

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 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

Corpus	Retrieval	nDCG@10 TREC DL 19 (judged only)
Anchor	BM25	0.41
Content	BM25	0.51
ORCAS	BM25	0.45

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

Corpus	Retrieval	nDCG@10 TREC DL 19 (judged only)
Anchor	BM25	0.41
Content	BM25	0.51
ORCAS	BM25	0.45

Results (Now)

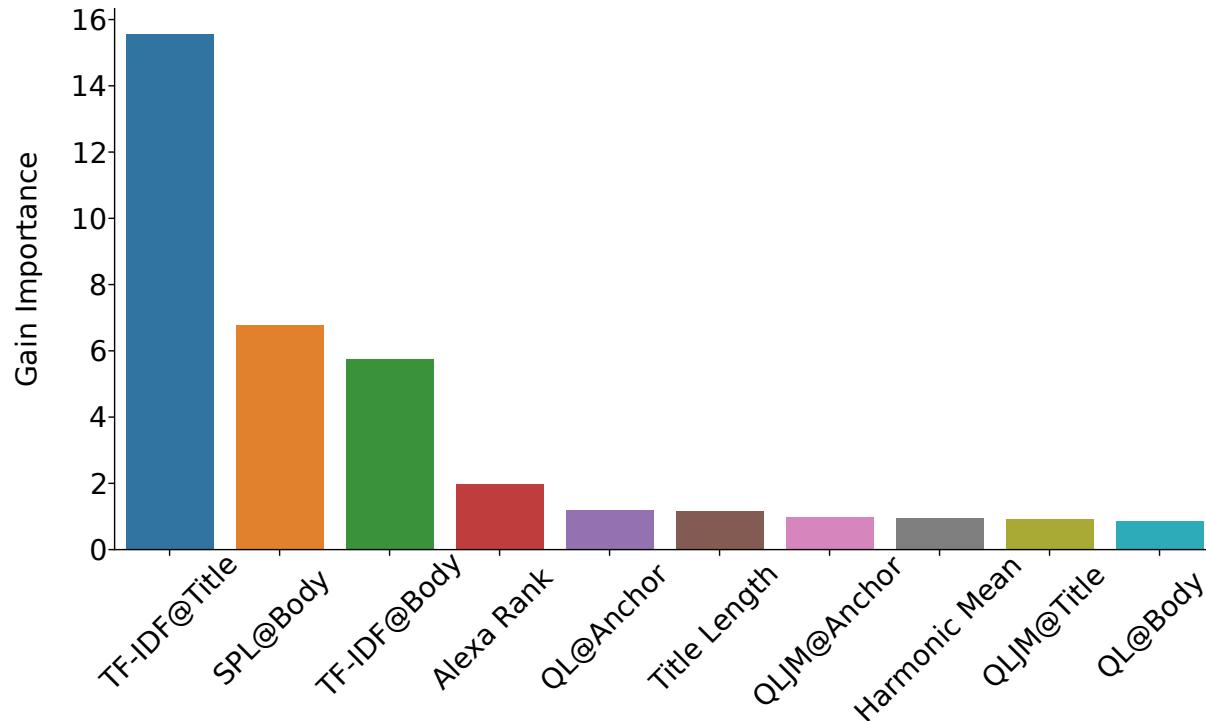
- ❑ We extracted more anchor texts
 - 6 Common Crawl snapshots
 - Version 1 and 2 of MS MARCO

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

Features	Trees	MRR@10	nDCG@10
With Anchors	5,000	0.9356	0.5831
Without Anchors	5,000	0.9396	0.5747
With Anchors	1,000	0.9488	0.5918
top-100 baseline	—	0.8367	0.5116

- 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



Alexander
Bondarenko



Maik
Fröbe



Michael
Völske*



Matthias
Hagen

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.

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).

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.

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

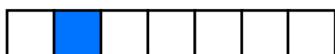
ArgUC Favor documents which contain more argumentative units.

Given:

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

Q  Arg 

D_1 

D_2 

Argumentative Axiomatic Re-Ranking

Axiom: ArgUC Argumentative Units Count

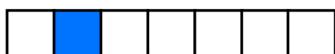
ArgUC Favor documents which contain more argumentative units.

