engineering, I submit the following thesis entitled:

"Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators: A Reproducibility Study".

This work was performed under the supervision of Dr. Guido Zuccon. I declare that the work submitted in this thesis is my own, except as acknowledged in the text and footnotes, and has not been previously submitted for a degree at The University of Queensland or any other institution.

Yours sincerely,

Krista Bradshaw

Abstract

This thesis explores the significance of query variations in information retrieval pipelines, replicating and expanding upon a seminal study by Gustavo Penha, Arthur Câmara, and Claudia Hauf. The study investigates how query variations, slightly modified versions of search queries, can affect retrieval efficiency.

Using ten established methods, query variations were automatically generated and applied to three datasets: ANTIQUE, TREC-DL-2019, and DL-TYPO. This revealed that synonym and paraphrasing methods sometimes produced queries with different meanings due to compounded misspellings.

The experiments involved using BM25 as an initial retriever and re-ranking the top 100 results with various models. This process was applied to original queries and their variations, with retrieval effectiveness measured using nDCG@10. Additionally, the study explored the impact of combining rankings from different query variation categories.

The results demonstrated that query variations significantly impact retrieval effectiveness. The replication phase confirmed the importance of exploring query variations while introducing the DL-TYPO dataset and highlighted the need to consider dataset-specific characteristics. In total, 51, 41, and 32 instances in ANTIQUE, TREC, and DL-TYPO, respectively, showed significant drops in effectiveness. Combining query variation categories often improved results but generally fell short of the original effectiveness.

In summary, this thesis enhances our understanding of how query variations affect retrieval pipeline robustness, confirming their crucial role in evaluating retrieval systems. The study successfully reproduced and expanded upon the original findings, offering insights into the impact of query variation generators. All code is available at `https://github.com/krista-b/reit4841`. update

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the digital age, access to vast information is at our fingertips. Retrieval pipelines, the core technology behind search engines and recommendation systems, are pivotal in helping us navigate this information landscape. These pipelines are tasked with finding the most relevant and accurate results in response to our queries, making them indispensable tools in our everyday lives.

However, the effectiveness of retrieval pipelines can be significantly influenced by variations in the way queries are formulated. Queries are rarely uniform; they can vary in wording, syntax, and structure while still aiming to convey the same information need. Understanding how these variations impact the performance of retrieval pipelines is crucial in ensuring that these systems continue to meet the ever-evolving needs of users.

1.1 Objectives and Scope

This thesis embarks on a comprehensive exploration of the robustness and generalisability of findings from an original study conducted in information retrieval. The study, authored by Gustavo Penha, Arthur Câmara, and Claudia Hauf, titled "Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators," and hereafter referred to as the "original study," has yielded valuable insights into the performance of retrieval pipelines in the presence of query variations. This research has highlighted their limitations and possible strategies for enhancement. The present study is guided by two fundamental research questions, each aimed at extending our comprehension of retrieval systems' behaviour and the implications of these insights on broader datasets. The following two research questions categorise the overarching themes of this study:

RQ1: To what extent can the findings and observations of the original study be successfully reproduced using the same datasets, models, and

methodologies?

RQ2: How do the conclusions drawn from the original study generalise when applied to additional datasets, and what insights can be gained from this broader perspective?

1.1.1 Research Question 1: Reproduction

The first research question assesses the reproducibility of findings and observations from the original study. Specifically, it seeks to answer to what extent the original study's results can be successfully reproduced using the same datasets, models, and methodologies. This investigation into the original study's findings' reproducibility is paramount. Reproducibility not only underscores the reliability of the original study but also lays the foundation for building upon its insights. By meticulously replicating the conditions of the original study, this research aims to validate the robustness of the observed phenomena, ultimately contributing to the establishment of more resilient retrieval pipelines.

The objective for Research Question 1 is twofold. Firstly, it aims to reproduce the critical findings of the original study using identical datasets, models, and methodologies, thereby evaluating the reliability and stability of the original observations. Secondly, it endeavours to identify any potential discrepancies or variations that may arise during the reproduction process. These discrepancies, if present, will be scrutinised to ascertain their underlying causes, which may encompass differences in dataset characteristics, model behaviour, or methodological nuances.

1.1.2 Research Question 2: Expansion

The second research question delves into the generalisability of the original study's conclusions when applied to additional datasets. This question posits that while the original study has provided valuable insights, its scope was delimited to specific datasets. Consequently, it raises the question of how well the conclusions drawn from the original study extend to a broader spectrum of datasets. This broader perspective is critical for understanding the phenomena under investigation, allowing for identifying overarching trends and patterns.

The objective of Research Question 2 is to investigate how the insights and conclusions of the original study hold when exposed to a diverse range of datasets beyond those initially considered. This entails applying the same methodologies and models used in the original study to these additional datasets. Through this process, the research aims to uncover insights and nuances that may emerge in different dataset contexts, shedding light on the generalisability of the original study's findings and providing a richer understanding of retrieval pipeline behaviour.

1.2 Thesis Structure

update if changes

The thesis is organised into several essential chapters, each dedicated to specific aspects of the research inquiry. It commences with the introductory chapter (Chapter 1), establishing the foundation by presenting two pivotal research questions and outlining the objectives. This chapter also offers an overview of the thesis's structure.

Following the introduction, Chapter 2 delves into the necessary background information to comprehensively understand information retrieval, natural language processing, ranking models, and evaluation metrics. This chapter lays the theoretical groundwork upon which subsequent analyses are constructed.

Chapter 3 explores the landscape of related work, focusing on previous research on large language models, enhancements in retrieval pipelines, evaluation metrics, and query variation techniques. This contextualisation provides insights into the existing body of knowledge and underscores the significance of the present research.

In Chapter 4, the methodology utilised for conducting experiments is elaborated upon. This includes the description of datasets used, ranking models, training processes, query generators, assessment of query generator quality, and the application of evaluation metrics. A thorough understanding of this methodology is crucial for comprehending the subsequent empirical findings.

Chapter 5 presents the results and discussions arising from the empirical investigations, divided into three sections, each corresponding to generation quality and one of the research questions posed in the introduction.

The final chapter, Chapter 6, serves as the conclusion and culmination of the thesis. It summarises the research's primary findings and conclusions, addressing the research questions. Furthermore, it reflects on future research avenues, identifying areas in information retrieval and query variation analysis that warrant further exploration.

Chapter 2

Background

This chapter lays the foundation for understanding the core concepts and context that underpin this thesis. It provides essential insights into information retrieval, large language models, and the evaluation of retrieval pipelines. Furthermore, it introduces the concept of query variations, which forms the basis of the subsequent research. This chapter serves as a critical backdrop for comprehending the challenges addressed and the solutions proposed in the following chapters.

2.1 Information Retrieval

Information Retrieval (IR) is the process of retrieving relevant information from extensive data collections, typically text documents or web pages. IR systems use various techniques and algorithms to analyse and index large datasets, such as search engine indexes, document repositories or databases. When a user enters a search query, the IR system searches through the indexed data and returns a list of documents or web pages in order of relevance.

2.1.1 Large Language Models

Large Language Models (LLMs) are machine learning models used in Natural Language Processing (NLP) applications that are trained on copious collections of text to generate natural language text or predictions [1]. LLMs have diverse real-world applications, including chatbots, language translation services, and voice assistants. Furthermore, they have facilitated new applications in content generation, language modelling, and text summarisation [2].

add diagram and ref.

Retrieval pipelines are a series of processing steps used in IR systems to retrieve

Figure 2.1: Caption

relevant information from a large data collection, such as a document database or a search engine index. Retrieval pipelines typically involve the following steps:

1. Query processing: The queries are processed to identify relevant terms, expand the query with synonyms, and apply other query optimisation techniques.
2. Document indexing: The documents in the collection are indexed, typically using techniques such as token embedding.
3. Ranking: The ranked documents are then sorted by relevance using models.
4. Presentation: The final step involves presenting the search results to the user in a user-friendly format, such as a list of documents with summaries [3, 4].

Retrieval pipelines can also include additional steps, such as filtering or classifying the search results based on domain-specific criteria or using machine learning techniques to improve relevance comparisons [5].

2.2 Natural Language Processing

2.2.1 Ranking

more - add a diagram.

Ranking refers to ordering a set of text documents or search results based on their relevance to a given query or topic. Ranking algorithms typically use various features to assign scores to each document. Then those documents are ranked in descending order of their score, having the most relevant documents appearing first [6]. The goal of ranking is to improve the performance of IR systems by ensuring that the most relevant documents appear to the user first. Ranking is critical to many applications, such as search engines, chatbots, and question-answering systems.

2.2.2 Rank Fusion

add my references - maybe add diagram or dot points.

In information retrieval, rank fusion is a valuable technique that amalgamates ranked document lists generated by multiple retrieval methods or models. The primary objective is to enhance the retrieval process's overall efficacy by integrating outcomes from various sources.

The fundamental concept underlying rank fusion revolves around assigning scores or weights to individual documents within the ranked lists acquired from diverse retrieval methods. These set values are subsequently employed in constructing a novel merged list where documents are arranged based on their rankings. A commonly utilised approach for rank fusion is Reciprocal Rank Fusion (RRF).

