


**Touché Shared Task 2**
Argument Retrieval for Comparative Questions

# LEVIRANK: Limited Query Expansion with Voting Integration for Document Retrieval and Ranking

*Ashish Rana*, Pujit Golchha, Roni Juntunen, Andreea Coajă, Ahmed Elzamarany, Chia-Chien Hung, Simone Paolo Ponzetto

UNIVERSITY OF MANNHEIM

LUT
Lappeenranta
University of Technology

*Data and Web Science Group, University of Mannheim, Germany*
*Lappeenranta-Lahti University of Technology LUT, Finland*

**TEAM**
CAPTAIN LEVI

# Table of Contents

# Introduction

## Motivation

People always use Web to find new things, And ***often*** they ***compare*** them as well! *(Turner et al. (2020), Bondarenko et al. (2020))*

For example, Which footballer has most goals? ***(Factual)***, Who is the best footballer? ***(Contextual)*** *(Trivedi et al. (2020))*

## Problem Statement

How can we find ***relevant information*** to such questions on the Web? Which also helps in the ***decision process***?

## Related Work

***Decision process, which is better?*** Given (q,d), get answer {'Obj. 1','Obj. 2','Neu.','None'} *(Bondarenko et al. (2022))*

***Retrieval process, which document is better?*** Extensively studied, in political (FEVER) & scientific discourses (SCIVER), and ad-hoc retrievals (MS-MARCO) *(Thorne et al. (2018), Wadden et al. (2020), Nguyen et al. (2016))*

***Combining retrieval, given document relevance/quality which object is better?*** *(Touché Overview Papers (2020, 2021) )*

# Problem and Dataset Description

### Problem Formulation

For given query retrieve relevant documents, classify the corresponding stance, and evaluate the system components

### Datasets Used

DocT5Query expanded corpora w/ 0.9 million text passages *(relevance)*, 956 comparative QA dataset containing Yahoo & Stack Exchange QA pairs *(stance). (Nogueira et al. (2019), Bondarenko et al. (2022))*

