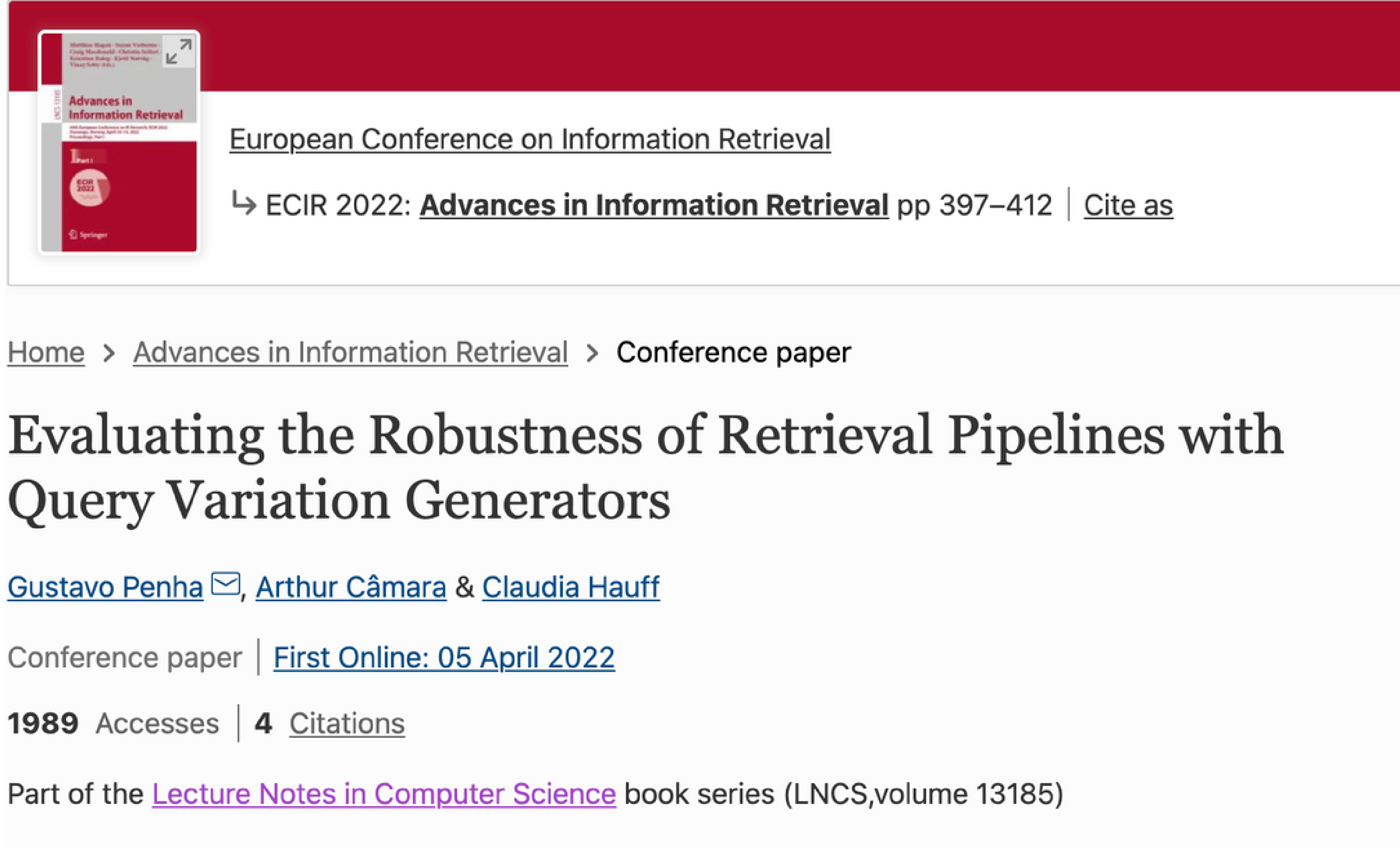
EVALUATING THE ROBUSTNESS OF RETRIEVAL PIPELINES WITH QUERY VARIATION GENERATORS

Krista Bradshaw 45285143
Supervisor: Guido Zuccon

THE TOPIC

A reproduction of the study by Penha, Câmara, and Hauff [1].


