Webis at TREC 2021: Deep Learning Track

November 16, 2021

Maik Fröbe
Sebastian Günther
Matthias Hagen

Martin-Luther-Universität Halle-Wittenberg
first.last@informatik.uni-halle.de

@webis_de www.webis.de
Webis at TREC 2021: Deep Learning Track

Overview

- Document ranking: Three feature-based LTR runs with LambdaMART
  - No deep learning, traditional approach/baseline

- Focus: Anchor text features
  - Extracted from a Common Crawl Snapshot

- Literature:
  - Anchor text useful ranking feature for certain types of queries
    [Craswell et al., SIGIR’01, Koolen and Kamps, SIGIR’10]
  - Anchor text useful for (pre-) training of models
    [Ma et al., CIKM’21, Dai and Callan, WWW’20]

- Research Question:
  - Should you use anchor text as feature or training data?
Extraction of Anchor Text

Results (DL’21)

- MS MARCO has sparse link structure
- We process the Common Crawl 2021-04 (3.40 b documents)
- After Filtering + Sampling:
  - 85 m anchors pointing to 3.20 m documents
Extraction of Anchor Text
Results (DL’21)

- MS MARCO has sparse link structure
- We process the Common Crawl 2021-04 (3.40 b documents)
- After Filtering + Sampling:
  - 85 m anchors pointing to 3.20 m documents

Effectiveness on MS MARCO v1

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Retrieval</th>
<th>nDCG@10 TREC DL 19 (judged only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor</td>
<td>BM25</td>
<td>0.41</td>
</tr>
<tr>
<td>Content</td>
<td>BM25</td>
<td><strong>0.51</strong></td>
</tr>
<tr>
<td>ORCAS</td>
<td>BM25</td>
<td>0.45</td>
</tr>
</tbody>
</table>
Extraction of Anchor Text

Results (DL’21)

- MS MARCO has sparse link structure
- We process the Common Crawl 2021-04 (3.40 b documents)
- After Filtering + Sampling:
  - 85 m anchors pointing to 3.20 m documents

Effectiveness on MS MARCO v1

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Retrieval</th>
<th>nDCG@10 TREC DL 19 (judged only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor</td>
<td>BM25</td>
<td>0.41</td>
</tr>
<tr>
<td>Content</td>
<td>BM25</td>
<td>0.51</td>
</tr>
<tr>
<td>ORCAS</td>
<td>BM25</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Results (Now)

- We extracted more anchor texts
  - 6 Common Crawl snapshots
  - Version 1 and 2 of MS MARCO
Overview Features

50 Features

- 36 Query-Document features calculated with Anserini
  - 9 scores (BM25, etc.) for 4 fields (Anchors, Title, Body, URL)
- 8 Document features (e.g., Alexa Rank)
- 6 Query features (e.g., length)

Feature Importance
Submissions and Results

Submissions

- We rerank the top-100 results
- Comparison LambdaMART with anchor features vs. without

Results

<table>
<thead>
<tr>
<th>Features</th>
<th>Trees</th>
<th>MRR@10</th>
<th>nDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Anchors</td>
<td>5,000</td>
<td>0.9356</td>
<td>0.5831</td>
</tr>
<tr>
<td>Without Anchors</td>
<td>5,000</td>
<td>0.9396</td>
<td>0.5747</td>
</tr>
<tr>
<td>With Anchors</td>
<td>1,000</td>
<td><strong>0.9488</strong></td>
<td><strong>0.5918</strong></td>
</tr>
<tr>
<td>top-100 baseline</td>
<td>—</td>
<td>0.8367</td>
<td>0.5116</td>
</tr>
</tbody>
</table>

- Interpretation: Our features are precision-oriented
  - Many documents have no anchors, no PageRank, etc.
  - Anchor text adds not much effectiveness
Conclusions