RRF is a technique used in information retrieval to combine the rankings generated by multiple retrieval methods or queries. It's advantageous when you have different sources of ranking scores, such as various retrieval models or query variations, and you want to aggregate them to improve overall retrieval performance. The core idea behind RRF is to assign a reciprocal rank score to each document in the individual rankings. Reciprocal rank is a measure that gives higher scores to documents that appear higher in the rankings. Then, these reciprocal rank scores from different rankings are summed up for each document, and the documents are re-ranked based on this aggregated score.

Rank fusion is advantageous when distinct retrieval techniques possess complementary strengths and weaknesses. Combining their findings makes it feasible to bolster overall performance in retrieving relevant documents while potentially enhancing quality standards. Nevertheless, meticulous design and evaluation of rank fusion methodologies remain crucial to effectively capture each source's advantages and generate meaningful merged rankings accordingly.

2.2.3 Training

more info.

Training refers to teaching a model to perform specific NLP tasks, such as text classification, sentiment analysis, or translation [7]. The training process involves feeding large amounts of text data into the model and corresponding target labels. The model then learns to recognise relationships and patterns based on the labelled examples provided [8]. The training process typically involves iterating over the data multiple times, using optimisation techniques to adjust the model's parameters and improve its performance [8]. The ultimate training goal is to develop models that can accurately and efficiently process natural language text for various applications.

2.3 Effectiveness Evaluation

The effectiveness of LLMs is evaluated by measuring performance on specific NLP tasks. Typically, this evaluation is conducted offline using a benchmark dataset and a well-defined task, such as language translation, sentiment analysis or text classification. LLMs are commonly evaluated using a variety of datasets and tasks, including the BM25 benchmark and other similar datasets. The results of evaluations like these can be used to compare the effectiveness of different LLMs and to identify areas that need strengthening [1].

Online evaluation is a technique to assess the effectiveness of an IR system in real-time using real-world applications. Real-world applications involve deploying LLM systems in actual settings and measuring their performance and impact on user outcomes, such as improved efficiency. In contrast to offline evaluation, which is performed on pre-existing data sets, online evaluation measures a system's performance by observing how users interact with it during live usage.

In addition to data-focused evaluations, the effectiveness of LLMs can also be assessed through user studies. User studies typically involve asking users to perform specific tasks or interact with a system and collecting data on their satisfaction and experience. Several types of user studies exist, including surveys and usability testing.

2.3.1 Metrics

refs, ai check

Evaluating retrieval systems in information retrieval is a pivotal aspect of assessing their performance. A diverse set of metrics is at the disposal of researchers, each shedding light on distinct facets of system effectiveness. This study employs nDCG (Normalized Discounted Cumulative Gain) as the primary evaluation metric. However, it's worth noting that the evaluation landscape includes other widely recognised metrics such as precision, recall, the F1 score, and several more, each providing a comprehensive view of system performance. These metrics collectively contribute to a nuanced understanding of the retrieval process and its outcomes.

- nDCG (Normalized Discounted Cumulative Gain): nDCG, a widely recognised metric in the field of information retrieval, holds particular relevance in situations where the order of search results bears significant importance. This metric is purposefully designed to assess the quality of ranked lists, considering both the relevance of documents and their respective positions within the list. nDCG achieves this by calculating the cumulative gain attributed to relevant documents while appropriately discounting the importance of items lower down the ranking. Through normalisation, the metric ensures that its

values fall within the range of 0 to 1, facilitating meaningful comparisons across different retrieval tasks and datasets. In evaluating model performance with query variations, nDCG shines by capturing subtle distinctions in ranking order, ultimately influencing user satisfaction.

- Other Common Metrics: Beyond nDCG, several other evaluation metrics serve distinct purposes in information retrieval evaluation [9]:
  - Accuracy: This metric measures the percentage of correctly classified instances out of all instances in the dataset.
  - Precision: This metric measures the proportion of true positives (correctly classified documents) versus all documents classified as positive.
  - Recall: This metric measures the proportion of true positives out of all positive documents.
  - fMeasure: This metric is the harmonic mean of recall and precision; it is commonly used when the dataset is imbalanced and to emphasise the importance of recall over precision.
  - MAP (Mean Average Precision): Unlike the above measures, accounts for all relevant document rankings by calculating the precision scores' mean across all queries.

2.4 Query Variations

more - add diagram.

Query variation refers to the different expressions or ways in which users can convey a specific information need or search query [4]. For example, a user might use different word order or sentence structure to express the same need. Consider the following two queries about a user's interest in baking a cake: "How to bake a cake from scratch?" "Homemade cake recipe?" Both queries express the same information need while having slight variations. Other examples of query variations may demonstrate specialisation, aspect change, misspelling, natural language, or paraphrasing [1].

Various techniques, such as query suggestion, reformulation, or expansion, can handle query variation in IR systems. Query expansion adds synonyms or extra terms to the original query to expand its scope and retrieve more relevant documents. Query reformulation is when the original query is augmented into a new one that reflects the information needed [10]. Addressing query variation is crucial in information retrieval since it can significantly impact the effectiveness of a search

system. By understanding and accounting for the diverse ways users express their information needs, search systems can retrieve more relevant documents and offer an improved user experience.

Chapter 3

Related Work

add more in-depth content and specific reference to gap/s.

This chapter delves into the existing body of literature that informs and supports the research presented in this thesis. It explores the domains of large language models, enhancing retrieval pipelines, and evaluating retrieval systems. Additionally, it delves into query variation, examining benchmark datasets, the impact on system performance, and various methods of generating query variations. This chapter highlights the current state of knowledge in the field and identifies gaps this research aims to fill.

Query variations can arise due to ambiguity in user search terms, different user intentions, or variations in the search context. By evaluating the existing literature on query variation generators and their place within LLMs, this review aims to identify the current state of research in this area and highlight any gaps or spaces for future research.

3.1 Information Retrieval

Recent studies focus on improving the performance and efficiency of LLMs in NLP tasks, particularly in text generation and understanding. One key trend in this field is using pre-trained models, such as GPT [2] and BERT [7], which achieve state-of-the-art results on various NLP tasks. Studies have also investigated the use of transfer learning techniques, such as fine-tuning [7] and domain adaptation [6], to improve the performance of LLMs in specific contexts. Future research is expected to explore language models and their use in more complex and diverse NLP tasks and develop new methods to improve the efficiency and scalability of these models [10, 11, 12].

3.1.1 Improvement of Retrieval Pipelines

Recent literature on retrieval pipelines has centred around improving the effectiveness and efficiency of handling large-scale datasets and diverse queries [1, 6, 13]. The main themes in this field of research include the use of deep learning techniques, including deep reinforcement learning and neural networks [1, 10, 14], the investigation of query expansion and query prediction techniques [4, 5, 13]. The development of novel evaluation metrics, such as diversity measures, to evaluate the relevance and novelty of search results [9, 13].

Several recent and related papers were mentioned in the original paper's introduction and related work sections [1]. One is "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by Devlin et al. [7], which introduced the BERT model that has achieved state-of-the-art results on various NLP tasks. BERT and GPT are pre-trained transformer models and can be used interchangeably in many NLP tasks. Therefore, future work could investigate the impact of using GPT instead of BERT in retrieval pipelines.

Overall, recent literature on retrieval pipelines aims to develop more effective and efficient methods for information retrieval that can enhance search quality and user satisfaction. Zendel et al.'s [4] work is a novel investigation into the relationship between query and information need, with their approach to query performance prediction (QPP) tasks outperforming the baseline. While research is abundant into improving these pipelines, there is still a way to go; there are studies that purposefully target the vulnerabilities to highlight weaknesses [14, 15]. Gao et al.'s [14] work in this field presents a novel framework designed to generate queries formulated to trick models into classifying them incorrectly. Their results show a drastic decrease in effectiveness.

3.1.2 Evaluation

Recent literature on evaluating LLMs has emphasised the need for more diverse and robust evaluation methods [1, 6, 10, 13]. Traditionally, LLMs have been evaluated using metrics such as accuracy, which measures the model's ability to predict text classification correctly. However, it has been theorised that in real-world applications, this metric may need to appropriately reflect the model's performance [9, 16]. Moffat et al. [13] introduced the C/W/L framework to better model actual user behaviour by considering the probability that a user will continue on the next while looking at a document. This study showed that by assessing an "expected goal of search", more accurate predictions of a user's need. Additionally, Ribeiro et al. [16] proposed a novel evaluation model that addresses the limitations of conventional evaluation measures and methods. Their experiments revealed previously

undetected shortcomings in NLP models that had undergone extensive testing.

3.2 Query Variations

3.2.1 Benchmark Datasets

Many recent studies have proved the importance of query variations in benchmark datasets [1, 4, 5]. This is important for several reasons: it makes the datasets more realistic and representative of the types of queries users use in real-world situations, including query variations ensures that benchmark datasets cover a wide range of queries and information needs, and it allows for better evaluation of the effectiveness of IR systems in handling variations in user queries [1, 17, 4]. Some datasets have been generated specifically to cater to these needs, such as UQV100 [18], TREC [19], and recently, DL-TYPO [17]. The study by Zhuang et al. specifically addressed the notion that current dense retrievers struggle to perform with unusual queries. Furthermore, as discovered by Penha et al., [1] following their investigation into retrieval pipelines robustness when using datasets with query variations, using these benchmark datasets can stimulate improvements in IR systems by highlighting areas of weaknesses. These studies have initiated the advancement of this field, and the study's experiments intend to carry on their research. Overall, the inclusion of query variations in benchmark datasets is critical for evaluating and enhancing the performance of IR systems in handling variations in user queries and information needs.