European Conference on Information Retrieval

↳ ECIR 2022: **Advances in Information Retrieval** pp 397–412 | Cite as

Home > Advances in Information Retrieval > Conference paper

Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators

Gustavo Penha, Arthur Câmara & Claudia Hauff

Conference paper | First Online: 05 April 2022

1989 Accesses | **4** Citations

Part of the Lecture Notes in Computer Science book series (LNCS,volume 13185)

THE AIMS

1. To conduct a thorough review of the existing literature
2. To accurately reproduce the experiments carried out in Penha et al.'s study
3. To expand on the original experiments

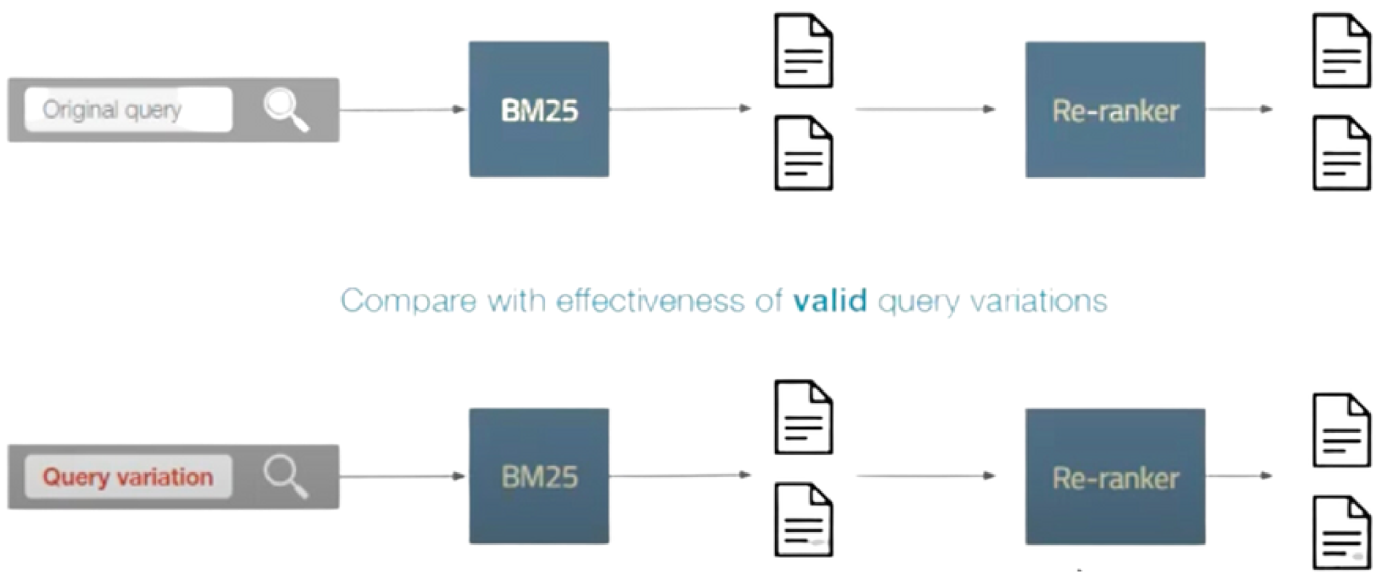
PROGRESS SO FAR

1. Become familiar with the original study
2. Conducted research into relevant background material and literature
3. Begun reproducing the dev environment and data processing

