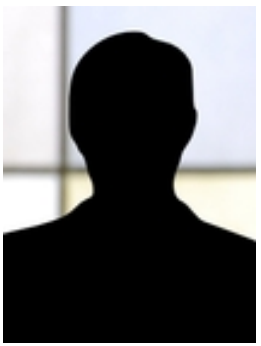
The Power of Anchor Text in the Neural Retrieval Era

April 13, 2022


Maik Fröbe


Sebastian Günther


Maximilian Probst


Martin Potthast

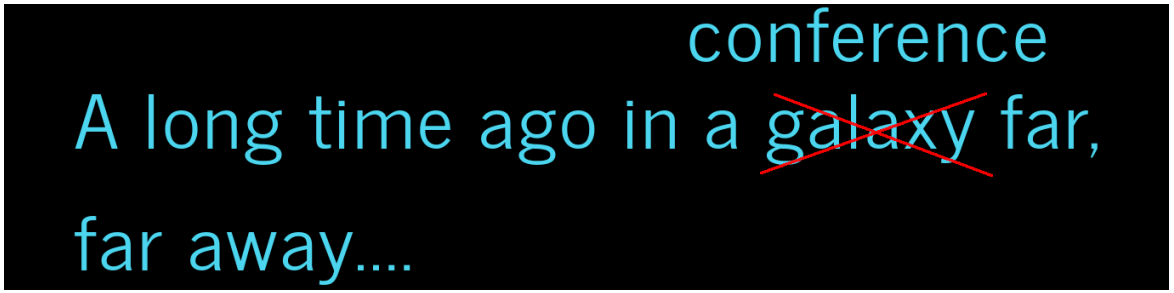

Matthias Hagen

Martin-Luther-Universität Halle-Wittenberg

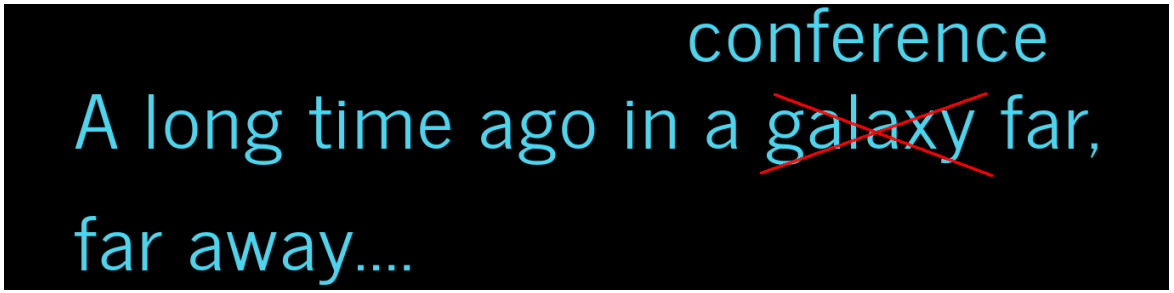
first.last@informatik.uni-halle.de

@webis_de www.webis.de

The Power of Anchor Text in the Neural Retrieval Era

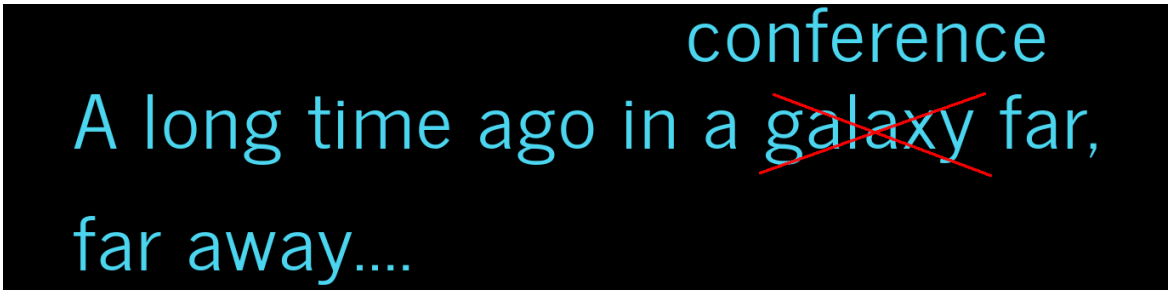


The Power of Anchor Text in the Neural Retrieval Era



"Anchors often provide more accurate descriptions of web pages than the pages themselves"
[Brin et al.;1998]