Summary

- Focus: extraction of anchor text features
- Anchor text did not substantially improve the effectiveness
- The extracted anchor text is available:
  
github.com/webis-de/ecir22-anchor-text
Conclusions

Summary

- Focus: extraction of anchor text features
- Anchor text did not substantially improve the effectiveness
- The extracted anchor text is available: 
  
  github.com/webis-de/ecir22-anchor-text

Future work

- Include anchor text into ir_datasets
- Study the parallel dataset: queries + anchor text
Conclusions

Summary

- Focus: extraction of anchor text features
- Anchor text did not substantially improve the effectiveness
- The extracted anchor text is available:
  
  github.com/webis-de/ecir22-anchor-text

Future work

- Include anchor text into ir_datasets
- Study the parallel dataset: queries + anchor text

thank you!
Webis @ TREC 2021: Health Misinformation Track

November 18, 2021

Martin-Luther-Universität Halle-Wittenberg
first.last@informatik.uni-halle.de

*Bauhaus-Universität Weimar
first.last@uni-weimar.de

@webis_de   www.webis.de
Webis @ TREC 2021: Health Misinformation Track

Overview

- **Task**: Given a search topic about potential treatment, retrieve from the C4 dataset documents that are credible and helpful.
- **Runs**: 2 initial rankings and 4 re-rankings.
Overview

- **Task:** Given a search topic about potential treatment, retrieve from the C4 dataset documents that are credible and helpful.
- **Runs:** 2 initial rankings and 4 re-rankings.

Initial rankings:
- Anserini’s BM25 (default $k=0.9$, $b=0.4$), removing stop words.
- Top 50 re-ranked with MonoT5 (PyGaggle’s monot5-base-msmarco).
Task: Given a search topic about potential treatment, retrieve from the C4 dataset documents that are credible and helpful.

Runs: 2 initial rankings and 4 re-rankings.

Initial rankings:
- Anserini’s BM25 (default k=0.9, b=0.4), removing stop words.
- Top 50 re-ranked with MonoT5 (PyGaggle’s monot5-base-msmarco).

Re-ranking:
- Using the argumentative axiomatic pipeline from previous years.
  [Bondarenko et al., TREC’18; Bondarenko et al., TREC’19; Bevendorff et al., TREC’20]
- Inspiration: Axiomatic IR—constraints a good retrieval model should fulfill.
Webis @ TREC 2021: Health Misinformation Track

Overview

- Task: Given a search topic about potential treatment, retrieve from the C4 dataset documents that are credible and helpful.
- Runs: 2 initial rankings and 4 re-rankings.

Initial rankings:
- Anserini’s BM25 (default k=0.9, b=0.4), removing stop words.
- Top 50 re-ranked with MonoT5 (PyGaggle’s monot5-base-msmarco).

Re-ranking:
- Using the argumentative axiomatic pipeline from previous years. [Bondarenko et al., TREC’18; Bondarenko et al., TREC’19, Bevendorff et al., TREC’20]
- Inspiration: Axiomatic IR—constraints a good retrieval model should fulfill.

Applied for argumentative queries (might benefit from arguments in documents) like “Should I apply ice to a burn?”. 
Argumentative Axiomatic Re-Ranking

Argumentative Units in Text

Simplified argumentative unit in a text document consists of:

- Premise $p_1$: Long-haired cats shed all over the house
- Premise $p_2$: Long-haired cats have a lot of fleas

Example:

Cats with long hair shed all over the house. I have heard that they also have lots of fleas. We should not get a long-haired cat.
Argumentative Axiomatic Re-Ranking

Argumentative Units in Text

Simplified argumentative unit in a text document consists of:

- **Premise** $p_1$: Long-haired cats shed all over the house
- **Premise** $p_2$: Long-haired cats have a lot of fleas
- **Conclusion** $c$: We should not get a long haired cat

Example:

Cats with long hair shed all over the house. I have heard that they also have lots of fleas. We should not get a long-haired cat.
Argumentative Axiomatic Re-Ranking
Argumentative Units Identification

TARGER
API Publication Source

Argument Tagger
Some text which is not an argument at all. Cats with long hair shed all over the house. I have heard that they also have lots of fleas. We should not get a long-haired cat. Some text which is not an argument at all.