# Initial Approach & Evaluation

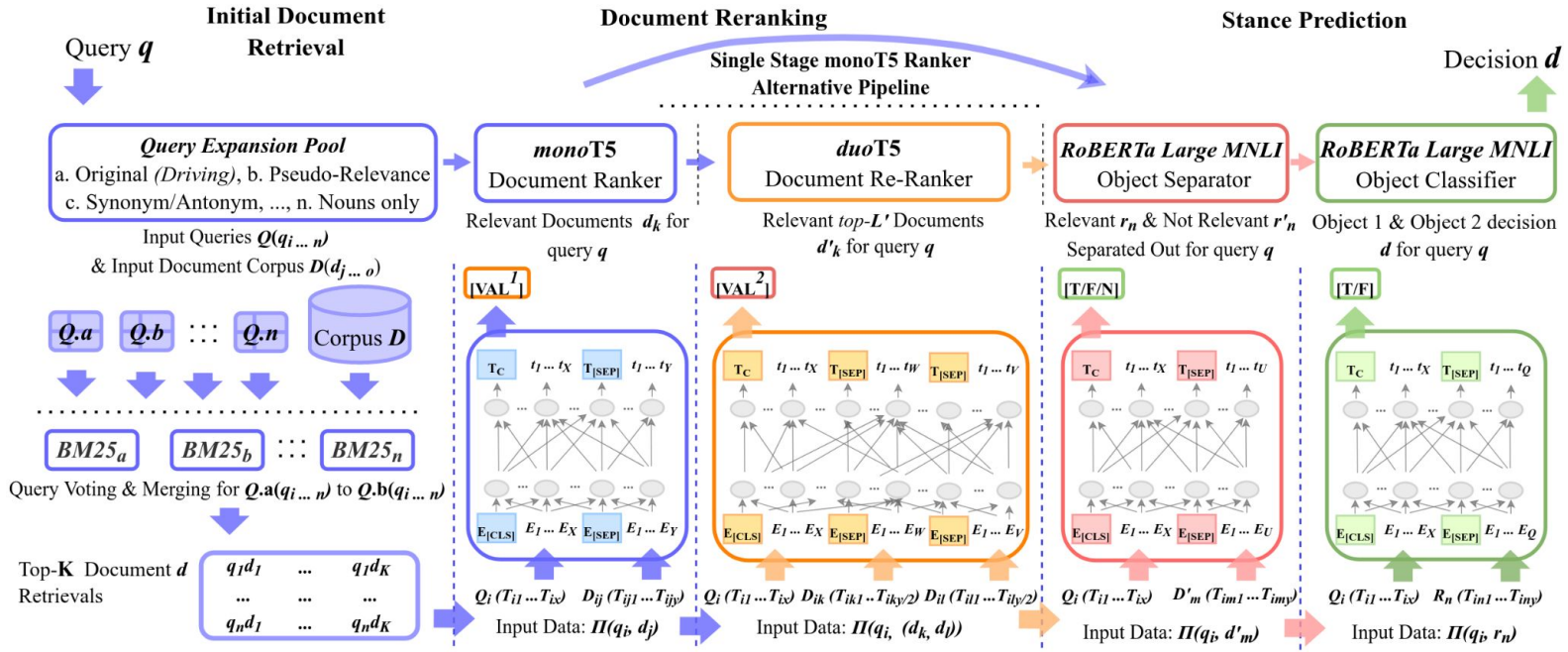
### Our System Approach

We tested *'Expando-Mono-Duo'* design pattern, and two-step stance prediction *(Pradeep et al. (2021), Zeng et al. (2021))*

### Initial Evaluation Approach

100 queries from past 2 iterations, ChatNoir urls of above corpora & merged sub-documents. For relevance we used nDCG@5 metric, and Macro-F1 was used for entailment detection

# LeviRANK: System Architecture




**A. Larger documents (≥ 512 tokens):** Initial Retrieval, Ranking (Mono-T5 only), Stance Prediction

**B. Smaller documents (< 512 tokens):** Initial Retrieval, Multi-stage Ranking (Mono-T5 & Duo-T5), Stance Prediction

## **Initial Retrieval:** Approaches Explored

**General Module Implementation Details for Submission**

- Preprocessing: lowercase, stopword removal, WordNet lemmatization
- DocT5Query expanded corpus used during submission
- But, below results reported on the merged document data
- Focus of initial retrieval stage to improve Recall@K values

Recall@K: Number of relevant documents present at K number of documents are retrieved by Initial Retrieval module

**Approaches Implemented & Tested**

- Previous Baselines: TF-IDF
- Probabilistic Approaches: BM25, BM25 + Pseudo-Relevance Feedback, LeviRank
- Dense Retrievals post larger BM25 retrieval: Cosine similarity on SimCSE's contrastive embeddings
- Dense index building & retrieval: TCT-ColBERT

*Performance Summary**: { TF-IDF < Contrastive Learning < **Dense Index** < BM25 < **LeviRANK Voting**}*

## **Initial Retrieval:** Result Findings

| Retrieval Approach | Recall@1000 | Recall@1500 | Recall@2000 |
|---|---|---|---|
| BM25 Baseline | **90,18** | 90.67 | 91.11 |
| Dense Retrieval | 85.70 | 86.56 | 87.56 |
| Pseudo-Relevance Feedback | 89.98 | 90.59 | 91.07 |
| **LeviRANK Voting** | 90.14 | **91.08** | **91.17** |

## **Document Ranking:** Approaches Explored & Result Findings

**General Module Implementation Details for Submission**

- Approaches explored: DistilBERT *(Previous Baseline)*, monoT5, monoT5-duoT5 multi-stage ranking. Scoring metric used, nDCG@5
- Results reported on merged document dataset against 100 topic queries from previous years, lower performance bound guarantee

| Ranking Approach | BM25 | monoT5-only | monoT5-duoT5 |
|---|---|---|---|
| nDCG@5 | 0.33 | **0.47** | 0.31 |

## **Stance Prediction:** Approach & Result Findings

**General Module Implementation Details for Submission**

- Two step multi-class classification approach: First, classifying (q,d) pairs {'None','Neu.','Obj.'} and secondly, {'First', 'Second'} objects
- RoBERTa-Large-MNLI pre-trained models used, fine-tuning on the given QA dataset, Macro-F1 score reporting for all classes

| Approach | No object | Neutral | Object 1 | Object 2 | Macro-F1 |
|---|---|---|---|---|---|
| Bondarenko et al. (2022) | 0.40 | **0.53** | 0.72 | 0.63 | 0.57 |
| LEVIRANK | 0.40 | 0.52 | **0.72** | **0.68** | **0.58** |

# Leaderboard Result Summary

**First Table**, reporting the nDCG@5 submitted systems & **Second Table**, highlighting stance prediction performance improvement scope

