LEVI RANK: 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

Data and Web Science Group, University of Mannheim, Germany
Lappeenranta-Lahti University of Technology LUT, Finnland
## Table of Contents

- **Introduction**: Motivation, Problem & Related Work
- **Approach Overview**: Task Introduction & Architecture Pipeline
  - **Initial Retrieval**: Approaches Explored, Initial Results & Findings
  - **Multi-stage Re-ranking**: Approaches Explored, Initial Results & Findings
  - **Stance Prediction**: Approaches Explored, Initial Results & Findings
- **Result Discussion**: Result Report & Discussion Summary
- **Conclusion**: Future Improvements & Conclusion
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
**LEVI\textsc{RANK}: System Architecture**

**A. Larger documents (\(\geq 512\) 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

**Initial Retrieval: Result Findings**

<table>
<thead>
<tr>
<th>Retrieval Approach</th>
<th>Recall@1000</th>
<th>Recall@1500</th>
<th>Recall@2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25 Baseline</td>
<td>90.18</td>
<td>90.67</td>
<td>91.11</td>
</tr>
<tr>
<td>Dense Retrieval</td>
<td>85.70</td>
<td>86.56</td>
<td>87.56</td>
</tr>
<tr>
<td>Pseudo-Relevance Feedback</td>
<td>89.98</td>
<td>90.59</td>
<td>91.07</td>
</tr>
<tr>
<td>LEVI RANK Voting</td>
<td>90.14</td>
<td>91.08</td>
<td>91.17</td>
</tr>
</tbody>
</table>

**Approaches Implemented & Tested**
- Previous Baselines: TF-IDF
- Probabilistic Approaches: BM25, BM25 + Pseudo-Relevance Feedback, LEVI RANK
- 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 < LEVI RANK Voting]
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

<table>
<thead>
<tr>
<th>Ranking Approach</th>
<th>BM25</th>
<th>monoT5-only</th>
<th>monoT5-duoT5</th>
</tr>
</thead>
<tbody>
<tr>
<td>nDCG@5</td>
<td>0.33</td>
<td>0.47</td>
<td>0.31</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Approach</th>
<th>No object</th>
<th>Neutral</th>
<th>Object 1</th>
<th>Object 2</th>
<th>Macro-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bondarenko et al. (2022)</td>
<td>0.40</td>
<td>0.53</td>
<td>0.72</td>
<td>0.63</td>
<td>0.57</td>
</tr>
<tr>
<td>LEVI-RANK</td>
<td>0.40</td>
<td>0.52</td>
<td>0.72</td>
<td>0.68</td>
<td>0.58</td>
</tr>
</tbody>
</table>
Leaderboard Result Summary

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

<table>
<thead>
<tr>
<th>Submitted Approaches</th>
<th>Recall@2K</th>
<th>Input Size for duoT5</th>
<th>nDCG@5 Relevance</th>
<th>nDCG@5 Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCT-ColBERT+monoT5+duoT5</td>
<td>92.05</td>
<td>100</td>
<td>0.758 (1)</td>
<td>0.744 (2)</td>
</tr>
<tr>
<td>BM25+monoT5+duoT5</td>
<td>98.23</td>
<td>100</td>
<td>0.755</td>
<td>0.742</td>
</tr>
<tr>
<td>LEVIANK+PR+monoT5+duoT5</td>
<td>97.96</td>
<td>50</td>
<td>0.753</td>
<td>0.730</td>
</tr>
<tr>
<td>LEVIANK+monoT5</td>
<td>98.34</td>
<td>0</td>
<td>0.727</td>
<td>0.706</td>
</tr>
<tr>
<td>Pseudo-Relevance(PR)+monoT5</td>
<td>97.16</td>
<td>0</td>
<td>0.722</td>
<td>0.695</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Training Approach</th>
<th>Prediction Annotation Set</th>
<th>Macro-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-shot Two-Step RoBERTa-MNLI</td>
<td>Whole stance dataset</td>
<td>0.303 (2)</td>
</tr>
<tr>
<td>Zero-shot Two-Step RoBERTa-MNLI</td>
<td>Worst 50 % topic queries</td>
<td>0.116 (6)</td>
</tr>
<tr>
<td>Zero-shot Two-Step RoBERTa-MNLI (fine tuned, 50 % best queries)</td>
<td>Worst 50 % topic queries</td>
<td>0.387 (1)</td>
</tr>
</tbody>
</table>

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?

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.
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 LEVI RANK 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 (Rana et al. 2022)
- Encouraging retrieving reduced document representations
- Better error analysis, for better decoding of model failures

Results Caveats

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

Thank you, Looking forward to discuss!
References

- E. Turner, L. Rainie, Most americans rely on their own research to make big decisions, and that often means online searches (2020).


References


References