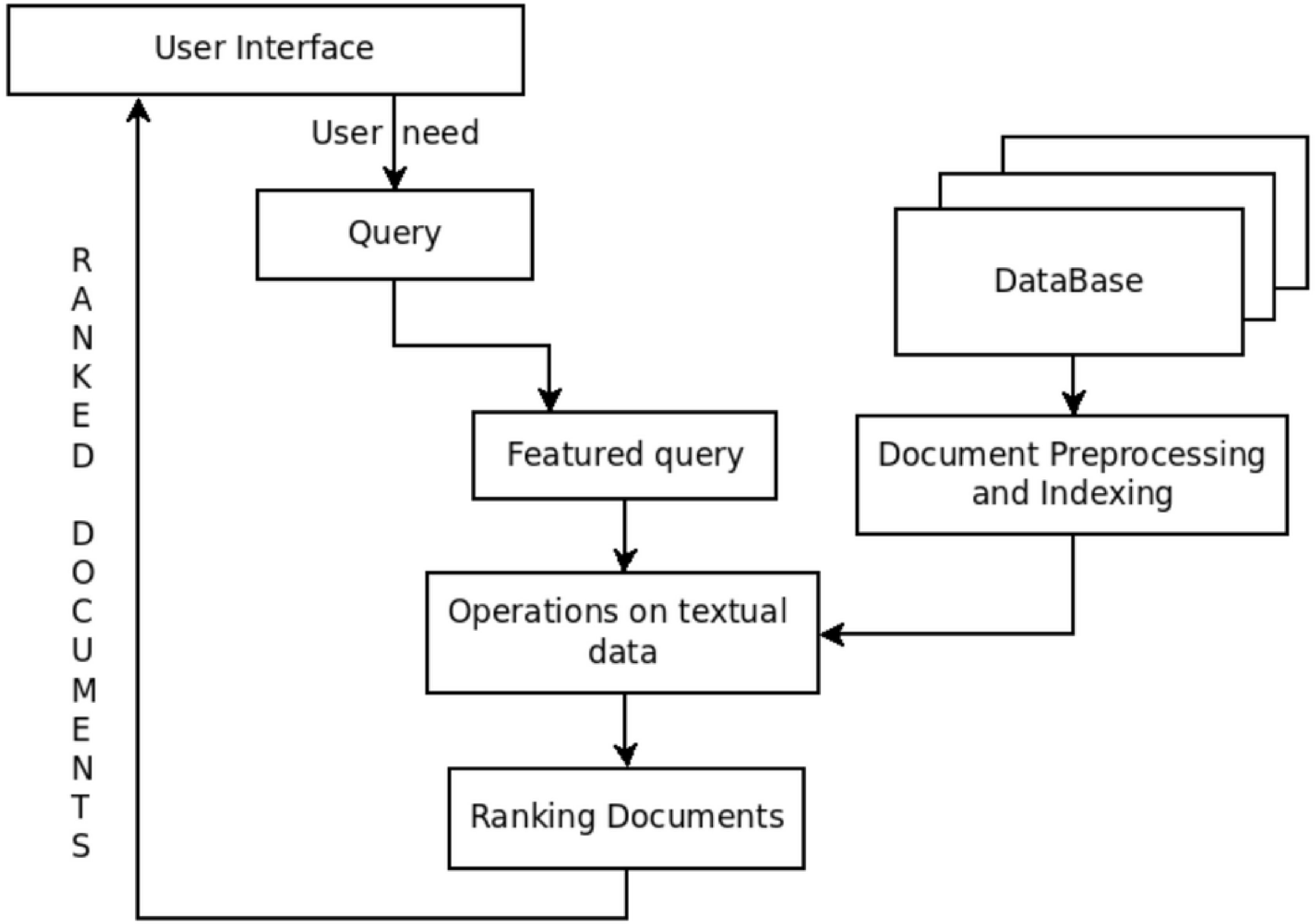
ORIGINAL STUDY

1. Analysed the UQV100 dataset to determine 6 types of query variations
2. Created 10 methods to automatically generate 4 of the 6 types
3. Generate one query variation for each of the proposed methods for each dataset (ANTIQUE and TREC-DL-2019)
4. Applied BM25 as a first stage retriever and then re-ranked the top 100 results with the neural ranking models (BM25, RM3, KNRM, CKNRM, EPIC, BERT, T5) for both the original queries and query variations
5. Compared the resulting ranked documents of the original query to each of its variations