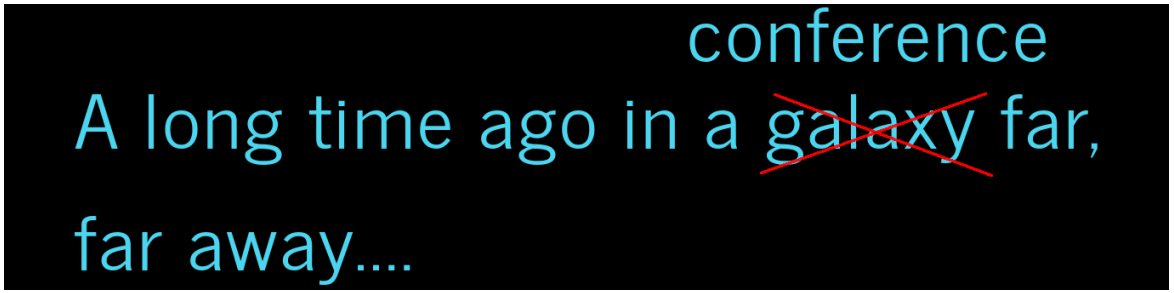
The Power of Anchor Text in the Neural Retrieval Era



"Anchors often provide more accurate descriptions of web pages than the pages themselves"
[Brin et al.;1998]

Commercial Web Search Engines	Academic Shared Tasks
Heavily use anchor text [Brin et al.;1998, Eiron et al., SIGIR'03]	No effectiveness gains with anchor text in TREC scenarios [Hawking et al., Westerveld et al.; TREC'99/01]

The Power of Anchor Text in the Neural Retrieval Era

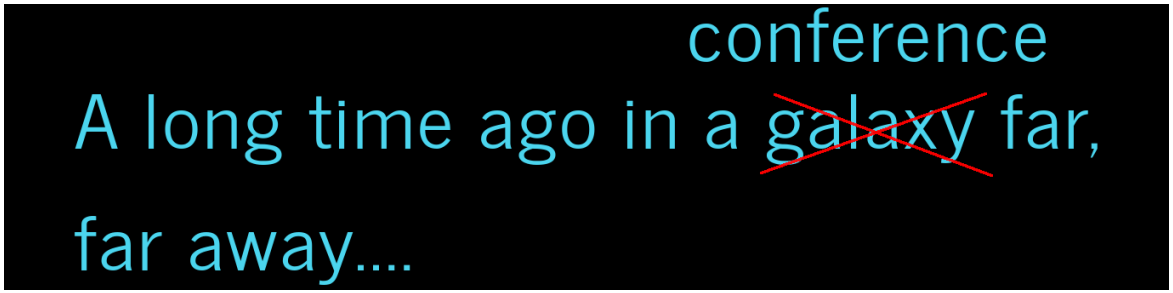


"Anchors often provide more accurate descriptions of web pages than the pages themselves"
[Brin et al.;1998]

Commercial Web Search Engines	Academic Shared Tasks
Heavily use anchor text [Brin et al.;1998, Eiron et al., SIGIR'03]	No effectiveness gains with anchor text in TREC scenarios [Hawking et al., Westerveld et al.; TREC'99/01]

- Craswell et al. provided an answer to this dichotomy [Craswell et al.; SIGIR'01]
  - Anchor texts are helpful for navigational queries
  - Hardly any TREC topic were navigational
  - $\geq$ 20% of the queries of commercial search engines were navigational
- Later: Relationship of anchor text and query logs [Eiron et al.; SIGIR'03]

The Power of Anchor Text in the Neural Retrieval Era



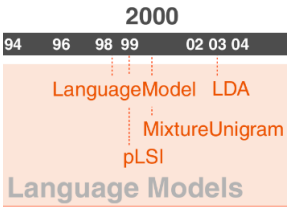
"Anchors often provide more accurate descriptions of web pages than the pages themselves"
[Brin et al.;1998]

Commercial Web Search Engines	Academic Shared Tasks
Heavily use anchor text [Brin et al.;1998, Eiron et al., SIGIR'03]	No effectiveness gains with anchor text in TREC scenarios [Hawking et al., Westerveld et al.; TREC'99/01]

We want to reproduce:

- Craswell et al. provided an answer to this dichotomy [Craswell et al.; SIGIR'01]
  - Anchor texts are helpful for navigational queries
  - Hardly any TREC topic were navigational
  - $\geq$ 20% of the queries of commercial search engines were navigational
- Later: Relationship of anchor text and query logs [Eiron et al.; SIGIR'03]