Model to label with
Combined dataset, fast1

Analyze

Argument Labels
- PREMISE
- CLAIM

Entity Labels
- PERSON
- PER
- NORP
- FACILITY
- ORG
- GPE
- LOC
- PRODUCT
- EVENT
+ more labels

Some text which is not an argument at all. Cats with long hair shed all over the house PREMISE. I have heard PREMISE that they also have lots of fleas PREMISE. We should not get a long-haired cat CLAIM. Some text which is not an argument at all.

https://demo.webis.de/targer/  [Chernodub et al.; ACL'19]
Argumentative Axiomatic Re-Ranking

Axiom: ArgUC Argumentative Units Count

ArgUC  Favor documents which contain more argumentative units.

Given:

- Query $Q$
- Documents $D_1, D_2$ with $|D_1| = |D_2|$
- $Arg_D$: set of argumentative units of a document $D$
Argumentative Axiomatic Re-Ranking

Axiom: ArgUC Argumentative Units Count

ArgUC Favor documents which contain more argumentative units.

Given:

- Query $Q$
- Documents $D_1, D_2$ with $|D_1| \approx_{10\%} |D_2|$
- $Arg_D$: set of argumentative units of a document $D$
Argumentative Axiomatic Re-Ranking

Axiom: ArgUC Argumentative Units Count

ArgUC    Favor documents which contain more argumentative units.

Given:

- Query $Q$
- Documents $D_1, D_2$ with $|D_1| \approx 10\% |D_2|$
- $Arg_D$: set of argumentative units of a document $D$

IF $count(Arg_{D_1}) > count(Arg_{D_2})$ THEN $rank(D_1, Q) > rank(D_2, Q)$
Argumentative Axiomatic Re-Ranking

Axiom: QTArg Query Term Occurrence in Argumentative Units

QTArg Favor documents with the query terms close to argumentative units.

Given:

- One-term query $Q = \{q\}$
- Documents $D_1, D_2$ with $|D_1| \approx_{10\%} |D_2|$
- $Arg_D$: set of argumentative units of a document $D$

---

Alexander Bondarenko
Argumentative Axiomatic Re-Ranking
Axiom: QTArg Query Term Occurrence in Argumentative Units

QTArg  Favor documents with the query terms close to argumentative units.

Given:
- One-term query $Q = \{q\}$
- Documents $D_1, D_2$ with $|D_1| \approx 10\% \ |D_2|$
- $Arg_D$: set of argumentative units of a document $D$

IF $q \in A_{D_1}$ for some $A_{D_1} \in Arg_{D_1}$ but $q \notin A_{D_2}$ for all $A_{D_2} \in Arg_{D_2}$

THEN $rank(D_1, Q) > rank(D_2, Q)$
Argumentative Axiomatic Re-Ranking

Axiom: QTPArg Query Term Position in Argumentative Units

QTPArg  Favor documents where the first appearance of a query term in an argumentative unit is closer to the beginning of the document.

[Troy, Zhang, SIGIR’07; Mitra, Diaz, Craswell, WWW’17]

Given:

- One-term query \( Q = \{q\} \)
- Documents \( D_1, D_2 \) with \(|D_1| \approx 10\% |D_2|\)
- \( 1^{st} \) position\((q, Arg_D)\): first position in an argumentative unit of document \( D \) where the term \( q \) appears

\[
\begin{array}{c}
Q & \text{Arg} \\
D_1 & \text{Arg} \\
D_2 & \text{Arg}
\end{array}
\]
Argumentative Axiomatic Re-Ranking

Axiom: QTPArg Query Term Position in Argumentative Units

QTPArg Favor documents where the first appearance of a query term in an argumentative unit is closer to the beginning of the document.