Category	**Method Name**	*M*(‘***what is durable medical equipment consist of***’)
Misspelling	`NeighbCharSwap`	*what is durable* ***mdeical*** *equipment consist of*
	`RandomCharSub`	*what is durable* ***medycal*** *equipment consist of*
	`QWERTYCharSub`	*what is durable medical equipment* ***xonsist*** *of*
Naturality	`RemoveStopWords`	~~***what is***~~ *durable medical equipment consist* ~~***of***~~
	`T5DescToTitle`	~~***what is***~~ *durable medical equipment* ~~***consist of***~~
Ordering	`RandomOrderSwap`	***medical*** *is durable* ***what*** *equipment consist of*
Paraphrasing	`BackTranslation`	*what is* ***sustainable*** *medical equipment* ~~*consist of*~~
	`T5QQP`	*what is durable medical equipment* ~~***consist of***~~
	`WordEmbedSynSwap`	*what is durable* ***medicinal*** *equipment consist of*
	`WordNetSynSwap`	*what is* ***long lasting*** *medical equipment consist of*



BACKGROUND
Information Retrieval



Retrieval Pipelines

QUERY PROCESSING

The queries are processed to identify relevant terms, expand the query with synonyms, or apply other query optimisation techniques.

DOCUMENT INDEXING

The documents in the collection are indexed, typically using techniques such as token embedding.

RANKING

The ranked documents are then sorted by relevance using models.

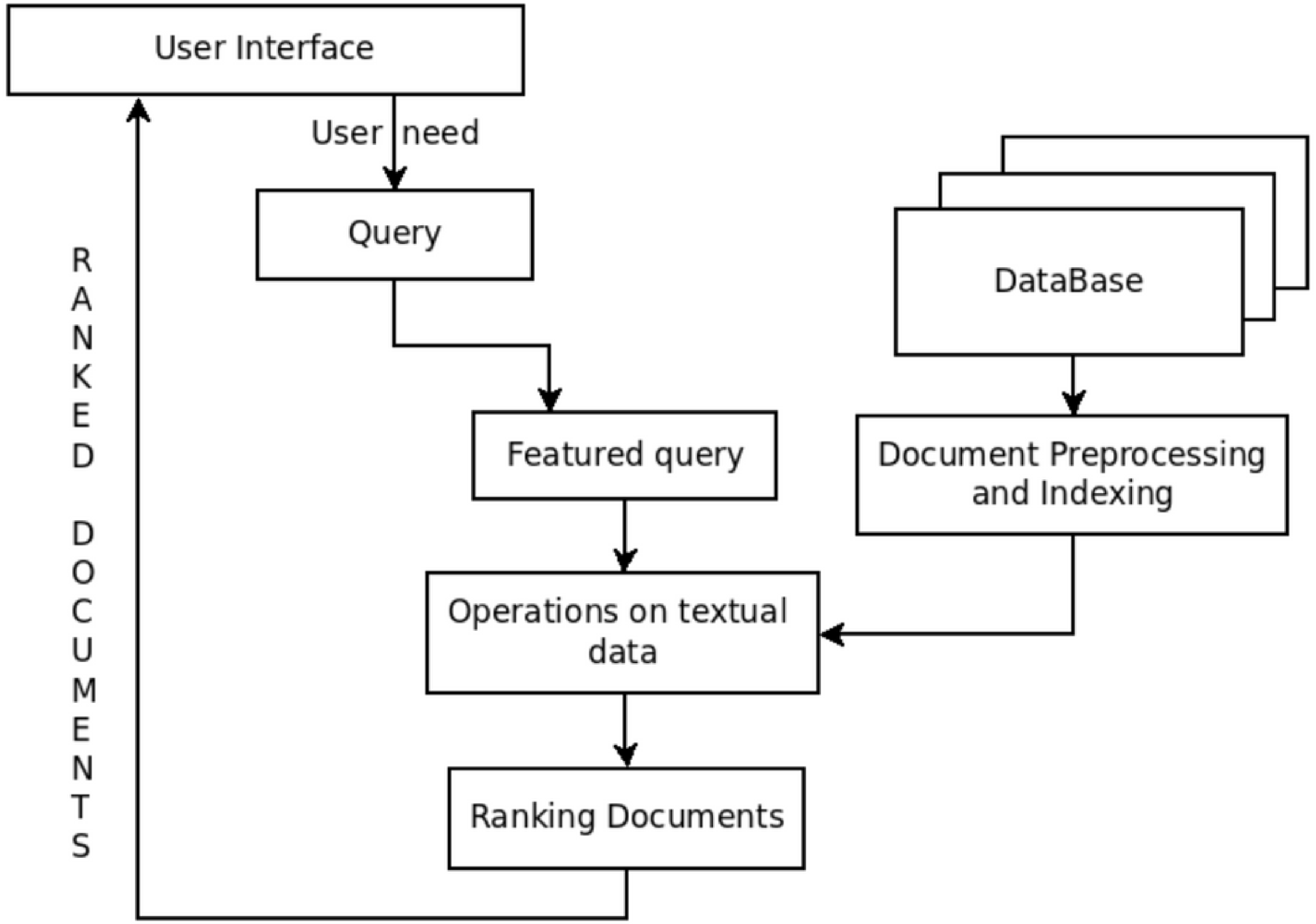
[2,3]