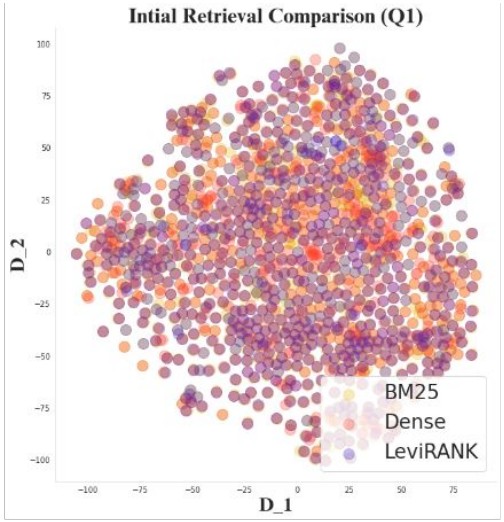
| Submitted Approaches | Recall@2K | Input Size for duoT5 | nDCG@5 Relevance | nDCG@5 Quality |
|---|---|---|---|---|
| TCT-ColBERT+monoT5+duoT5 | 92.05 | 100 | 0.758 *(1)* | 0.744 *(2)* |
| BM25+monoT5+duoT5 | 98.23 | 100 | 0.755 | 0.742 |
| LEVIRANK+PR+monoT5+duoT5 | 97.96 | 50 | 0.753 | 0.730 |
| LEVIRANK+monoT5 | **98.34** | 0 | 0.727 | 0.706 |
| Pseudo-Relevance(PR)+monoT5 | 97.16 | 0 | 0.722 | 0.695 |

DuoT5 *(small documents size attribution)* & TCT-ColBERT perform surprisingly better, LEVIRANK approach can outperform the TCT-ColBERT.

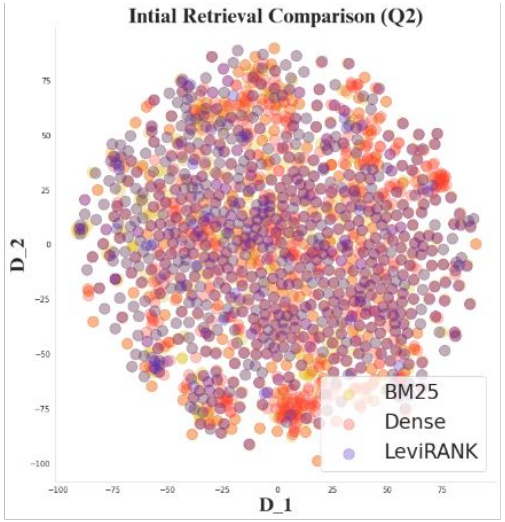
| Training Approach | Prediction Annotation Set | Macro-F1 |
|---|---|---|
| Zero-shot Two-Step RoBERTa-MNLI | Whole stance dataset | 0.303 *(2)* |
| Zero-shot Two-Step RoBERTa-MNLI | Worst 50 % topic queries | 0.116 *(6)* |
| Zero-shot Two-Step RoBERTa-MNLI (fine tuned, 50 % best queries) | Worst 50 % topic queries | **0.387** *(1)* |

Approaches explored: DistilBERT *(Previous Baseline)*, monoT5, monoT5-duoT5 multi-stage ranking. Scoring metric used, nDCG@5

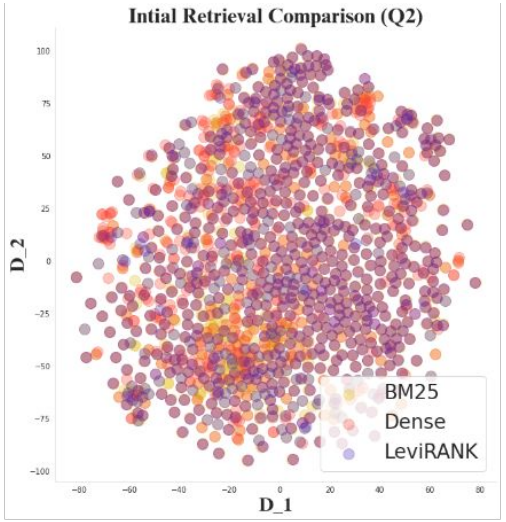
# Result Summary




Q1. What is better Google search or Yahoo search?




Q2. Which is better MAC or PC?