[Troy, Zhang, SIGIR’07; Mitra, Diaz, Craswell, WWW’17]

Given:

- One-term query $Q = \{q\}$
- Documents $D_1, D_2$ with $|D_1| \approx 10\% |D_2|$
- $1^{st} position(q, Arg_D)$: first position in an argumentative unit of document $D$ where the term $q$ appears

IF $1^{st} position(q, Arg_{D_1}) < 1^{st} position(q, Arg_{D_2})$ THEN $\text{rank}(D_1, Q) > \text{rank}(D_2, Q)$
Argumentative Axiomatic Re-Ranking

Pipeline

- Retrieve initial top 20 results with some model.

Diagram:

1. Corpus
   Retrieval model R
   +
   Query
   1
   .
   .
   k
Argumentative Axiomatic Re-Ranking

Pipeline

- Retrieve initial top 20 results with some model.
- Calculate pairwise preferences of the argumentative axioms.
- Preferences: swap (or not) document positions in the ranking.
- Aggregate **weighted** re-ranking preferences.
Argumentative Axiomatic Re-Ranking

Pipeline

1. Retrieve initial top 20 results with some model.
2. Calculate pairwise preferences of the argumentative axioms.
3. Preferences: swap (or not) document positions in the ranking.
4. Aggregate weighted re-ranking preferences.
5. Re-rank the initial top 20 retrieved results.
Argumentative Axiomatic Re-Ranking

Results

- **ax1**: at least 1 axiom decides to swap document positions.

- **ax3**: all 3 axioms decide to swap document positions.

Results as provided by the organizers. Abbreviations: U: useful, Co: Correct, Cr: credible. Incor: incorrect).

<table>
<thead>
<tr>
<th>Run</th>
<th>Compatibility</th>
<th>nDCG (binary)</th>
<th>P@10 (binary)</th>
<th>nDCG (graded)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Help</td>
<td>Harm</td>
<td>U/Co</td>
<td>U/Cr</td>
</tr>
<tr>
<td>webis-bm25 (initial)</td>
<td>0.1292</td>
<td>0.1454</td>
<td>0.4275</td>
<td>0.4856</td>
</tr>
<tr>
<td>webis-bm25-ax1</td>
<td>0.1339</td>
<td>0.1474</td>
<td><strong>0.4325</strong></td>
<td><strong>0.4877</strong></td>
</tr>
<tr>
<td>webis-bm25-ax3</td>
<td>0.1318</td>
<td>0.1445</td>
<td>0.4285</td>
<td>0.4859</td>
</tr>
<tr>
<td>webis-t5 (initial)</td>
<td>0.1314</td>
<td>0.1447</td>
<td>0.2383</td>
<td>0.2618</td>
</tr>
<tr>
<td>webis-t5-ax1</td>
<td>0.1297</td>
<td>0.1449</td>
<td>0.2362</td>
<td>0.2645</td>
</tr>
<tr>
<td>webis-t5-ax3</td>
<td>0.1327</td>
<td><strong>0.1438</strong></td>
<td>0.2392</td>
<td>0.2632</td>
</tr>
</tbody>
</table>
Argumentative Axiomatic Re-Ranking

Results

Soboroff 2021. TREC Overview.
Argumentative Axiomatic Re-Ranking

Summary

- Axiom-based re-ranking framework for *any* retrieval model.
- Directly incorporating axiomatic “thinking” in the retrieval process.
- Axioms are easy to understand / rankings are more explainable.

WIP

- Argumentative query classification.
- New axioms capturing further different angles of argumentativeness.
- Improve the weighting scheme through large-scale training.
- Better detect (define) argumentative units.
Argumentative Axiomatic Re-Ranking

Summary

- Axiom-based re-ranking framework for any retrieval model.
- Directly incorporating axiomatic “thinking” in the retrieval process.
- Axioms are easy to understand / rankings are more explainable.