QUERY EXPANSION The process of adding synonyms or extra terms to the original query [4]

QUERY REFORMULATION The original query is augmented into a new query that better reflects the information need [4]

BACKGROUND

Information Retrieval



Retrieval Pipelines

QUERY PROCESSING

The queries are processed to identify relevant terms, expand the query with synonyms, or apply other query optimisation techniques.

DOCUMENT INDEXING

The documents in the collection are indexed, typically using techniques such as token embedding.

RANKING

The ranked documents are then sorted by relevance using models.

[2,3]

QUERY EXPANSION The process of adding synonyms or extra terms to the original query [4]

QUERY REFORMULATION The original query is augmented into a new query that better reflects the information need [4]

BACKGROUND

Query Variations

To handle query variation in IR systems, various techniques such as query suggestion, reformulation, or expansion can be employed.

Table 2: Taxonomy of query variations derived from a sample of the UQV100 dataset. Last column is the count of each query variation found on UQV100 based on manual annotation of tuples of queries for the same information need. * spelling errors were already fixed for the UQV100 pairs.

Category	Definition	Changes Semantics	$\{q_i, q_j\}$ Examples from UQV100			Count (%)
Gen./specialization	Generalizes or specializes within the same information need.	✓	american civil war	↔	number of battles in south carolina during civil war	172 (26.34%)
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	111 (17.00%)
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	118 (18.07%)
Ordering	Changes the order of words		carotid cavernous fistula treatment.	↔	treatment carotid cavernous fistula	37 (5.67%)
Paraphrasing	Rephrases the query by modifying one or more words.		cures for a bald spot	↔	cures for baldness	215 (32.92%)

[1]

LITERATURE

- The overarching goal is to develop better methods for information retrieval that can enhance search quality and user satisfaction.
- Explores deep learning techniques, query expansion and prediction techniques, and new evaluation metrics.
- Benchmark datasets has been emphasised and hence, datasets like UQV100, TREC, and DL-typo have been specifically generated to meet these needs.

ZENDEL ET AL. Conducted a novel investigation into the relationship between queries and information needs, with their approach to query performance prediction outperforming the baseline [3].