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

Q  Arg 

D_1 

D_2 

Argumentative Axiomatic Re-Ranking

Axiom: ArgUC Argumentative Units Count

ArgUC Favor documents which contain more argumentative units.

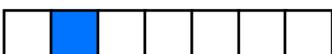
Given:

- 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 $rank(D_1, Q) > rank(D_2, Q)$

Q  Arg 

D_1 

D_2 

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

Q 

Arg 

D_1  A horizontal sequence of seven boxes. The second box from the left is filled blue, representing the query term q . The other six boxes are white, representing argumentative units.

D_2  A horizontal sequence of seven boxes. The first box is filled blue, representing the query term q . The other six boxes are white, representing argumentative units.

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 $rank(D_1, Q) > rank(D_2, Q)$

Q  Arg 

D_1 

D_2 

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

Q 

Arg 

D_1 

D_2 

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

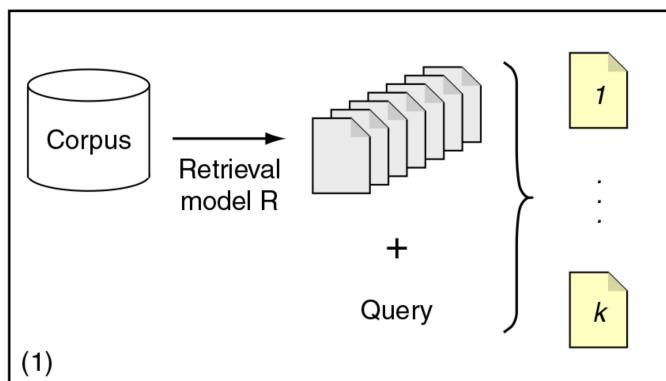
IF $1^{st}position(q, Arg_{D_1}) < 1^{st}position(q, Arg_{D_2})$ THEN $rank(D_1, Q) > rank(D_2, Q)$

Q 

Arg 



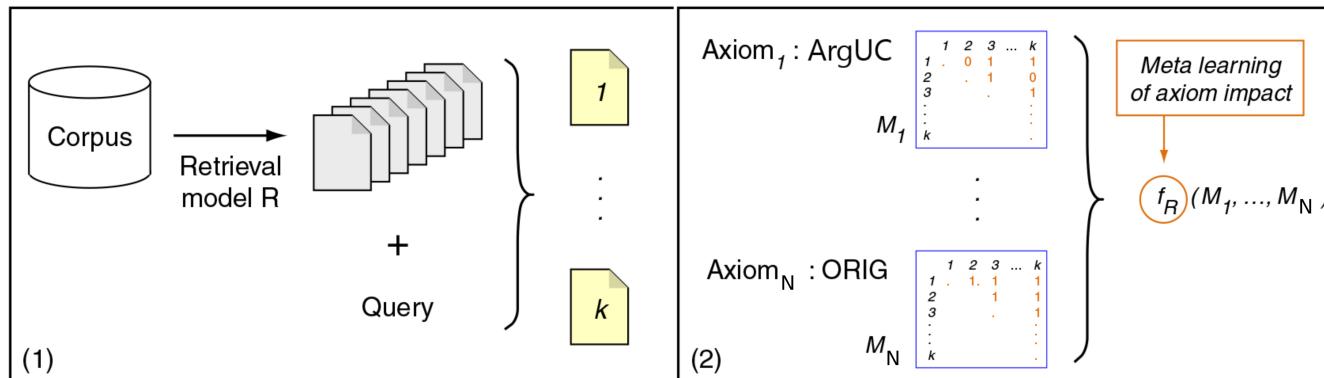
Argumentative Axiomatic Re-Ranking Pipeline



- Retrieve initial top 20 results with some model.