3.2.2 Effect on Performance

Research on query variation is crucial because it is well-established that such variations have a detrimental effect on the efficacy of LLMs. Many studies have shown that text variations cause under-performing IR systems [17, 3, 16]. The study by Penha et al. [1] modelled seven approaches to ranking with two datasets. Their findings showed that when faced with four different query variations, the effectiveness dropped an average of 20%. This proves the importance of further study and ways to improve the existing models.

Further, Lu et al.'s [5] study on relevance modelling with multiple queries shows that the utilisation of query variations substantially outperforms that of a single query. Their experiments involved performing fusion at the term, query, and document level, which is a new concept in this field and shows promising results. LLMs can better capture the nuances and complexities of user queries and information needs by incorporating query variations. This results in more accurate and relevant search results, as provided in several recent studies [1, 17, 4, 5]. Zuccon et al.'s [11]

study focused on examining the role of query variations in comparing system effectiveness and proposed a framework that explicitly incorporates query variations. Unlike similar studies, their analysis considers not only the mean efficacy of the system but also its variance across different query variations and topics. The findings reveal a significant impact of query variations on comparing other systems.

3.2.3 Methods of Generation

Recent research has explored methods to generate queries that better match user intent and handle variations such as data augmentation, fusion, query reformulation, and adversarial training [5, 10, 12]. Namely, Bailey et al. [10] prove that query formulation is critical to effectively studying query variability in LLM evaluation. Benham et al. [12] also investigate building query variations using fusion and weighted random sampling processes, making their retrieval effectiveness competitive with the state-of-the-art.

Overall, recent literature on query variation in IR highlights the critical role of datasets, the potential impact on performance, the promising approach of query generation, and the need for more appropriate evaluation metrics. Accounting for query variation is essential for developing effective systems that understand and respond to users' information needs.

Chapter 4

Methodology

This chapter serves as the blueprint for the empirical study conducted in this thesis. It outlines the critical components of the research, including datasets, ranking models, training procedures, query generators, and evaluation metrics. This chapter offers transparency into the experimental framework and how the research questions will be addressed by delineating the methodology employed. It lays the groundwork for comprehending the subsequent results and their implications.

4.1 Datasets

The original study leverages two distinct datasets to investigate the reproducibility and expansion of prior findings. The first dataset, TREC-DL-2019 [19], is a benchmark for passage retrieval tasks and comprises a test set featuring 43 diverse queries. The second dataset, ANTIQUE [20], focuses on non-factoid question answering and includes a test set with 200 distinct queries. These datasets, also employed in the original report, facilitate comparing and replicating findings, providing a consistent foundation for experimentation.

Incorporating both the datasets from the original study and introducing new ones for expansion was a reasonable approach to ensure the thoroughness and generalisability of the research findings. Using the datasets from the original study provided continuity and allowed for the replication of experimental conditions. This ensures that the reproducibility aspect of the research question can be rigorously examined, as any variations in results can be attributed to the models and methodologies rather than differences in datasets. On the other hand, introducing new datasets broadened the scope of the study and extended its applicability to a broader range of information retrieval scenarios. This expansion allows for exploring how the conclusions and observations drawn from the original datasets generalise to different contexts and data characteristics. By combining existing and novel datasets, the research

balances between validating previous findings and advancing our understanding of query variation impacts on retrieval pipelines in diverse settings.

The DL-TYPO dataset, introduced by Zuccon and Zhuang, is a valuable resource for investigating the robustness of retrieval systems when dealing with real-world query variations, precisely, queries containing typos. This dataset comprises human relevance judgments and feature pairs of queries, one with typos and one with those typos corrected. What sets DL-TYPO apart is that it contains queries mined from a substantial search engine query log, offering authentic representations of queries as they appear in practice and the produced misspellings. Relevance assessments are provided against passages from the MSMARCO dataset, an established resource for information retrieval research. Investigating the performance of retrieval models on this dataset allows for meaningful conclusions to be drawn about the effectiveness and robustness of such models in handling actual typos. Moreover, the statistical analysis and benchmarking conducted on this dataset, using standard evaluation metrics, provide a solid foundation for comparative research. The dataset topics and qrels were sourced from `https://github.com/ielab/CharacterBERT-DR`.

4.2 Ranking Models and Training

mention my used values + max iter.

The selection of models in the original study was driven by the need to comprehensively assess the robustness of retrieval pipelines when exposed to query variations. To achieve this goal, a diverse range of ranking models was chosen to represent different paradigms and approaches within the field of information retrieval.

- Traditional Models (Trad): BM25 [21] and RM3 [22] utilised default hyperparameters and the implementation provided by the PyTerrier toolkit [23].
- Neural Ranking Models (NN): Kernel-based ranking models KNRM [24] and CKNRM [25] were trained on the training sets of TREC-DL-2019 and ANTIQUE, employing default settings from the OpenNIR [26] implementation.
- Transformer-Based Models (TNN): BERT-based methods, including EPIC [27] and BERT [7], underwent fine-tuning using the bert-base-uncased model on the respective training datasets. T5 [28] models were implemented using the monoT5 [29] framework from the PyTerrier T5 plugin, incorporating pre-trained weights for MSMarco [30] by the original authors of monoT5.

Using the same models from the original study without introducing any new ones was a deliberate choice to ensure the research's reproducibility and comparability.

By maintaining consistency with the models used in the original study, this thesis focused primarily on reproducing the findings and expanding upon them through additional datasets, thus allowing for a more direct evaluation of the study's original observations. Furthermore, the original study had already carefully selected a diverse set of ranking models that represented various paradigms within information retrieval, making them well-suited for a reproducibility study and the subsequent expansion. This approach ensures that any differences or improvements observed in the new datasets can be more confidently attributed to variations in query generation and the nature of the datasets rather than introducing new ranking models. It also facilitates a direct comparison of the results with those of the original study, enabling a comprehensive assessment of the generalisability of the findings across different contexts.

4.3 Query Variations

4.3.1 Taxonomy

add refs and table expl.

The original study investigated query variations using the UQV100 [18] dataset, which contains query variations for 100 subtopics from the TREC 2013 and 2014 web tracks. From 365,000 query variation pairs, 100 pairs were randomly selected for manual annotation. During annotation, the authors categorised these pairs into six transformation categories, differentiating between those that changed query semantics and those that did not. Table 4.1 provides an example and description of each of the six types.

Table 4.1: Query Variation Taxonomy with examples from UQV100 Dataset.

Changes semantics	**Category**	**Definition**	**Examples from UQV100**		
Yes	Gen./specialization	Generalizes or specializes within the same information need.	american civil war	↔	number of battles in south carolina during civil war
	Aspect Change	Moves between related but different aspects within the same information need.	what types of spiders can bite you while gardening	↔	signs of spider bite
No	Misspelling	Adds or removes spelling errors.	raspberry pi	↔	raspeberry pi
	Naturality	Moves between keyword queries and natural language queries.	how does zinc relate to wilson's disease	↔	zinc wilson's disease
	Ordering	Changes the order of words.	carotid cavernous fistula treatment	↔	treatment carotid cavernous fistula
	Paraphrasing	Rephrases the query by modifying one or more words.	cures for a bald spot	↔	cures for baldness

An additional 550 query variation pairs were labelled to determine category distribution, with inter-annotator agreement found to be moderate. The study revealed that most query variations (57%) altered query syntax without affecting semantics. This finding led the study to focus on syntax-changing query variations, deferring exploration of query variation generators for future research.

4.3.2 Variation Generators

add refs and table expl, more detail in defn.

The study explored various methods for generating query variations within four syntax-changing categories. Initial investigations encompassed different query generator techniques for each category, followed by a filtering process. This filtering aimed to retain only those approaches capable of producing valid query variations while eliminating those with high correlation to one another. Ultimately, the research utilised ten distinct methods, each listed in Table 4.2 along with an illustrative example of transformation. Each method takes an input query and produces a query variation. Although most methods have the potential to generate multiple variations for a single input query, the study opted to use a single query variation per method to provide ample data for analysis.

Table 4.2: Outline of each of the ten variation generator methods used.

Category	**Method**	**Definition**
Misspelling	NeighbCharSwap	Swaps two neighbouring characters from a random query term.
	RandomCharSub	Replaces a random character from a random query term with a randomly chosen new ASCII character.
	QWERTYCharSub	Replaces a random character of a random query term with another character from the QWERTY keyboard
Naturality	RemoveStopWords	Removes all stopwords from the query.
	T5DescToTitle	Applies an encoder-decoder transformer model (T5) that is fine-tuned on the task of generating the title based on a description.
Ordering	RandomOrderSwap	Randomly swap two words of the query.
Paraphrasing	BackTranslation	Applies a translation method to the query to a new language (de) and back again (en).
	T5QQ	Applies an encoder-decoder transformer model (T5) that is fine-tuned on the task of generating a paraphrase question from the original question.
	WordEmbedSynSwap	Replaces a non-stop word by a synonym as defined by the nearest neighbour word in the embedding space.
	WordNetSynSwap	Replaces a non-stop word by a the first synonym found on WordNet.

