A Comparative Analysis of Human and Automatic Query Variants

Binsheng Liu, Nick Craswell, Xiaolu Lu, Oren Kurland and J. Shane Culpepper

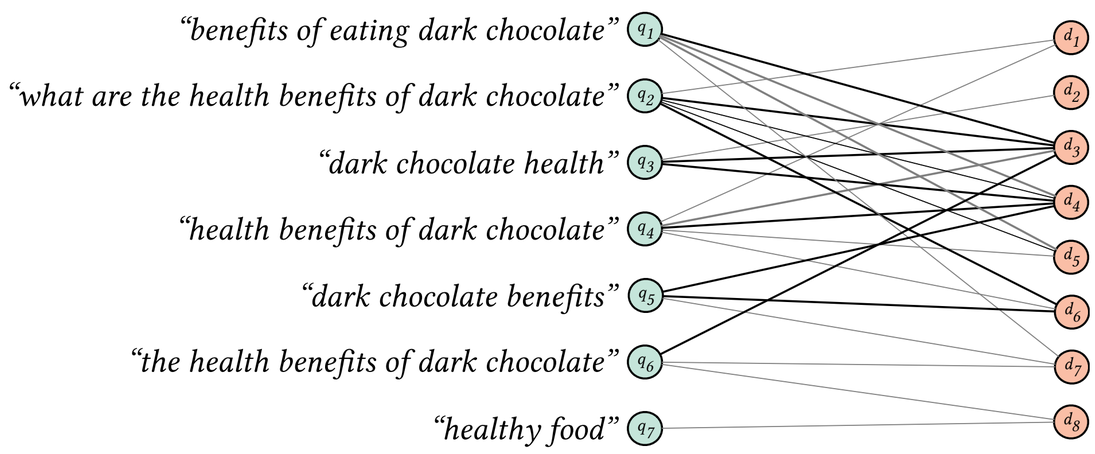
RMIT UNIVERSITY

TECHNION Israel Institute of Technology

1. Contribution

- We present a comprehensive analysis of two sets of queries for the same information need: **human written variants** and **automatically generated variants**.
- We show that both sets of variants can achieve comparable performance, while they can be appreciably different in several important respects.
- We empirically show that remarkable effectiveness gains are still possible based purely on the query formulation of an information need.

2. Automatic Query Variants



A bipartite click-graph, showing the associations of document clicks from queries. The thickness of each line represents the frequency of clicks for that query and document pair.