WIP

- Argumentative query classification.
- New axioms capturing further different angles of argumentativeness.
- Improve the weighting scheme through large-scale training.
- Better detect (define) argumentative units.

thank you!
Bibliography


M. Hagen, M. Völske, S Göring, B. Stein. Axiomatic Result Re-Ranking. *In CIKM 2016.*

M. Markel. Technical Communication. 9th ed. *Bedford/St Martin’s (2010).*


Webis at TREC 2021: Podcasts Track

November 17, 2021

Marcel Gohsen¹, Johannes Kiesel¹, Jakob Schwerter², Shahbaz Syed², Martin Potthast², Benno Stein¹

Bauhaus-Universität Weimar¹, Leipzig University²

@webis_de www.webis.de
Webis at TREC 2021: Podcasts Track
Retrieval Task

- Four runs for podcast retrieval, all with BM25
- Classification for re-ranking
  - SVM trained on own annotations (Entertaining, Subjective, Discussion)
  - Multiplying confidence with BM25 score

### Runs

<table>
<thead>
<tr>
<th>Run</th>
<th>Criterion</th>
<th>nDCG@30</th>
<th>nDCG@1000</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>webis_pc_bs</td>
<td>Entertaining</td>
<td>0.1182</td>
<td>0.2330</td>
<td>0.0975</td>
</tr>
<tr>
<td></td>
<td>cola</td>
<td>0.0522</td>
<td>0.1748</td>
<td>0.0450</td>
</tr>
<tr>
<td></td>
<td>rob</td>
<td>0.0351</td>
<td>0.1584</td>
<td>0.0275</td>
</tr>
<tr>
<td></td>
<td>co_rob</td>
<td>0.0332</td>
<td>0.1620</td>
<td>0.0275</td>
</tr>
<tr>
<td>webis_pc_cola</td>
<td>Subjective</td>
<td>0.1725</td>
<td>0.3435</td>
<td>0.2000</td>
</tr>
<tr>
<td></td>
<td>cola</td>
<td>0.0591</td>
<td>0.2443</td>
<td>0.0600</td>
</tr>
<tr>
<td></td>
<td>rob</td>
<td>0.0371</td>
<td>0.2250</td>
<td>0.0350</td>
</tr>
<tr>
<td></td>
<td>co_rob</td>
<td>0.0430</td>
<td>0.2320</td>
<td>0.0550</td>
</tr>
<tr>
<td>webis_pc_rob</td>
<td>Discussion</td>
<td>0.1619</td>
<td>0.3208</td>
<td>0.1600</td>
</tr>
<tr>
<td></td>
<td>cola</td>
<td>0.0598</td>
<td>0.2289</td>
<td>0.0625</td>
</tr>
<tr>
<td></td>
<td>rob</td>
<td>0.0399</td>
<td>0.2101</td>
<td>0.0400</td>
</tr>
<tr>
<td></td>
<td>co_rob</td>
<td>0.0475</td>
<td>0.2193</td>
<td>0.0550</td>
</tr>
<tr>
<td>webis_pc_co_rob</td>
<td>both concatenated</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Webis at TREC 2021: Podcasts Track
Summarization Task

- Two runs: abstractive and extractive
- Using the entertainment ranking from the combined model

Runs

**webis_pc_abstr:**
DistilBART abstractive summarization
Input: 5 most entertaining sentences + their 5 previous and following ones

**webis_pc_extr:**
TextRank extractive summarization
Output: 10 sentences with highest entertainment-biased TextRank

<table>
<thead>
<tr>
<th>Run</th>
<th>EGFB score</th>
<th>E</th>
<th>G</th>
<th>F</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>abstr</td>
<td>0.2332</td>
<td>0</td>
<td>6</td>
<td>33</td>
<td>154</td>
</tr>
<tr>
<td>extr</td>
<td>0.2604</td>
<td>0</td>
<td>6</td>
<td>38</td>
<td>148</td>
</tr>
<tr>
<td>Baseline (one-minute)</td>
<td><strong>0.8083</strong></td>
<td>7</td>
<td>26</td>
<td>76</td>
<td>84</td>
</tr>
</tbody>
</table>