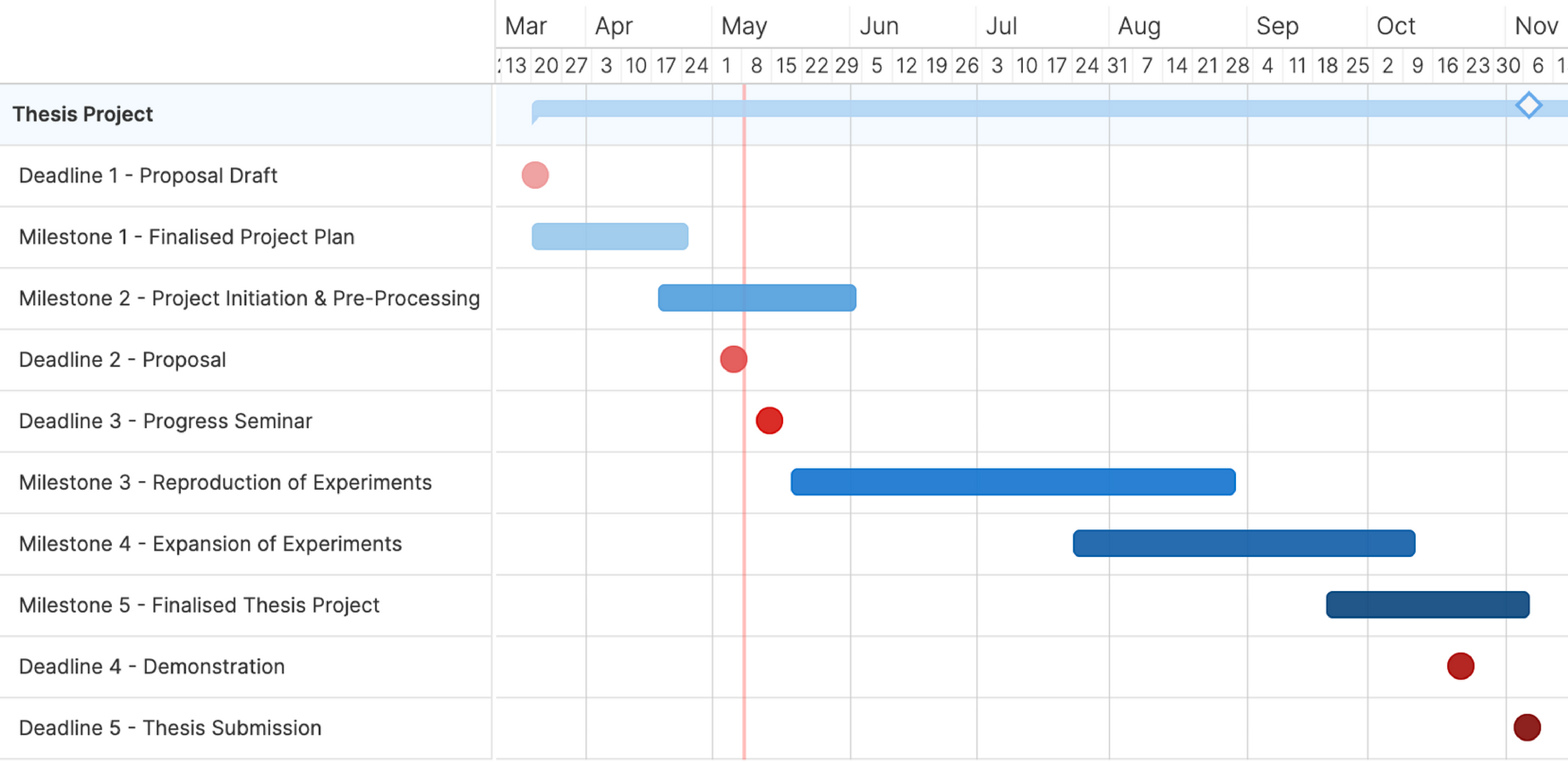
GAO ET AL. Developed a framework that generates queries designed to trick models into incorrectly classifying them. Their results showed a significant decrease in effectiveness [5].