For T5DescToTitle and T5QQP, pre-trained T5 models (t5-base) are fine-tuned using the Huggingface transformers library [31]. The facebook/m2m100_418M pre-

trained model from the transformers library is employed for BackTranslation. Other query variation methods use the TextAttack library [32].

4.4 Evaluation Metrics

The primary evaluation metric employed in this study was nDCG@10 (normalised Discounted Cumulative Gain at rank 10). The choice of nDCG as the evaluation metric in the original study stemmed from its suitability for information retrieval tasks, particularly in scenarios where ranking results play a critical role. It was used to measure the effectiveness of ranking models in retrieving relevant documents.

The exclusive use of nDCG as the evaluation metric in the thesis was a deliberate choice driven by several compelling reasons. Firstly, nDCG is a widely recognised and accepted metric in information retrieval, making it suitable for comparison with existing research and ensuring consistency in the evaluation process. Secondly, nDCG's ability to consider both document relevance and ranking position offers a holistic assessment of retrieval pipeline performance, aligning with the research objectives of evaluating the impact of query variations on the quality of ranked results. Using additional metrics could have introduced complexity and potentially conflicting results, making it harder to draw clear conclusions. Lastly, the focus on nDCG allowed for a more in-depth exploration of the nuances and specific effects of query variations on retrieval pipeline performance, providing a comprehensive understanding of the research questions. In essence, the decision to exclusively use nDCG was made to maintain methodological rigour, ensure meaningful and interpretable results, and facilitate a focused investigation into the reproducibility and generalisability of the original study's findings.

Chapter 5

Results & Discussion

This chapter provides a comprehensive account of the research outcomes, their significance, and the insights gained. The ensuing results and discussions are organised into three key sections. The first section focuses on generating query variations and assessing their validity, which directly contributes to addressing both research questions. The second and third section, categorised according to the relevant research question, delves into analysing the primary experiments conducted in the original study concerning model robustness. The first research question delves into reproducing the original study's main experiments, scrutinising the robustness of retrieval pipelines to query variations. The second research question extends the investigation by exploring the effects of query variations on an additional dataset, DL-TYPO. This structured approach allows for a comprehensive exploration of the impact of query variations on retrieval pipelines. It offers valuable insights into the challenges and opportunities presented by this dynamic aspect of information retrieval.

5.1 Generator Quality

Given the automatic nature of query generation methods, the research assesses the quality of the generated query variations. The authors conducted manual annotations for 1,371 pairs of {original query, developed query variation} from the test sets of TREC-DL-2019 and ANTIQUE. This quality assessment ensures that only valid query variations, aligning with the intended query variation categories, are considered in the experiments, enhancing the reliability of the study's findings. For consistency with the original study, each ANTIQUE data set pair was re-annotated and compared to that of the original authors to understand their judgement.

This method was closely followed in this study when analysing the variations generated for the DL-TYPO dataset. This dataset comprises 60 queries, each subjected to query variation generation. The methods used for query variation gen-

eration remained consistent with those employed in the original study, albeit with slight adjustments because it is not sourced from IR_datasets. The process can be summarised as follows:

1. **Query Variations Generation:** Following the methodology used in the original study, a query variation, q^, was created using each method, M, for each query, q.

2. **Variation Editing:** Punctuation and capital letters were consistently removed from query variations, aligning with the procedures outlined in the original study.

3. **Automatic Annotation:** All variations originating from misspelling and ordering were automatically designated as valid, as these transformations are considered rule-based and retain the same query semantics. Additionally, any transformations resulting in a query identical to the input query (q^ = M(q) = q) were automatically deemed invalid.

4. **Manual Annotation:** For the remaining 320 query-variation pairs, manual annotation was performed based on criteria similar to those employed in the original study. These criteria encompassed two primary considerations: (I) Whether q^ maintained the same semantics as q and (II) Whether the syntax difference between q and q^ could be attributed to category C.

Furthermore, observations from DL-TYPO variation annotations indicated several similarities to the original study, as well as novel insights:

5.1.1 Parallel Findings (RQ1):

(I) **T5DescToTitle Effects:** The T5DescToTitle method occasionally resulted in the removal of essential query terms, thereby altering query semantics (e.g. "things to do when broed" to "broeding" (T5DescToTitle), "how money does the show seinfled make" to "seinfled money" (T5DescToTitle)) (total of 5 occurrences).

(II) **Replicating Identical Queries:** The BackTranslation and T5QQP methods were found to generate identical copies of the input query, which were automatically labelled as invalid (a total of 56 occurrences).

(III) **Inaccurate Synonym Replacements:** Transformations that replaced words with presumed synonyms, such as WordEmbedSynSwap and WordNetSynSwap, sometimes introduced words that were not synonymous in the query

context (e.g. "how do i clean my computor monitor screen" to "how do i clean my computor supervises screen" (WordEmbedSynSwap), "what kind of medicine is zytec" to "what kind of music is zytec" (WordNetSynSwap)). (total of 24 occurrences).

5.1.2 Novel Findings (RQ2):

(IV) **Proper Noun Alterations:** An expansion of the previous observation revealed that transformations occasionally replaced proper nouns or titles, causing them to no longer refer to the correct entity (e.g. "facts about chirs brown" to "facts about chirs chocolate-brown" (WordNetSynSwap), "singer axel rose" to "singer axel soars" (WordEmbedSynSwap)) (total of 11 occurrences).

(V) **Substituting Non-English Words:** WordEmbedSynSwap replaced a word with its equivalent in another language (e.g. "flee market buildings" to "flee mercado buildings" (WordEmbedSynSwap)). This does not occur in the other datasets (total of 1 occurrence).

(VI) **Misspelling Compounding:** Synonym and paraphrasing-related methods were observed to compound misspellings, generating queries with different semantics, despite the method functioning correctly (e.g. "medal taste in mouth symptoms" to "decoration taste in mouth symptoms" (WordNetSynSwap), "flee market buildings" to "the escape market building" (BackTranslation)) (total of 14 occurrences).

(VII) **Fixing Spelling Errors:** Interestingly, five of the ten query variation methods were identified as fixing the original spelling errors (e.g. "alchol and drug rehab" to "alcohol and drug rehabilitation" (BackTranslation), "car accident lawers" to "what are car accident lawyers" (T5QQP), "los angelel unified school district" to "los angeles unified school" (T5DescToTitle)). BackTranslation was the method most often responsible for this correction (occurring 14 times), as well as T5QQP (7 times) (total of 25 occurrences).

possibly bar - show pos and neg

Upon rigorous manual annotation, 477 valid query-variation DL-TYPO pairs were identified and retained for further analysis. These pairs were scrutinised based on their semantic integrity and alignment with the original query, offering insights into the effectiveness and challenges posed by various query variation methods. The results of this annotation process for all three datasets, as summarised in Table 5.1, provide valuable insights into the effectiveness of ten query variation methods across the three distinct datasets. RandomOrderSwap, NeighbCharSwap, and QWERTYCharSub consistently produced 100% valid variations for the DL-TYPO and

Table 5.1: Total number of valid query variations and percentage of valid variations compared to the total queries for dataset and generation method, categorised by the category.

Method	**DL-TYPO**	**ANTIQUE**	**TREC-DL-2019**
NeighbCharSwap	60 (100%)	199 (100%)	43 (100%)
RandomCharSub	59 (98%)	197 (98%)	42 (98%)
QWERTYCharSub	60 (100%)	182 (91%)	42 (98%)
RemoveStopWords	40 (67%)	199 (100%)	37 (86%)
T5DescToTitle	42 (70%)	136 (68%)	35 (81%)
RandomOrderSwap	60 (100%)	200 (100%)	43 (100%)
BackTranslation	34 (57%)	93 (46%)	23 (53%)
T5QQP	52 (87%)	105 (52%)	26 (60%)
WordEmbedSynSwap	41 (68%)	124 (62%)	27 (63%)
WordNetSynSwap	29 (48%)	71 (36%)	16 (37%)
Total	**477 (80%)**	**1506 (75%)**	**334 (78%)**

ANTIQUE datasets, demonstrating their robustness in creating semantically sound queries. In the TREC-DL-2019 dataset, these methods also performed well, with QWERTYCharSub achieving the lowest but still notable percentage of 97.7%. RandomCharSub achieved high percentages of valid variations, reaching 98.3% for the DL-TYPO dataset and 98.5% for ANTIQUE, while slightly lagging for the TREC-DL-2019 dataset with 97.7%. RemoveStopWords showed variable results, with a lower 66.7% success rate for the DL-TYPO dataset. However, it consistently attained a high 99.5% for the ANTIQUE dataset and a respectable 86.0% for the TREC-DL-2019 dataset. T5DescToTitle performed moderately, achieving a 70% success rate for the DL-TYPO dataset, 68% for ANTIQUE, and a more promising 81.4% for the TREC-DL-2019 dataset. WordEmbedSynSwap, T5QQP, and BackTranslation exhibited varying percentages of valid variations across datasets, with WordEmbedSynSwap and T5QQP showing moderate success rates and BackTranslation achieving slightly lower percentages. Conversely, WordNetSynSwap consistently produced the lowest percentage of valid variations across all datasets, indicating significant challenges in generating semantically valid query variations using this method. It is essential to acknowledge that variations in the validity of DL-TYPO annotations can be attributed, to some extent, to the change in the annotator's identity. The consistency and reliability of these annotations could be enhanced by adopting the original study's practice of engaging multiple annotators for this task. Multiple annotators mitigate the potential influence of individual annotator biases and contribute to a more comprehensive and well-rounded assessment of the dataset's quality and reliability. This approach aligns with established best prac-

tices in dataset curation, ensuring that the dataset remains a robust and valuable resource for research.

