ROBUSTNESS OF RETRIEVAL PIPELINES

AUTHORED BY KRISTA BRADSHAW, SUPERVISED BY GUIDO ZUCCON

OVERVIEW

A reproduction of the study of retrieval pipeline robustness when faced with automatically generated query variations by Penha, Câmara, and Hauff [1].

Research Question 1

To what extent can the findings and observations of the original study be successfully **reproduced** using the same datasets, models, and methodologies?

Research Question 2

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?

BACKGROUND

Retrieval Pipelines

- **Query Processing:** Involves tokenising, stemming, and other preprocessing to prepare for matching against corpus docs.
- **Document Scoring**: Based on their relevance to the query.
- **Ranking:** Rank docs by their scores in descending order.

Query Variations

Original	how sun rises
Misspelling	how **us**n rises
Naturality	~~**how**~~ sun rises
Ordering	**rises** sun how
Paraphrase	how **does the** sun rise
Synonym	how sun **soars**

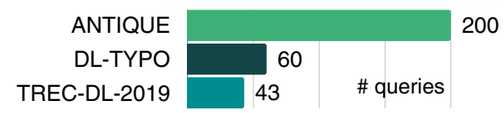
Examples from ANTIQUE

RELATED WORKS

- Zendel et al. — Query performance prediction [2]
- Zuccon et al. — Investigation into query variations and their effect [3]
- Zhuang et al. — Model that incorporates query variations (CharacterBERT) [4]
- Lu et al. — Relevance modelling with query variations and rank fusion [5]

METHODOLOGY

Datasets



Models

Traditional	Neural	Transformer
BM25, RM3	KNRM, CKNRM	BERT, EPIC, T5

Methods

Misspelling: NeighbCharSwap, RandomCharSub, QWERTYCharSub
Naturality: RemoveStopWords, T5DescToTitle
Ordering: RandomOrderSwap
Paraphrase: BackTranslation, T5QQ, WordEmbedSynSwap, WordNetSynSwap

Variation Generator Quality

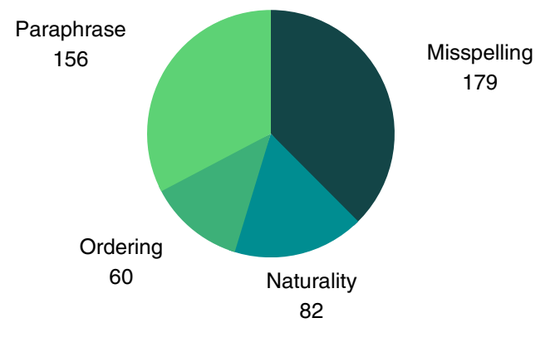


Figure 1: Number of valid variations per category in DL-TYPO (477 valid of 600 total)

RESULTS

ANTIQUE

Table 1: Effectiveness (nDCG@10) summary when faced with different query variations.

Category	BM25	RM3	KNRM	CKNRM	EPIC	BERT	T5
Original	0.23	0.22	0.22	0.21	0.27	0.36	0.33
Highest	OG	OG	OG	0.22 RemoveStopWords	0.2	OG	OG
Lowest	0.16 NeighbCharSwap	0.15	0.16	0.14 WordNetSynSwap	0.18 NeighbCharSwap	0.24 T5DescToTitle	0.25 RandomCharSwap
RRF Highest	0.28	0.27	0.3	0.28	0.34	0.45	0.39

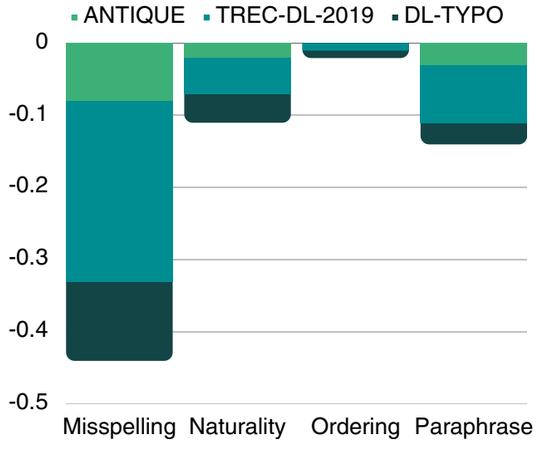
TREC-DL-2019 results omitted due to similarity

DL-TYPO

Table 2: Effectiveness (nDCG@10) summary when faced with different query variations.

Category	BM25	RM3	KNRM	CKNRM	EPIC	BERT	T5
Original	0.19	0.18	0.21	0.18	0.19	0.2	0.29
Highest	OG	OG	OG	OG	0.2	0.22 T5QQP	0.29
Lowest	0.07 RandomCharSub	0.06	0.08	0.06 QWERTYCharSub	0.07	0.06 WordEmbedSynSwap	0.13
RRF Highest	0.32	0.3	0.28	0.32	0.31	0.29	0.42

Figure 2: Average nDCG@10 Δ values per variation category for each dataset.



CONCLUSIONS

- Query variations do **negatively impact** retrieval effectiveness, confirming the findings of the original study (occurred in 51, 41, and 32 out of 70 instances per dataset)
- Fusion of query variation categories showed **improvements in some instances** but did not consistently surpass the retrieval outcomes of the original queries.
- Misspellings caused the highest effectiveness drops on average compared to other methods.
- **DL-TYPO consistently underperformed** in comparison with ANTIQUE and TREC-DL-2019 due to the additional misspellings.

1. G. Penha, A. Cˆamara, and C. Hauff, “Evaluating the robustness of retrieval pipelines with query variation generators.”
2. O. Zendel, A. Shtok, F. Raiber, O. Kurland, and J. S. Culpepper, “Information needs, queries, and query performance prediction.”
3. G. Zuccon, J. Palotti, and A. Hanbury, “Query variations and their effect on comparing information retrieval systems.”
4. S. Zhuang and G. Zuccon, “Characterbert and self-teaching for improving the robustness of dense retrievers on queries with typos.
5. Lu, O. Kurland, J. S. Culpepper, N. Craswell, and O. Rom, “Relevance modeling with multiple query variations.”