ZHUANG ET AL. Conducted a study to tackle the issue of current dense retrievers struggling with unusual queries [6].

LU ET AL. Study on relevance modelling with multiple queries also demonstrated that utilising query variations performs substantially better than using a single query. Their experiments involved fusion at the term, query, and document level, which is a new concept in this field and shows promising results [7].

ZUCCON ET AL. Study focused on examining the role of query variations in comparing system effectiveness, and proposed a framework that explicitly incorporates query variations. Their analysis considers not only the mean effectiveness of the system but also its variance across different query variations and topics. The findings reveal a significant impact of query variations on the comparison of different systems [8].

PLAN FOR DEVELOPMENT



MILESTONE 1 Focused on nailing down the project plan in preparation for submission of Deadline 1.

MILESTONE 2 Focuses on gathering and pre-processing the required datsets and models.

MILESTONE 3 Focuses on conducting all experiments as per the original study.

MILESTONE 4 Focuses on making adjustments to the resources and expanding on the original experiments to address gaps in this field of study.

MILESTONE 5 Focuses on finalising all experiments and analysation of data in preparation for submission of the written thesis report.

PROGRESS

Generate one query variation for each of the proposed methods for each dataset (ANTIQUE and TREC-DL-2019)

antique-train-split200-valid_weakly_supervised_variations_sample_5.csv

q_id	original_query	variation	method	transformation_type
1158088	why my neighbor's dog doesn't bark?	a neighbor's dog doesn	summarization_with_t5-base	naturality
1158088	why my neighbor's dog doesn't bark?	neighbor's dog barks	summarization_with_t5-base_from_description_to_title	naturality
1158088	why my neighbor's dog doesn't bark?	Why doesn't my neighbor's dog bark?	ramsrigouthamg/t5_paraphraser	paraphrase
1184520	why be an "a:theist ?	why be an "a:thiest ?	WordSwapNeighboringCharacterSwap	mispelling
1184520	why be an "a:theist ?	why be an "a:thwist ?	WordSwapQWERTY	mispelling
1184520	why be an "a:theist ?	why be an "a:thVist ?	WordSwapRandomCharacterSubstitution	mispelling
1184520	why be an "a:theist ?	"a:theist ?	naturality_by_removing_stop_words	naturality
1184520	why be an "a:theist ?	a:theist :	summarization_with_t5-base	naturality
1184520	why be an "a:theist ?	a:theist	summarization_with_t5-base_from_description_to_title	naturality
1184520	why be an "a:theist ?	a be an "why:theist ?	WordInnerSwapRandom	ordering
1184520	why be an "a:theist ?	Why be a ,Äúa:theist,Äù?	back_translation_pivot_language_de	paraphrase
1184520	why be an "a:theist ?	Why should I be an "atheist"?	ramsrigouthamg/t5_paraphraser	paraphrase
1184520	why be an "a:theist ?	why be an "a:theist ?	WordSwapEmbedding	synonym
1184520	why be an "a:theist ?	why be an "a:theistic ?	WordSwapWordNet	synonym
1398838	what does "crunching numbers" mean?	what does "crunchign numbers" mean?	WordSwapNeighboringCharacterSwap	mispelling
1398838	what does "crunching numbers" mean?	what does "crunching numbers" mewn?	WordSwapQWERTY	mispelling