5.2 Robustness to Query Variations

5.2.1 Research Question 1: Reproduction

A comprehensive set of experiments was conducted using the same datasets, models, and methodologies to evaluate the robustness of retrieval pipelines to query variations and investigate the extent to which the original study's findings can be successfully reproduced. This section presents the key results of the reproduction study, structured to mirror the organisation of the original study's results. The

Table 5.2: Effectiveness (nDCG@10) of each method and model for ANTIQUE when faced with different query variations. Bold indicates the highest values observed for each model. Subscripts, ↓/↑, signify statistically significant decreases/increases obtained through a two-sided paired Student's T-Test conducted at a 95% confidence level when comparing the model's performance with the original queries.

Category	**Variation Method**	**BM25**	**RM3**	**KNRM**	**CKNRM**	**EPIC**	**BERT**	**T5**
-	Original Query	**0.2286**	**0.217**	0.2181	0.2065	0.266	**0.363**	**0.3333**
Misspelling	NeighbCharSwap	0.1559↓	0.1469↓	0.159↓	0.1444↓	0.184↓	0.2608↓	0.2509↓
Misspelling	QWERTYCharSub	0.1613↓	0.1525↓	0.162↓	0.1553↓	0.192↓	0.2619↓	0.2652↓
Misspelling	RandomCharSub	0.1623↓	0.1593↓	0.1602↓	0.1476↓	0.1879↓	0.2563↓	0.2458↓
Naturality	RemoveStopWords	0.227	0.2161	**0.2232**	**0.2153**	**0.2693**	0.3391↓	0.32
Naturality	T5DescToTitle	0.1673↓	0.1646↓	0.1647↓	0.1672↓	0.2006↓	0.2402↓	0.2393↓
Ordering	RandomOrderSwap	**0.2286**	0.2169	0.2181	0.1978	0.2661	0.3566	0.3255
Paraphrase	BackTranslation	0.1618↓	0.1546↓	0.1609↓	0.1438↓	0.2032↓	0.274↓	0.2581↓
Paraphrase	T5QQP	0.2201	0.2063	0.2085	0.1957	0.2617	0.3389↓	0.3214
Paraphrase	WordEmbedSynSwap	0.1759↓	0.1715↓	0.1915↓	0.1689↓	0.2139↓	0.2809↓	0.2814↓
Paraphrase	WordNetSynSwap	0.1791↓	0.175↓	0.1933↓	0.1763↓	0.212↓	0.2829↓	0.2734↓

first objective was to assess the robustness of different ranking models to query variations, categorised into lexical traditional models (Trad), neural ranking models (NN), and transformer-based language models (TNN). The aim was to determine the effectiveness of these models when original queries were replaced with their corresponding query variations. The results are outlined in Table 5.2 and Table 5.3 for the ANTIQUE and TREC-DL-2019 datasets, respectively. These tables show the resulting nDCG@10 value for each method and model, grouped by their variation category.

The findings exhibit a substantial degree of alignment with the original study. The conducted experiments, encompassing various query variations and model combinations, consistently reveal a statistically significant decline in retrieval effectiveness. For the TREC-DL-2019 dataset, this reduction is observed in 41 out of 70

Table 5.3: Effectiveness (nDCG@10) of each method and model for TREC-DL-2019 when faced with different query variations. Bold indicates the highest values observed for each model. Subscripts, ↓/↑, signify statistically significant decreases/increases obtained through a two-sided paired Student's T-Test conducted at a 95% confidence level when comparing the model's performance with the original queries.

Category	**Variation Method**	**BM25**	**RM3**	**KNRM**	**CKNRM**	**EPIC**	**BERT**	**T5**
-	Original Query	**0.4795**	**0.5156**	**0.4941**	**0.4931**	**0.624**	**0.6358**	0.6998
Misspelling	NeighbCharSwap	0.2747↓	0.2748↓	0.3078↓	0.308↓	0.3893↓	0.3812↓	0.4944↓
Misspelling	QWERTYCharSub	0.2435↓	0.2504↓	0.2563↓	0.2965↓	0.3496↓	0.3657↓	0.4461↓
Misspelling	RandomCharSub	0.2314↓	0.2347↓	0.2349↓	0.2263↓	0.295↓	0.3075↓	0.3963↓
Naturality	RemoveStopWords	0.4778	0.5113	0.4769	0.4756	0.6214	0.615	0.6862
Naturality	T5DescToTitle	0.4215	0.4344	0.3693↓	0.3928	0.5061↓	0.533↓	0.5717↓
Ordering	RandomOrderSwap	**0.4795**	**0.5156**	**0.4941**	0.4708	0.6227	0.6268	0.697
Paraphrase	BackTranslation	0.3964	0.4195	0.3954	0.3605↓	0.5301	0.4874↓	0.6058
Paraphrase	T5QQP	0.4722	0.5043	0.4525	0.4609	0.604	0.6222	**0.7045**
Paraphrase	WordNetSynSwap	0.3488↓	0.365↓	0.3615↓	0.3605↓	0.449↓	0.459↓	0.5457↓
Paraphrase	WordEmbedSynSwap	0.353↓	0.3539↓	0.3767↓	0.368↓	0.4749↓	0.4816↓	0.5603↓

instances, representing a slight deviation from the original study's results by 8 cases. Similarly, for the ANTIQUE dataset, a decrease in effectiveness is recorded in 51 instances, deviating by merely 3 cases from the original study's outcomes. Furthermore, the analysis of valid queries illustrates that, on average, the models experience a reduction in effectiveness of 22.02% for TREC-DL-2019, a figure marginally distinct from the original study's 20.62%. For ANTIQUE, the average decline amounts to 17.92%, contrasting with the original study's 19.21%. These consistent trends reaffirm the original study's pivotal assertion that retrieval pipelines exhibit limited robustness in the face of query variations. These outcomes substantiate previous evidence suggesting that query variations introduce considerable variability into diverse information retrieval systems.

Discrepancies emerge when comparing specific values between this study and the original research. On average, the discrepancies are minor, less than 0.08, particularly evident in the results for the BERT model. One potential explanation for these differences is the value of training iterations. Although the exact training iteration value used in the original study remains undisclosed, this reproduction study had to make do with a value of 50 due to computational constraints. Notably, these minor deviations in results underline the intricate nature of replication work, where subtle variations in experimental settings, environmental conditions, or data pre-processing procedures can lead to distinctions in experimental outcomes.

The replication study effectively addresses Research Question 1 (RQ1). Through a meticulous replication of experiments and methodologies from the original study, this reproduction study has offered critical insights and fortified the original findings.

The consistently corroborated results from the reproduction study underscore the limited robustness of retrieval pipelines in query variations, consequently reinforcing the pivotal conclusion of the original study. This coherence in findings between the two independent studies signifies the robustness and broader applicability of the original study's observations. This outcome robustly attests to the validity of the research outcomes in the original study's context and their wider relevance, accentuating the significant role of query variations in evaluating retrieval systems.

5.2.2 Research Question 2: Expansion

The outcomes of this extended study provide a broader perspective on the conclusions derived from the original study. To assess the generalisability of the original findings, the study replicated the experiments using the additional DL-TYPO dataset. Furthermore, in line with the original research's assertion regarding the diminished robustness of neural ranking models to query variations, even in the context of contemporary collections, DL-TYPO emerged as an apt candidate for testing the transference of this inquiry. The selection of DL-TYPO was motivated by its novel and unexplored nature, thereby substantiating its suitability for broadening the research endeavour.

validity of new compared to old - newest plot - maybe pie or bar

In alignment with the original study's focus on the TREC-DL-2019 and ANTIQUE datasets, this study applied the same models and methodologies to the DL-TYPO dataset. The primary objective was to ascertain if the conclusions regarding retrieval pipeline robustness held when confronted with this real query dataset. The results are outlined in Table 5.4 for the DL-TYPO dataset. This table shows the resulting nDCG@10 value for each method and model, grouped by their variation category.

As anticipated, the results closely paralleled the original research, albeit with a discernibly diminished effectiveness. A pronounced reduction in the retrieval models' performance manifested across most query variation scenarios, thus substantiating the initial investigation's findings. Notably, the extent of the overall effectiveness decline, averaging at 31.81%, surpassed that observed in the context of TREC-DL-2019 and, notably, ANTIQUE. This observation underscores the resilience of the conclusions derived from the original study when transposed to the novel and distinct DL-TYPO dataset. The research's core inferences and insights possess a certain degree of generalisability, albeit with the caveat of nuanced dataset-specific variations in the extent of the observed effects.