Argumentative Axiomatic Re-Ranking

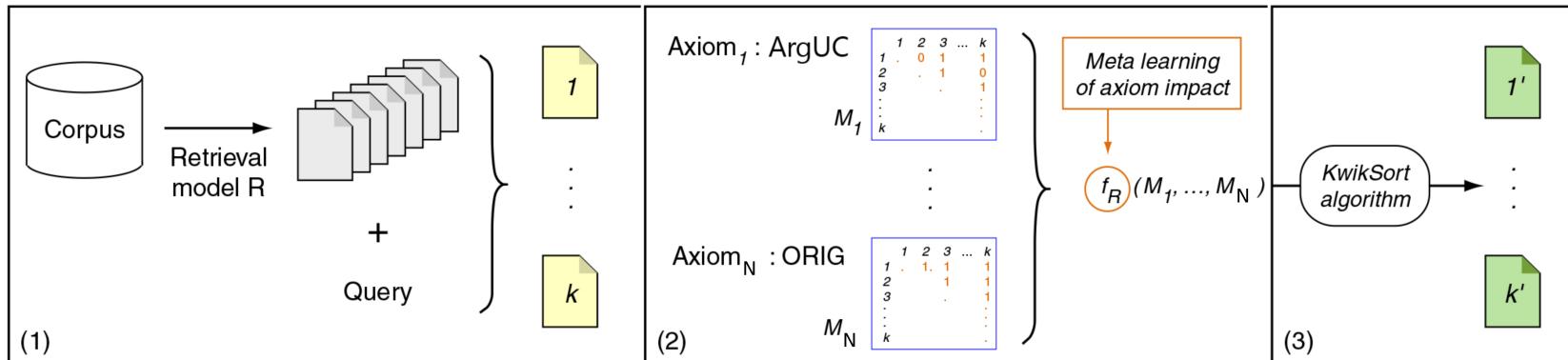
Pipeline



- ❑ 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



- 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.
- 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.

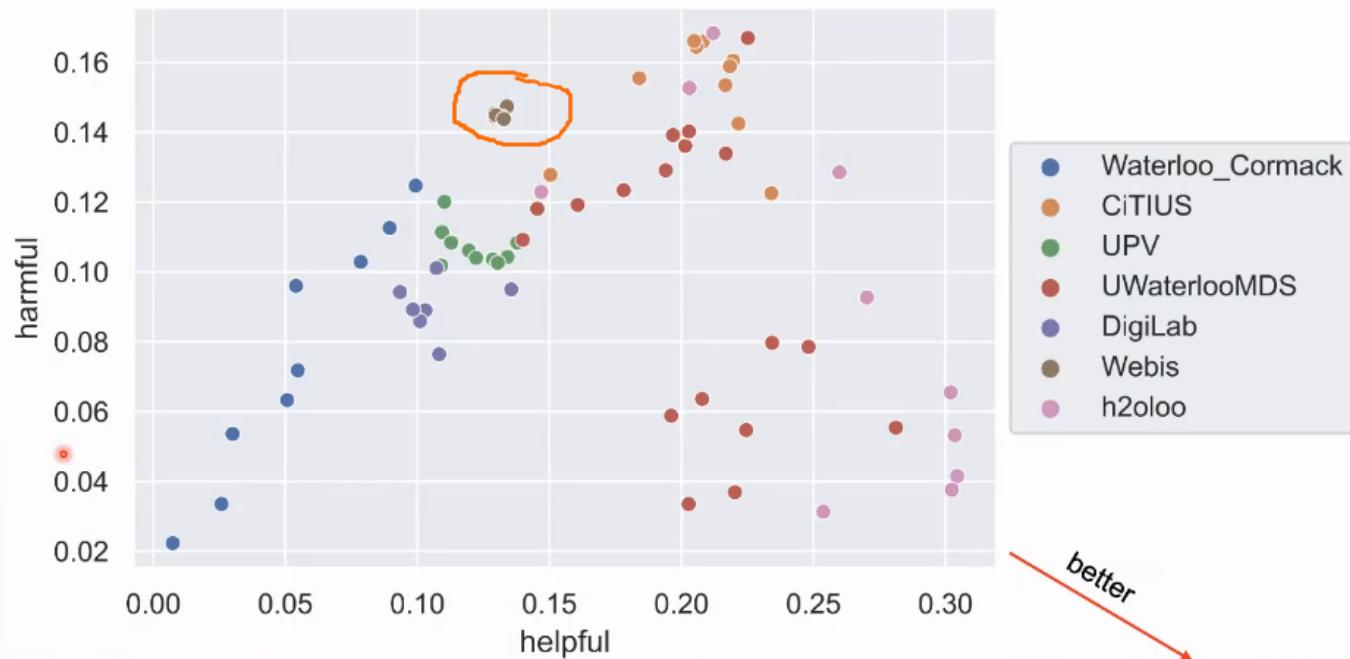
Results as provided by the organizers. Abbreviations: U: useful, Co: Correct, Cr: credible. Incor: incorrect).

