Stance-Aware Re-Ranking for Non-factual Comparative Queries

December 7, 2023

Jan Heinrich Reimer
Alexander Bondarenko
Maik Fröbe
Matthias Hagen

Friedrich-Schiller-Universität Jena
https://webis.de
Which tower is taller, the Eiffel Tower or the Big Ben?
Stance-Aware Re-Ranking for Non-factual Comparative Queries

Factual comparisons

Which tower is taller, the Eiffel Tower or the Big Ben?

→ comparison based on facts, no arguments required
Stance-Aware Re-Ranking for Non-factual Comparative Queries

Non-factual comparisons

Which city is better, Paris or London?

Source: https://netivist.org/debate/paris-vs-london
Non-factual comparisons

Source: https://netivist.org/debate/paris-vs-london

Which city is better, Paris or London?

→ choice depends on personal opinions, arguments required
Stance-Aware Re-Ranking for Non-factual Comparative Queries

Argument retrieval

Existing argument search engines

- Examples: args.me [Wachsmuth et al. EMNLP’17] or ArgumenText [Stab et al. NAACL-HLT’18]

- Goal: Relevance to query, presence/quality of arguments

- Often focus on controversial topics, e.g., ‘nuclear energy’

→ No search engine focused on comparative topics

Touché shared tasks [Bondarenko et al. CLEF’22]

- Queries compares two options (e.g., Paris vs. London)

- Goal: Arguments for/against either option

- Approaches: BM25, neural (re-)ranking, arg. quality

→ Stance not exploited by Touché participants
Stance-Aware Re-Ranking for Non-factual Comparative Queries

Argument retrieval

Existing argument search engines

- Examples: args.me [Wachsmuth et al. EMNLP’17] or ArgumenText [Stab et al. NAACL-HLT’18]

- Goal: Relevance to query, presence/quality of arguments

- Often focus on controversial topics, e.g., ‘nuclear energy’
  → No search engine focused on comparative topics

Touché shared tasks [Bondarenko et al. CLEF’22]

- Queries compares two options (e.g., Paris vs. London)

- Goal: Arguments for/against either option

- Approaches: BM25, neural (re-)ranking, arg. quality
  → Stance not exploited by Touché participants
Stance-Aware Re-Ranking for Non-factual Comparative Queries

Re-ranking based on the stance

Stance for comparative questions (e.g. ‘Which city is better, Paris or London?’)

- pro London: I like the popular West End musicals.
- pro Paris: To me, Paris is the most romantic town in the world.
- both equal: My favorite art museums are in London and Paris.
- no stance: London has 9M inhabitants, and Paris has 10M.

Our approach

- Intuition: Helpful results should express a stance towards the options
- Idea: Move results without stance to the end of the ranking
- Evaluation: Re-rank top-5, measure retrieval effectiveness (nDCG@5)

<table>
<thead>
<tr>
<th>Rank</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FACT</td>
<td>London</td>
<td>FACT</td>
<td>both</td>
<td>FACT</td>
</tr>
<tr>
<td>Original run</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our re-ranking</td>
<td>London</td>
<td>both</td>
<td></td>
<td>FACT</td>
<td>FACT</td>
</tr>
</tbody>
</table>
Stance-Aware Re-Ranking for Non-factual Comparative Queries

Re-ranking based on the stance

Stance for comparative questions (e.g. ‘Which city is better, Paris or London?’)

<table>
<thead>
<tr>
<th>pro London</th>
<th>pro Paris</th>
<th>both equal</th>
<th>no stance</th>
</tr>
</thead>
<tbody>
<tr>
<td>I like the popular West End musicals.</td>
<td>To me, Paris is the most romantic town in the world.</td>
<td>My favorite art museums are in London and Paris.</td>
<td>London has 9M inhabitants, and Paris has 10M.</td>
</tr>
</tbody>
</table>

Our approach

- **Intuition:** Helpful results should express a stance towards the options
- **Idea:** Move results without stance to the end of the ranking
- **Evaluation:** Re-rank top-5, measure retrieval effectiveness (nDCG@5)

<table>
<thead>
<tr>
<th>Rank</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original run</td>
<td>FACT</td>
<td>London</td>
<td>FACT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our re-ranking</td>
<td></td>
<td>London</td>
<td></td>
<td>FACT</td>
<td>FACT</td>
</tr>
</tbody>
</table>
Stance-Aware Re-Ranking for Non-factual Comparative Queries

Re-ranking scenarios

1. “Perfect” stance labels
   - Touché ground truth labels (artificial scenario)
   - Results: Significantly improved nDCG@5 for all runs, close to optimal re-rank.

2. Predicted stance (Touché participants)
   - Use Touché participants’ stance predictions (real-world scenario)
   - Results: No improvements, stance detection not effective enough ($F_1 \leq 0.31$)

3. Improved stance detection (ours)
   - RoBERTa ($F_1=0.34$), Flan-T5 (zero-shot, $F_1=0.39$), GPT-3.5 (4-shot, $F_1=0.49$)
   - Results: improved nDCG@5, top-3 runs would reach rank 1 in leaderboard

4. Simulated stance detection (perturbed ground truth)
   - Stopping thresholds: $F_1 = 0.95, 0.90, \ldots, 0.20$
   - Results: higher nDCG@5 improv. with higher $F_1$, top runs “harder” to improve (higher $F_1$)
Stance-Aware Re-Ranking for Non-factual Comparative Queries

Summary

- Simple, yet effective re-ranking approach for non-factual comparative queries
- Significantly improves all Touché 2022 runs with “perfect” stance labels
- Still changes the top-3 system order with our GPT-3.5 stance detection
- Future work: improving stance detection, separate result lists per stance

Code and Data

🔗 github.com/webis-de/ArgMining-23

Contributions are welcome!
Stance-Aware Re-Ranking for Non-factual Comparative Queries

Summary

- Simple, yet effective re-ranking approach for non-factual comparative queries
- Significantly improves all Touché 2022 runs with “perfect” stance labels
- Still changes the top-3 system order with our GPT-3.5 stance detection
- Future work: improving stance detection, separate result lists per stance

Code and Data

- [github.com/webis-de/ArgMining-23](https://github.com/webis-de/ArgMining-23)
- Contributions are welcome!

Thank you!