The variation in results between the DL-TYPO dataset and the previously studied datasets can be attributed to several underlying factors. Firstly, the DL-TYPO dataset inherently features more query variations (misspellings, specifically) than the

Table 5.4: Effectiveness (nDCG@10) of each method and model for DL-TYPO when faced with different query variations. Bold indicates the highest value observed for each model. Subscripts, ↓/↑, signify statistically significant decreases/increases obtained through a two-sided paired Student's T-Test conducted at a 95% confidence level when comparing the model's performance with the original queries.

Category	**Variation Method**	**BM25**	**RM3**	**KNRM**	**CKNRM**	**EPIC**	**BERT**	**T5**
-	Original Query	0.1911	0.1757	**0.2117**	**0.1798**	0.1862	0.1951↓	0.2882
Misspelling	NeighbCharSwap	0.1145	0.1086	0.1224↓	0.0961↓	0.0987↓	0.0865↓	0.1922
Misspelling	QWERTYCharSub	0.0772↓	0.0746↓	0.0813↓	0.0572↓	0.0707↓	0.0683↓	0.1139↓
Misspelling	RandomCharSub	0.068↓	0.0626↓	0.1083↓	0.079↓	0.0844↓	0.0733↓	0.148↓
Naturality	RemoveStopWords	0.1911	0.1757	0.2022	0.1666	0.1825	0.1762	0.2794
Naturality	T5DescToTitle	0.1347	0.1244	0.1351↓	0.1228	0.1434	0.1454	0.2055
Ordering	RandomOrderSwap	0.1911	0.1757	**0.2117**	0.1443	0.1875↓	0.1719	0.2707
Paraphrase	BackTranslation	0.1468	0.141	0.1578	0.121	0.1702	0.1625	0.2245
Paraphrase	T5QQP	**0.2138**	**0.2052**	0.1738	0.1546	**0.1956**	**0.218**	**0.2897**
Paraphrase	WordEmbedSynSwap	0.0809	0.075	0.106↓	0.0901↓	0.0976↓	0.0638↓	0.1326↓
Paraphrase	WordNetSynSwap	0.1097↓	0.0978↓	0.1255↓	0.1004↓	0.1033↓	0.1018↓	0.1726↓

synthetically generated counterparts. The diversity and unpredictability of these authentic query variations could pose a more substantial challenge to the robustness of retrieval pipelines, leading to a potentially higher decline in effectiveness. Furthermore, the specific characteristics of the queries in the DL-TYPO dataset, such as the frequency and types of typos, differ significantly from those in the other datasets. These idiosyncrasies interact with the retrieval models and the query variations in a distinct manner, further contributing to variations in observed effectiveness. Additionally mentioned in the original study, dataset-specific properties, such as query lengths, language styles, and topical diversity, play a role in the differing outcomes. Lastly, the performance of retrieval models can also be influenced by the dataset's composition in terms of judged and unjudged documents. If the query variations in the DL-TYPO dataset significantly increase unjudged documents' ranking highly, this could impact the perceived effectiveness, even if some of these documents might be contextually relevant.

5.2.3 Robustness by Query Variation Category

Mirroring the methodology employed in the original study, an in-depth investigation was undertaken to scrutinise the influence of different query variation categories on model effectiveness. These variations were systematically classified into four distinct groups: misspellings, paraphrasing, naturality, and ordering. The depicted Fig. 5.2 illustrates the distribution of nDCG@10 Δ values across all models and variation categories, and these findings are corroborated by the mean nDCG@10 Δ values per category and dataset, delineated in Fig. 5.1. The findings accentuate that,

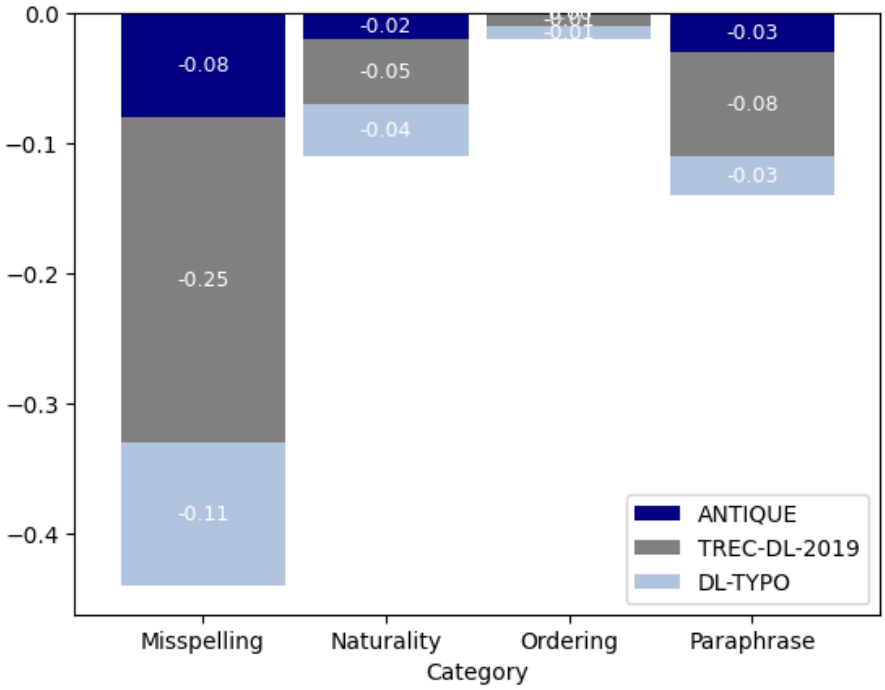

Figure 5.1: Average nDCG@10 Δ values per variation category for each dataset.

on average, misspelling variations yielded the most pronounced negative impact, registering nDCG@10 Δ values of -0.08 for ANTIQUE, -0.25 for TREC-DL-2019, and -0.11 for DL-TYPO. These results are consistent with the observations made in the original study, lending further credence to the notion that misspelling variations consistently undermine retrieval effectiveness. In contrast, variations belonging to the paraphrasing and naturality categories exhibited more modest nDCG@10 Δ values, with select queries manifesting a positive effect, mitigating the overall dip in retrieval performance. Notably, ordering variations had a negligible impact on traditional models, given their inherent nature as bag-of-words models, culminating in an average nDCG@10 Δ close to zero. This comprehensive assessment elucidates the divergent impacts of distinct query variation categories on retrieval model efficacy. DL-TYPO's consistently lower performance than ANTIQUE and TREC-DL-2019 further confirms this finding that misspellings lead to worse retrieval effectiveness. This can be seen when comparing original query values in Table 5.4 versus Table 5.2 and Table 5.3.

Having a closer look at the specific methods within each category, Table **??** provides an overview of how the methods affect nDCG@10 Δ when applied to an example query from DL-TYPO. Notably, the T5DescToTitle method exhibits a significant 75% reduction in nDCG@10. Other methods like NeighbCharSwap, RandomCharSub, and QWERTYCharSub result in roughly 49% decreases in nDCG@10, highlighting their detrimental influence on query variations. This analysis underscores

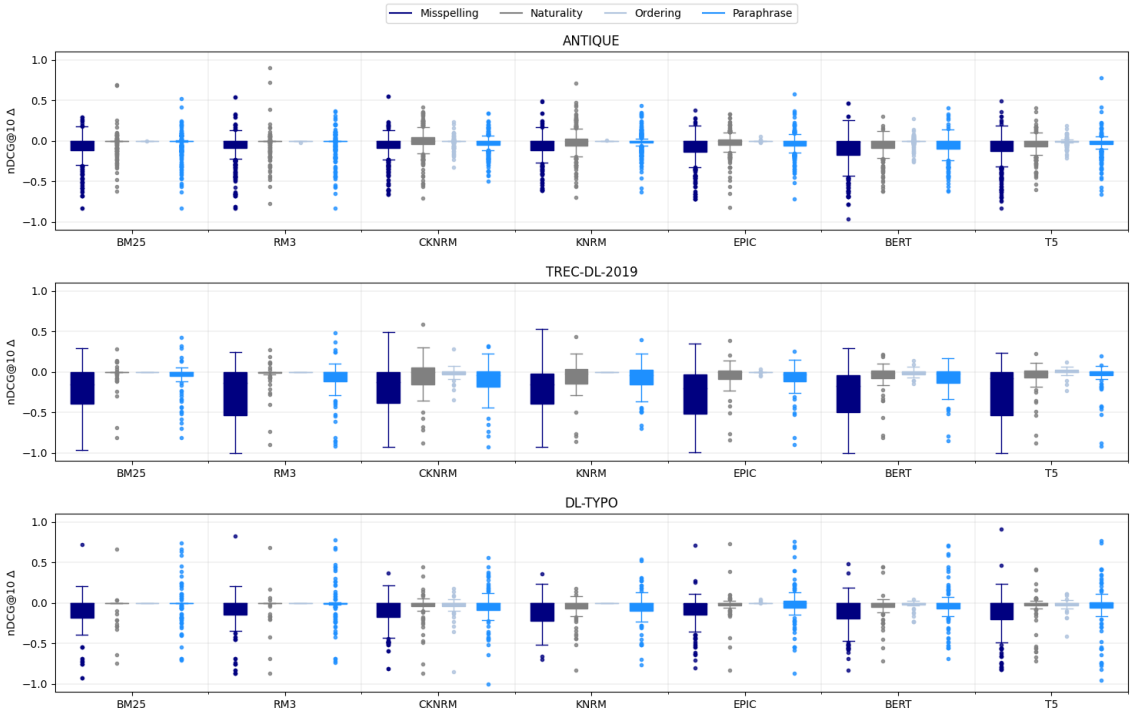

Figure 5.2: Distribution of nDCG@10 Δ when replacing the original query by the methods of each category.

the need to carefully choose variation methods, as their effects can vary significantly, impacting the quality of retrieval results.