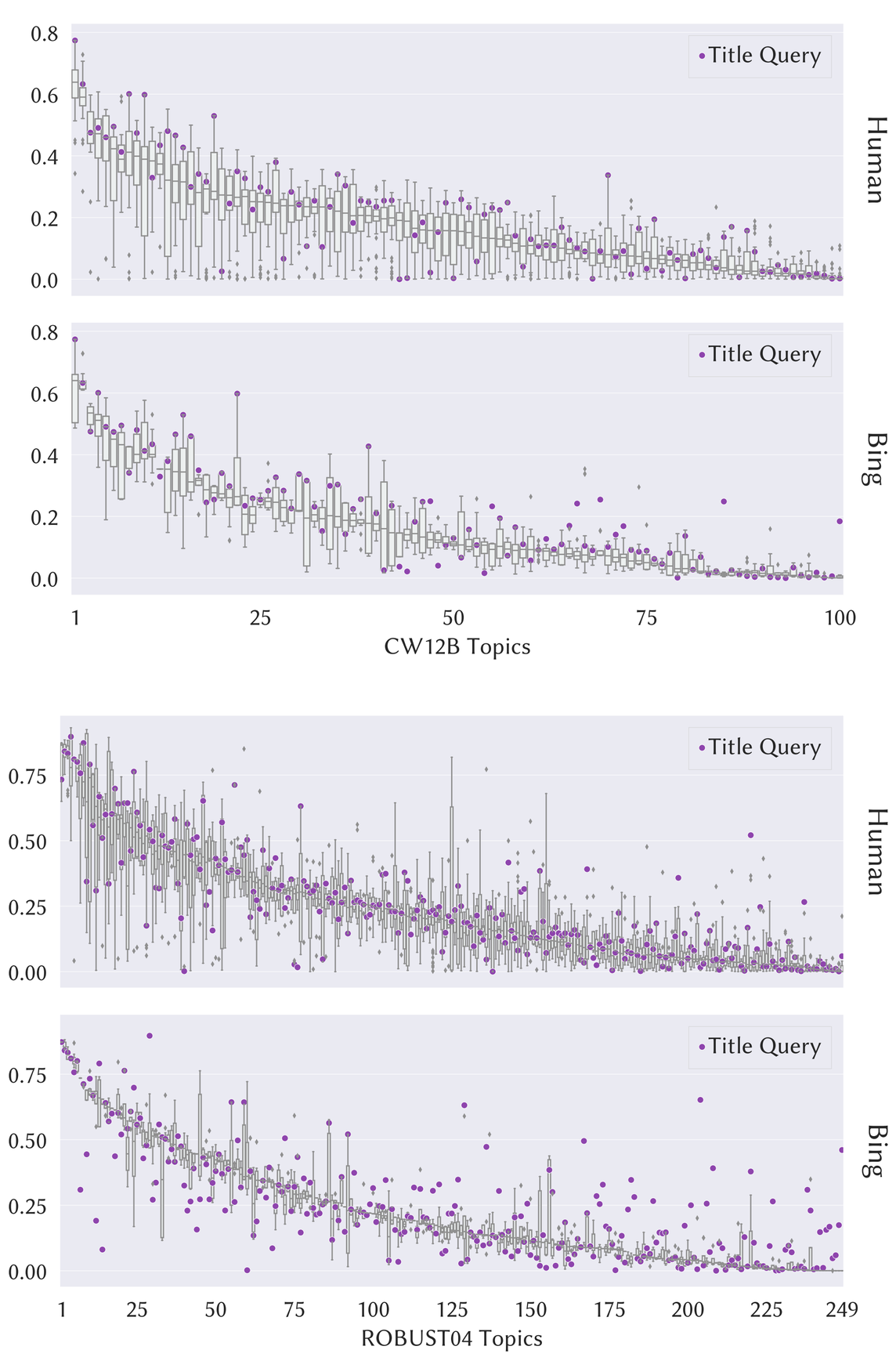
3. Retrieval Performance

Query Set		CW12B MAP	CW12B NDCG@10	CW12B RBP@0.95
Title query	-	0.201	0.192	0.360+0.213
Human	Median	0.178	0.190	0.351+0.185
Bing	0.5	0.166‡	0.192‡	0.323+0.313†
	0.7	0.194†	0.210	0.366+0.271
Human	Best	0.286	0.304	0.501+0.118
Bing	Best	0.239	0.252	0.428+0.215
Combined	Best	0.288^{b}	0.303^{b}	0.503+0.120^{b}

Query Set		ROBUST MAP	ROBUST NDCG@10	ROBUST RBP@0.95
Title query	-	0.247	0.426	0.308+0.035
Human	Median	0.239	0.421	0.294+0.124
Bing	0.5	0.201†	0.358†	0.248+0.249†
	0.7	0.228‡	0.402	0.281+0.216‡
Human	Best	0.373	0.604	0.422+0.078
Bing	Best	0.282	0.481	0.338+0.170
Combined	Best	0.389^{h,b}	0.621^{h,b}	0.436+0.081^{h,b}