Q3. Which is better Family Guy or The Simpsons?

*Geometric Interpretation of retrieved results, LeviRANK system's initial retrieval attempts to increase the variation in different retrievals from multiple newly spawned queries with restricted {updated, removed, added} keywords*

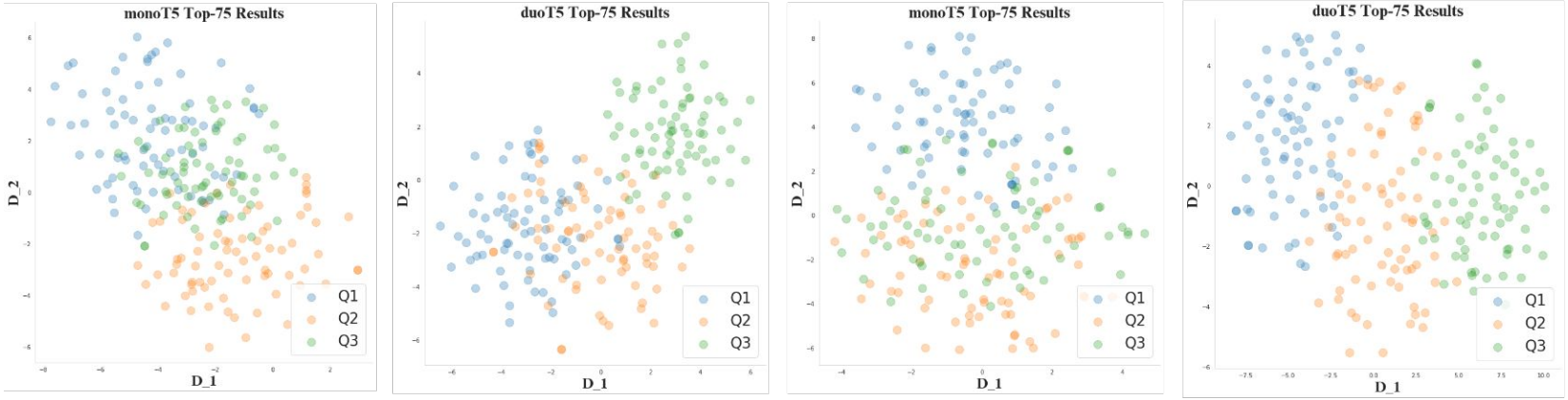
# Result Summary

Similar Topic Queries

Q1. What is better, a laptop or a desktop?
Q2. What is better, MAC or PC?
Q3. Why is Linux better than Windows?

Dissimilar Topic Queries

Q1. What is better, Canon or Nikon?
Q2. What city is better, London or Paris?
Q3. Who is stronger, Hulk or Superman?




a. monoT5 Ranker on similar queries. b. duoT5 Ranker on similar queries. c. monoT5 Ranker on dissimilar queries. d. duoT5 Ranker on dissimilar queries.

*Retrieval document set comparison for the monoT5 & monoT5-duoT5 multi-stage ranking systems. Here, duoT5 system presents strong discriminative qualities for the top retrieved documents for both similar and dissimilar queries*

**Software,** github.com/softgitron/LeviRank

## Conclusion & Approach Improvements

- The ‘Expando-Mono-Duo’ design even without fine-tuning captures argumentation structure via self-attention
- duoT5 model was great for smaller documents, TCT-ColBERT retrieval for LeviRANK showed more relative success
- Stance prediction suffers stark performance decrease, but can perform better with further fine-tuning