5.3 Fusing Query Variations

Although the occurrences are fewer, Fig. 5.2 shows that in some instances, the query variation performs better than its original query (a positive nDCG@10 Δ). As per the original study, this motivates the investigation of fusing various variation types.

Table 5.5, Table 5.6, and Table 5.7 outline the resulting effectiveness when combining the rankings of each variation category. Each row labelled as RFC_C indicates that the results obtained from the query variations obtained after applying M_C methods using the Reciprocal Rank Fusion (RRF) method are fused, and RFC_{All} fuses the results obtained by all query variation methods $RFC_{Ordering}$ is excluded as it only has one method.

5.3.1 Research Question 1: Reproduction

As evidenced by the best queries, incorporating RRF yields substantial improvements in nDCG@10. Furthermore, a comparative analysis with Table 5.2 and Table 5.3 suggests that in numerous instances, the outcomes are either equivalent or

fix bert

fix bert, epic, ck-nrm

Table 5.5: Effectiveness (nDCG@10) of different methods for ANTIQUE when employing rank fusion (RRF) of the rankings obtained by using different sets of queries.

	BM25	**RM3**	**KNRM**	**CKNRM**	**EPIC**	**BERT**	**T5**
Original Query	0.2286	0.217	0.2182	0.2065	0.266	0.3947	0.3333
$RRF_{Misspelling}$	0.1714	0.1653	0.1804	0.1664	0.2065	0.2675	0.2441
$RRF_{Naturality}$	0.1842	0.1865	0.2058	0.2039	0.2407	0.3002	0.271
$RRF_{Paraphrase}$	0.1924	0.1861	0.1955	0.1765	0.2401	0.3223	0.2894
$RRF_{Synonym}$	0.1831	0.177	0.2009	0.1847	0.2184	0.2944	0.2678
RRF_{All}	0.2076	0.206	0.219	0.2121	0.2547	0.3197	0.2867
Best Query	0.4149	0.2716	0.2836	0.3369	0.3019	0.2681	0.3911

Table 5.6: Effectiveness (nDCG@10) of different methods for TREC-DL-2019 when employing rank fusion (RRF) of the rankings obtained by using different sets of queries.

	BM25	**RM3**	**KNRM**	**CKNRM**	**EPIC**	**BERT**	**T5**
Original Query	0.4795	0.5156	0.4941	0.4931	0.624	0.6358	0.6998
$RRF_{Misspelling}$	0.3035	0.3073	0.3175	0.3175	0.3838	0.3941	0.4636
$RRF_{Naturality}$	0.4744	0.4972	0.4668	0.464	0.5925	0.6005	0.6643
$RRF_{Paraphrase}$	0.4742	0.4865	0.4883	0.433	0.5847	0.577	0.6616
$RRF_{Synonym}$	0.4247	0.4066	0.4117	0.4048	0.4975	0.4902	0.5632
RRF_{All}	0.4752	0.4958	0.4965	0.4951	0.5908	0.5939	0.6442
Best Query	0.6964	0.5401	0.6116	0.6983	0.5939	0.5784	0.7598

superior to those achieved solely by employing query variations, occasionally rivalling the performance associated with the unaltered original query. This pattern of results aligns with the original study's findings, reinforcing the notion that RRF can enhance retrieval effectiveness. However, it is essential to note that while RRF augmentation demonstrates the potential for improvement, it does not consistently surpass the retrieval outcomes associated with the unaltered original queries. This observation underscores the nuanced nature of the impact of RRF in the context of query variations and retrieval performance.

5.3.2 Research Question 2: Expansion

Table 5.7 presents the effectiveness (nDCG@10) results for the DL-TYPO dataset when employing RRF. It is evident that, compared to the results for the ANTIQUE and TREC-DL-2019 datasets presented in Table 5.6 and Table 5.7, the effectiveness in DL-TYPO is notably lower. The original queries in DL-TYPO demonstrate lower nDCG@10 values than the other datasets, indicating a more challenging retrieval scenario. The impact of query variations and subsequent fusion using RRF

Table 5.7: Effectiveness (nDCG@10) of different methods for DL-TYPO when employing rank fusion (RRF) of the rankings obtained by using different sets of queries.

	BM25	**RM3**	**KNRM**	**CKNRM**	**EPIC**	**BERT**	**T5**
Original Query	0.1911	0.1757	0.1994	0.1798	0.1862	0.1951	0.2882
$RRF_{Misspelling}$	0.0979	0.0933	0.1147	0.0855	0.1006	0.0782	0.1373
$RRF_{Naturality}$	0.1657	0.154	0.1592	0.1498	0.161	0.1557	0.2277
$RRF_{Paraphrase}$	0.1965	0.1919	0.1839	0.1722	0.202	0.1931	0.2479
$RRF_{Synonym}$	0.0883	0.0795	0.1214	0.0945	0.1051	0.0932	0.1477
RRF_{All}	0.1546	0.1516	0.184	0.159	0.1656	0.156	0.2098
Best Query	0.3156	0.3024	0.275	0.317	0.3052	0.2913	0.421

is consistent with the results in ANTIQUE and TREC-DL-2019, as it demonstrates improvements over the original queries. However, the magnitude of improvements is smaller in DL-TYPO, underscoring the dataset's distinct characteristics and the more incredible difficulty in achieving substantial enhancements through query variations and RRF.

This discrepancy is due to the DL-TYPO dataset presenting more challenging queries with typos, making it inherently more complex to achieve high retrieval effectiveness. Consequently, the improvements resulting from query variations and RRF are less pronounced than in the ANTIQUE and TREC-DL-2019 datasets. These variations in results emphasise the dataset-specific nature of the research and the importance of considering different datasets' unique characteristics and challenges when investigating query variations and retrieval performance.

5.3.3 Query Variations Performance

Fig. 5.3 demonstrates that the naturality and paraphrase methods improve the variation effectiveness over that of the original more so than the other methods. The paraphrase category shows the highest number of variations with improved effectiveness across all models and datasets. This suggests that paraphrasing query variations have the potential to outperform the original queries. The counts range from 13 to 19, indicating substantial variability in successful paraphrased queries across models and datasets. This is due to paraphrasing methods often improving queries by fixing spelling errors (e.g. See 5.1.2) and replacing words with better synonyms (e.g. "real gohst pics" to "authentic gohst pics" (WordEmbedSynSwap)).

In contrast, the ordering category demonstrates limited variations with improved effectiveness. Some models (such as Trad and NN) on specific datasets (TREC-DL-2019 and ANTIQUE) have no variations that surpass the original query. This is attributed to the fact that changes in word order have little impact on retrieval

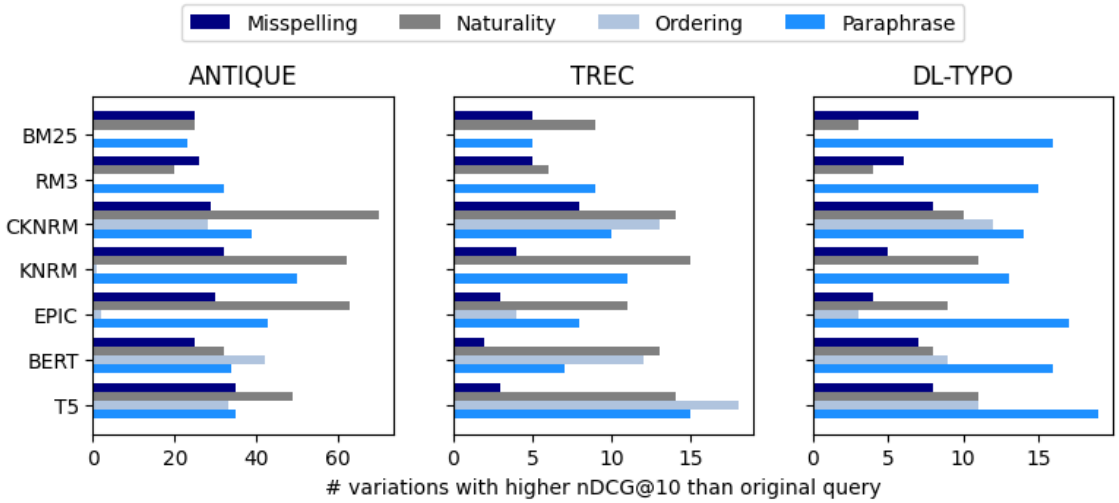

Figure 5.3: Number of variations with higher effectiveness (nDCG@10) than original query per model for each variation category.

effectiveness, specifically in traditional bag-of-words models. Overall, the data suggests that the effectiveness of query variations is highly dependent on the specific category of variation and the model dataset combination. Paraphrasing variations tend to show the most promise in improving retrieval performance, while ordering variations appear to have a limited impact. The analysis emphasises the importance of considering query variations' specific nature and interactions with different retrieval models and datasets.