Run	Compatibility		nDCG (binary)			P@10 (binary)		nDCG (graded)
	Help	Harm	U/Co	U/Cr	U/Co/Cr	U/Co	Incor.	Useful
webis-bm25 (initial)	0.1292	0.1454	0.4275	0.4856	0.3796	0.3088	0.2906	0.5809
webis-bm25-ax1	0.1339	0.1474	0.4325	0.4877	0.3880	0.3088	0.2844	0.5807
webis-bm25-ax3	0.1318	0.1445	0.4285	0.4859	0.3802	0.3088	0.2844	0.5810
webis-t5 (initial)	0.1314	0.1447	0.2383	0.2618	0.1912	0.3235	0.2969	0.3400
webis-t5-ax1	0.1297	0.1449	0.2362	0.2645	0.1896	0.3471	0.3344	0.3407
webis-t5-ax3	0.1327	0.1438	0.2392	0.2632	0.1907	0.3412	0.3344	0.3410

Argumentative Axiomatic Re-Ranking

Results

Ad Hoc Retrieval Task



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

- N. Ailon, M. Charikar, A. Newman. Aggregating Inconsistent Information: Ranking and Clustering. *Journal of the ACM* 2008.
- A. Bondarenko, M. Völske, A. Panchenko, C. Biemann, B. Stein, M. Hagen. Webis at TREC 2018: Common Core Track. *In TREC 2018*.
- A. Bondarenko, M. Fröbe, V. Kasturia, M. Völske, B. Stein, M. Hagen. Webis at TREC 2019: Decision Track. *In TREC 2019*
- J. Bevendorff, A. Bondarenko, M. Fröbe, S. Günther, M. Völske, B. Stein, M. Hagen. Webis at TREC 2020: Health Misinformation Track. *TREC 2020*.
- A. Chernodub, O. Oliynyk, P. Heidenreich, A. Bondarenko, M. Hagen, C. Biemann, A. Panchenko. TARGER: Neural Argument Mining at Your Fingertips. *In ACL 2019*.
- H. Fang, Tao Tao, C. X. Zhai. A Formal Study of Information Retrieval heuristics. *In SIGIR 2004*.
- 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).
- B. Mitra, F. Diaz, and N. Craswell. Learning to match using local and distributed representations of text for web search. *In WWW 2017*.
- Newell, C.: Editing Tip: Sentence Length (2014).
- Adam D. Troy and Guo-Qiang Zhang. Enhancing relevance scoring with chronological term rank. *In SIGIR 2007*.

Webis at TREC 2021: Podcasts Track

November 17, 2021

Marcel Gohsen¹, **Johannes Kiesel**¹, Jakob Schwerter², Shahbaz Syed², Martin Potthast²,
Benno Stein¹

Bauhaus-Universität Weimar¹, Leipzig University²

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

webis_pc_bs:

no re-ranking

webis_pc_cola:

COLA audio embeddings

webis_pc_rob:

RoBERTa text embeddings

webis_pc_co_rob:

both concatenated

Criterion	Run	nDCG@30	nDCG@1000	P@10
Entertaining	bs	0.1182	0.2330	0.0975
	cola	0.0522	0.1748	0.0450
	rob	0.0351	0.1584	0.0275
	co_rob	0.0332	0.1620	0.0275
Subjective	bs	0.1725	0.3435	0.2000
	cola	0.0591	0.2443	0.0600
	rob	0.0371	0.2250	0.0350
	co_rob	0.0430	0.2320	0.0550
Discussion	bs	0.1619	0.3208	0.1600
	cola	0.0598	0.2289	0.0625
	rob	0.0399	0.2101	0.0400
	co_rob	0.0475	0.2193	0.0550

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

Run	EGFB score	E	G	F	B
abstr	0.2332	0	6	33	154
extr	0.2604	0	6	38	148
Baseline (one-minute)	0.8083	7	26	76	84