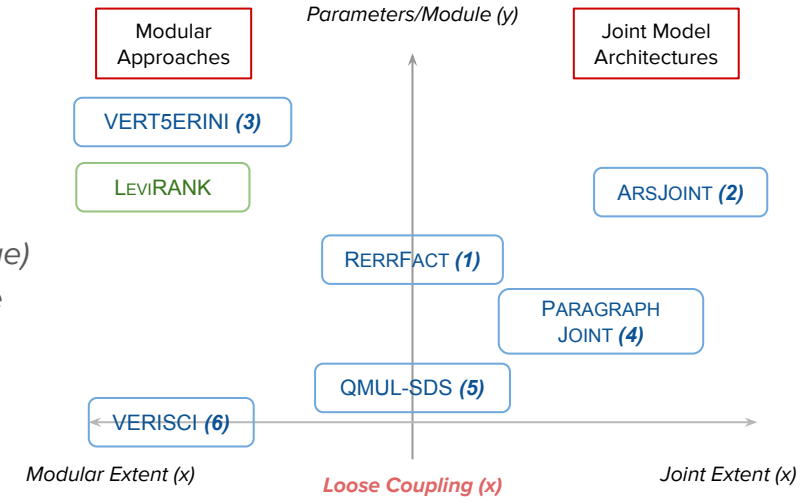
## Approach Improvements Suggestions

### Improvements

- Further fine-tuning the retrieval & stance models
- Combining limited query spawning with TCT-ColBERT
- Encouraging *‘loosely coupled’* retrieval & ranking designs
- Encouraging retrieving reduced document representations *(large)*
- Better error analysis, & Using query description in ranking stage

### Results Caveats

- TCT-ColBERT & duoT5 performance limited to small documents
- Additionally, the {‘None’, ‘No’} class distinguishing capabilities really not good during stance prediction




*SCIVER Shared Task system paradigms, RERRFACT’s (Rana et al. 2022) simplistic design & devset performance gains.*

**Thank you,** Looking forward to discussion!

# References


- E. Turner, L. Rainie, Most americans rely on their own research to make big decisions, and that often means online searches (2020).
- A. Bondarenko, P. Braslavski, M. Völske, R. Aly, M. Fröbe, A. Panchenko, C. Biemann, B. Stein, M. Hagen, Comparative web search questions, in: Proceedings of the 13th International Conference on Web Search and Data Mining, 2020, pp. 52–60.
- H. Trivedi, H. Kwon, T. Khot, A. Sabharwal, N. Balasubramanian, Repurposing entailment for multi-hop question answering tasks, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2948–2958. URL: https://aclanthology.org/N19-1302. doi:10.18653/v1/N19-1302.
- A. Bondarenko, Y. Ajjour, V. Dittmar, N. Homann, P. Braslavski, M. Hagen, Towards Understanding and Answering Comparative Questions, in: K. S. Candan, H. Liu, L. Akoglu, X. L. Dong, J. Tang (Eds.), 15th ACM International Conference on Web Search and Data Mining (WSDM 2022), ACM, 2022, pp. 66–74. URL: https://dl.acm.org/doi/10.1145/3488560. 3498534. doi:10.1145/3488560.3498534.
- J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074. doi:10.18653/v1/N18-1074.
- D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7534–7550. URL: https://aclanthology.org/2020.emnlp-main.609. doi:10. 18653/v1/2020.emnlp-main.609.

# References


- T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, Ms marco: A human generated machine reading comprehension dataset, in: CoCo@ NIPS, 2016.
- A. Bondarenko, M. Fröbe, M. Beloucif, L. Gienapp, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2020: Argument Retrieval, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 11th International Conference of the CLEF Association (CLEF 2020), volume 12260 of Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2020, pp. 384–395. URL: https://link.springer.com/chapter/10.1007/978-3-030-58219-7_26. doi:10.1007/978-3-030-58219-7_26.
- A. Bondarenko, L. Gienapp, M. Fröbe, M. Beloucif, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2021: Argument Retrieval, in: K. Candan, B. Ionescu, L. Goeuriot, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the CLEF Association (CLEF 2021), volume 12880 of Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2021, pp.450–467. URL: https://link.springer.com/chapter/10.1007/978-3-030-85251-1_28. doi:10.1007/978-3-030-85251-1_28.
- T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, Ms marco: A human generated machine reading comprehension dataset, in: CoCo@ NIPS, 2016.
- R. Pradeep, R. Nogueira, J. Lin, The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models, 2021. URL: https://arxiv.org/abs/2101.05667. doi:10.48550/ARXIV.2101.05667.

# References


- R. Nogueira, J. Lin, A. Epistemic, From doc2query to doctttttquery, Online preprint 6 (2019).
- R. Pradeep, R. Nogueira, J. Lin, The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models, 2021. URL: https://arxiv.org/abs/2101.05667. doi:10.48550/ARXIV.2101.05667.
- X. Zeng, A. Zubiaga, Qmul-sds at sciver: Step-by-step binary classification for scientific claim verification, arXiv preprint arXiv:2104.11572 (2021).
- Rana, A., Khanna, D., Singh, M., Ghosal, T., Singh, H. and Rana, P.S., 2022. RerrFact: Reduced Evidence Retrieval Representations for Scientific Claim Verification. arXiv preprint arXiv 2202.02646.