The data highlights notable differences in the effectiveness of query variations between the DL-TYPO dataset and the other datasets, ANTIQUE and TREC-DL-2019. DL-TYPO has a lower occurrence of variations with higher effectiveness (nDCG@10) than the original query compared to the different datasets. This suggests that query variations may be less effective in DL-TYPO. The variability in the number of improvements is also evident across query variation categories. The variability is also model-dependent in DL-TYPO, where some models exhibit more successful query variations than others. These differences indicate the unique challenges posed by the real-world queries with typos from the DL-TYPO dataset compared to the synthetic data in the other datasets. These observations underscore the significance of dataset-specific characteristics when working with query variations and retrieval models and emphasise the need for further investigation to comprehend the distinct challenges presented by diverse datasets.

Chapter 6

Conclusion

6.1 Summary and Conclusions

This thesis undertook a comprehensive journey to explore query variations' impact on retrieval pipelines' robustness. The overarching objective drove the research to understand how query variations influence the retrieval effectiveness of diverse information retrieval systems. The study generated valuable insights and outcomes that collectively address the core research questions by extensively exploring various datasets, models, and query variations. Here, a synthesis of what has been accomplished, the pivotal conclusions drawn from the results, and reflections on potential avenues for future research are provided.

In this study, two primary research questions guided the investigation—the first research question aimed to reproduce an original study's experiments, methodologies, and findings. The primary focus was to investigate the robustness of retrieval pipelines to query variations, explicitly emphasising the effect of query variations on different ranking models. The results largely align with the original study's findings, reaffirming the lack of robustness of retrieval pipelines to query variations. Overall, the successful reproduction of the original study's results enhances the credibility of the conclusions drawn regarding the impact of query variations on retrieval pipelines and provides a strong foundation for further exploration in this critical area of information retrieval research.

The second research question, regarding expansion, delved into the broader perspective on the original study's conclusions by repeating the work on a new and unexplored dataset, DL-TYPO. The outcomes not only upheld the original findings but also revealed the dataset's unique characteristics, accentuating the need for careful consideration of dataset-specific attributes when assessing the impact of query variations. With compelling consistency across multiple datasets and models, the original study's findings were validated, emphasising the integral role of query

variations in evaluating retrieval systems. To gain a nuanced understanding of the influence of different query variation categories, variations were systematically categorised into misspellings, paraphrasing, naturality, and word order, uncovering that misspelling variations consistently had the most detrimental impact on retrieval effectiveness, further corroborating the significance of the original study's insights.

Additionally, the robustness of different model categories was explored, showing that models within the same category exhibit similar behaviour when confronted with query variations. A preference among transformer-based language models for natural language queries was identified in this context. In contrast, word order had a minimal impact on these models, aligning with recent research trends.

To enhance the retrieval outcomes influenced by query variations, rank fusion techniques were experimented with, highlighting the potential for query variation methods to augment retrieval effectiveness significantly. The results were consistent with the original study. Combining query variations using RRF generally mitigated the decreases in retrieval effectiveness observed when using query variations individually. However, it was noteworthy that RRF did not consistently outperform using the original query. This reaffirms that while rank fusion techniques can improve retrieval performance under certain circumstances, they need to prove to be a solution for the challenges posed by query variations.

6.2 Possible Future Work

The findings of this research signify that query variations indeed wield a substantial impact on retrieval pipeline robustness. These findings extend the realm of information retrieval, revealing the intricacies of query variations in different datasets, models, and variation category contexts. Future research in this domain may focus on several promising directions:

- Optimising Query Variation Methods: Investigation of advanced techniques for generating query variations more tailored to specific information retrieval systems, thereby mitigating the negative impact of specific variations.
- Leveraging Neural Models: Further exploration of the effectiveness of neural models for handling query variations and developing more robust deep learning architectures for this task.
- Real-world Applications: Extension of this research to explore the practical implications of query variations in real-world information retrieval systems and user experiences.

In conclusion, this study adds significant insights to the information retrieval domain, emphasising the importance of query variations and their substantial impact on retrieval pipeline robustness. As the field evolves, these findings will guide researchers, system developers, and practitioners in crafting more effective and adaptive information retrieval systems. The journey of understanding query variations has only just begun, and the path forward promises further exploration and discovery in the ever-evolving landscape of information retrieval.

Bibliography

[1] G. Penha, A. Câmara, and C. Hauff, "Evaluating the robustness of retrieval pipelines with query variation generators," in *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, pp. 397–412, Springer, 2022.

[2] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[3] X. Chen, B. He, K. Hui, L. Sun, and Y. Sun, "Dealing with textual noise for robust and effective bert re-ranking," *Information Processing & Management*, vol. 60, no. 1, p. 103135, 2023.

[4] O. Zendel, A. Shtok, F. Raiber, O. Kurland, and J. S. Culpepper, "Information needs, queries, and query performance prediction," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 395–404, 2019.

[5] X. Lu, O. Kurland, J. S. Culpepper, N. Craswell, and O. Rom, "Relevance modeling with multiple query variations," in *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 27–34, 2019.

[6] C. Wu, R. Zhang, J. Guo, Y. Fan, and X. Cheng, "Are neural ranking models robust?," *ACM Transactions on Information Systems*, vol. 41, no. 2, pp. 1–36, 2022.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] K. Chowdhary and K. Chowdhary, "Natural language processing," *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[9] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," in *AI*

2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings 19, pp. 1015–1021, Springer, 2006.

[10] P. Bailey, A. Moffat, F. Scholer, and P. Thomas, "User variability and ir system evaluation," in *Proceedings of The 38th International ACM SIGIR conference on research and development in Information Retrieval*, pp. 625–634, 2015.

[11] G. Zuccon, J. Palotti, and A. Hanbury, "Query variations and their effect on comparing information retrieval systems," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 691–700, 2016.

[12] R. Benham, J. S. Culpepper, L. Gallagher, X. Lu, and J. M. Mackenzie, "Towards efficient and effective query variant generation.," in *DESIRES*, pp. 62–67, 2018.

[13] A. Moffat, P. Bailey, F. Scholer, and P. Thomas, "Incorporating user expectations and behavior into the measurement of search effectiveness," *ACM Transactions on Information Systems (TOIS)*, vol. 35, no. 3, pp. 1–38, 2017.

[14] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 50–56, IEEE, 2018.

[15] S. Garg and G. Ramakrishnan, "Bae: Bert-based adversarial examples for text classification," *arXiv preprint arXiv:2004.01970*, 2020.

[16] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of nlp models with checklist," *arXiv preprint arXiv:2005.04118*, 2020.

[17] S. Zhuang and G. Zuccon, "Characterbert and self-teaching for improving the robustness of dense retrievers on queries with typos," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1444–1454, 2022.

[18] P. Bailey, A. Moffat, F. Scholer, and P. Thomas, "Uqv100: A test collection with query variability," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 725–728, 2016.

[19] C. Buckley and J. A. Walz, "The trec-8 query track.," in *TREC*, 1999.

[20] H. Hashemi, M. Aliannejadi, H. Zamani, and W. B. Croft, "Antique: A non-factoid question answering benchmark," in *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pp. 166–173, Springer, 2020.

[21] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pp. 232–241, Springer, 1994.

[22] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade, "Umass at trec 2004: Novelty and hard," *Computer Science Department Faculty Publication Series*, p. 189, 2004.

[23] C. Macdonald and N. Tonellotto, "Declarative experimentation in information retrieval using pyterrier," in *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pp. 161–168, 2020.

[24] C. Xiong, Z. Dai, J. Callan, Z. Liu, and R. Power, "End-to-end neural ad-hoc ranking with kernel pooling," in *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, pp. 55–64, 2017.

[25] Z. Dai, C. Xiong, J. Callan, and Z. Liu, "Convolutional neural networks for soft-matching n-grams in ad-hoc search," in *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 126–134, 2018.

[26] S. MacAvaney, "Opennir: A complete neural ad-hoc ranking pipeline," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 845–848, 2020.

[27] S. MacAvaney, F. M. Nardini, R. Perego, N. Tonellotto, N. Goharian, and O. Frieder, "Expansion via prediction of importance with contextualization," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 1573–1576, 2020.

[28] R. Nogueira, Z. Jiang, and J. Lin, "Document ranking with a pretrained sequence-to-sequence model," *arXiv preprint arXiv:2003.06713*, 2020.

[29] R. Nogueira, Z. Jiang, and J. Lin, "Document ranking with a pretrained sequence-to-sequence model," *arXiv preprint arXiv:2003.06713*, 2020.

[30] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "Ms marco: A human generated machine reading comprehension dataset," *choice*, vol. 2640, p. 660, 2016.

[31] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

[32] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp," *arXiv preprint arXiv:2005.05909*, 2020.

Appendix A

Name

A.1 Name