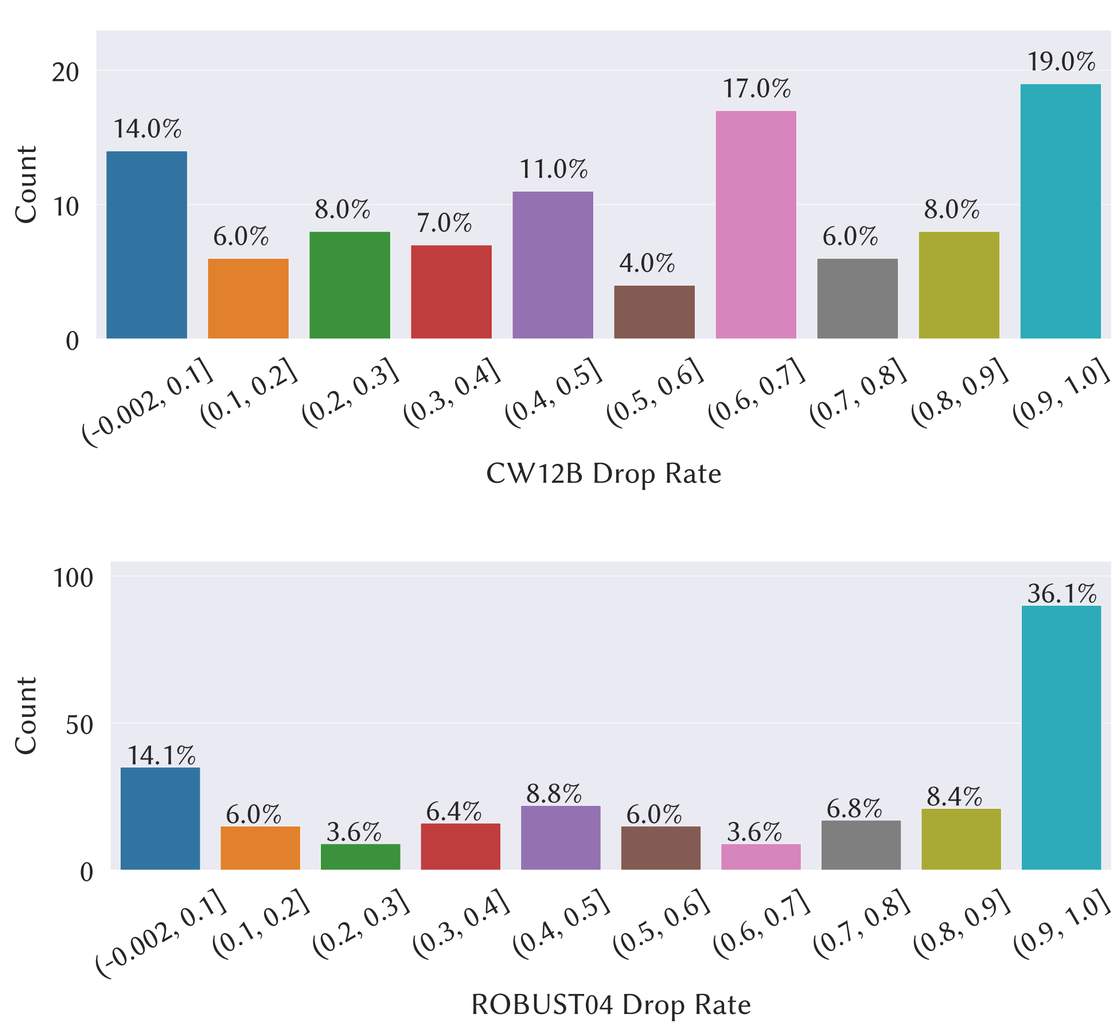
All statistical significance tests are performed against median queries in both query sets. † and ‡ mean $p < 0.05$ in the t-test and TOST test (ΔAP $= 0.05$), respectively.

4. Results Breakdown



Per-topic comparisons, ordered by the median of corresponding variants. Automatic query variants are in the pruned set, where the pruning percentage are 50% and 70% on CW12B and ROBUST, respectively.

5. Automatic Variants Drop Rate



Per-topic drop rate of automatic query variants needed to achieve performance comparable to that of human variants. The x-axis is the drop rate and the y-axis is the number of dropped variants.

6. Summary

- Automatically generated variants and human written variants can achieve comparable performance, while subtle differences between the queries still exist.
- Automatic variants and human variants have their own strengths in representing an information need and can complement each other.
- Understanding how query variants affect other ranking algorithms, LTR, query expansion, fusion, etc. is an interesting future research question.

7. Funding

This work was supported in part by the Israel Science Foundation (grant no. 1136/17), the Australian Research Council's *Discovery Projects* Scheme (DP170102231), an Amazon Research Award, and a Google Research Award.