msmarco-passage-trec-dl-2019-judged_weakly_supervised_variations_sample_5.csv

q_id	original_query	variation	method	transformation_type
1037798	who is robert gray	robert gray is	summarization_with_t5-base	naturality
1037798	who is robert gray	robert gray	summarization_with_t5-base_from_description_to_title	naturality
1037798	who is robert gray	Who is Robert Gray?	ramsrigouthamg/t5_paraphraser	paraphrase
1063750	why did the us volunterilay enter ww1	why did the su volunterilay enter ww1	WordSwapNeighboringCharacterSwap	mispelling
1063750	why did the us volunterilay enter ww1	why did the us vopunterilay enter ww1	WordSwapQWERTY	mispelling
1063750	why did the us volunterilay enter ww1	why did the Ms volunterilay enter ww1	WordSwapRandomCharacterSubstitution	mispelling
1063750	why did the us volunterilay enter ww1	us volunterilay enter ww1	naturality_by_removing_stop_words	naturality
1063750	why did the us volunterilay enter ww1	why did the us volunteri	summarization_with_t5-base	naturality
1063750	why did the us volunterilay enter ww1	volunterilay ww	summarization_with_t5-base_from_description_to_title	naturality
1063750	why did the us volunterilay enter ww1	why did the us ww1 enter volunterilay	WordInnerSwapRandom	ordering
1063750	why did the us volunterilay enter ww1	Why the U.S. Volunterilay entered WW1	back_translation_pivot_language_de	paraphrase
1063750	why did the us volunterilay enter ww1	Why did the US enter WW1?	ramsrigouthamg/t5_paraphraser	paraphrase
1063750	why did the us volunterilay enter ww1	why did the usa volunterilay enter ww1	WordSwapEmbedding	synonym
1063750	why did the us volunterilay enter ww1	why did the us volunterilay figure ww1	WordSwapWordNet	synonym
1110199	what is wifi vs bluetooth	what is wifi vs bleutooth	WordSwapNeighboringCharacterSwap	mispelling
1110199	what is wifi vs bluetooth	what is sifi vs bluetooth	WordSwapQWERTY	mispelling

Code snippit:

```
logging.info("Generating weak supervision for task {} and saving results in {}."\
    .format(args.task, args.output_dir))

dataset = ir_datasets.load(args.task)
queries = [t[1].lower() for t in dataset.queries_iter()]
q_ids = [t[0] for t in dataset.queries_iter()]

pa = ParaphraseActions(queries, q_ids, args.output_dir)
transformed_queries_paraphrase_models = pa.seq2seq_paraphrase(sample=args.sample)
transformed_queries_back_translation = pa.back_translation_paraphrase(sample=args.

na = NaturalityActions(queries, q_ids)
transformed_queries_trec_desc_to_title = na.naturality_by_trec_desc_to_title(model
transformed_queries_stop_word_removal = na.remove_stop_words(sample=args.sample)
# transformed_queries_stop_word_and_stratified_removal = na.remove_stop_words_and_
transformed_queries_summarizer = na.naturality_by_summarization(sample=args.sample

sa = SynonymActions(queries, q_ids)
transformed_queries_syn = sa.adversarial_synonym_replacement(sample=args.sample)

oa = OrderingActions(queries, q_ids)
transformed_queries_shuffled_order = oa.shuffle_word_order(sample=args.sample)

ma = MispellingActions(queries, q_ids)
transformed_queries_mispelling = ma.mispelling_chars(sample=args.sample)


transformed_queries =  transformed_queries_mispelling +\
  transformed_queries_shuffled_order  +\
  transformed_queries_syn + \
  transformed_queries_paraphrase_models + \
  transformed_queries_back_translation +  \
  transformed_queries_stop_word_removal +  \
  transformed_queries_summarizer + \
  transformed_queries_trec_desc_to_title

transformed_queries = pd.DataFrame(transformed_queries, columns =
                                   ["q_id", "original_query", "variation", "method
transformed_queries.sort_values(by=["q_id", "transformation_type", "method"]).to_c
    "{}/{}_weakly_supervised_variations_sample_{}.csv".format(args.output_dir,
    args.task.replace("/",'-'), args.sample), index=False)
```

PROGRESS

Variation Generating Methods

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.

THANK YOU

References

1. G. Penha, A. Cˆamara, 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. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.

ANY QUESTIONS?