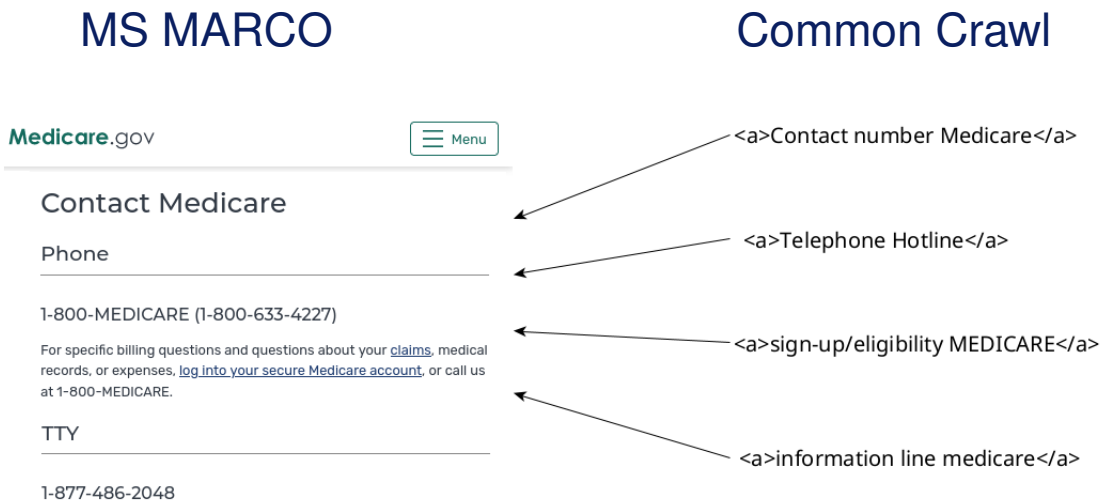
Why Reproduce?

	2 Decades Ago	Now
Computers		
Retrieval Models[1]		
Internet Users[2]	413 million	3.2 billion
Number of Websites[2]	17 million	1.6 billion

[1] https://webis.de/research/retrieval-models/retrieval-models.html
[2] https://www.internetlivestats.com/total-number-of-websites/

The Webis MS MARCO Anchor Text 2022 Dataset

- Goal: Enrich MS MARCO (V1 + V2) with anchor text
- MS Marco itself has sparse link structure
- We extract anchor text from six Common Crawl snapshots



- We re-use existing filter rules and pre-processing steps for anchor text
  [Chen et al.; WWW'20, Metzler et al.; SIGIR'09]
- Cherry-picked example
  - "medicare eligibility contact number"

	MRR@1000
BM25@Content	0.0
BM25@Anchor	1.0

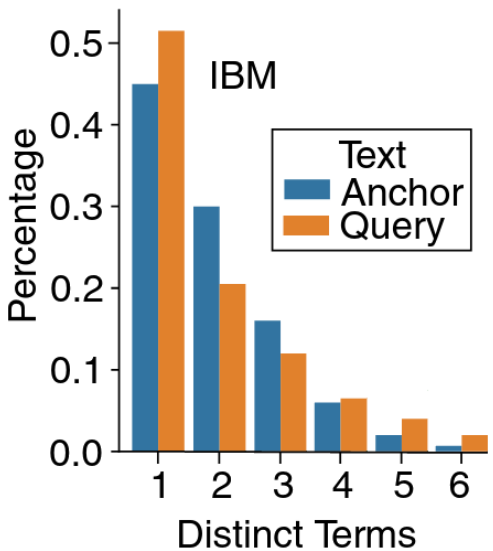
The Webis MS MARCO Anchor Text 2022 Dataset

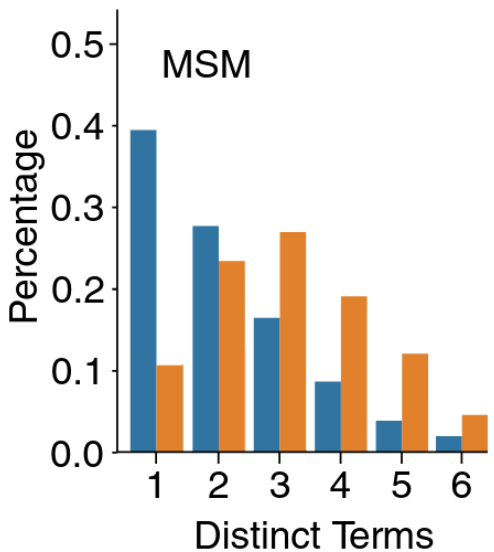
CC snapshot	Version 1		Version 2	
	Anchors	Docs. covered	Anchors	Docs. covered
2016	54.05 m	0.83 m	65.04 m	1.49 m
2017	61.19 m	1.18 m	94.35 m	2.34 m
2018	81.24 m	1.27 m	116.59 m	2.45 m
2019	65.60 m	1.16 m	90.18 m	2.83 m
2020	78.46 m	1.24 m	108.16 m	3.10 m
2021	60.62 m	1.14 m	84.93 m	3.18 m
$\sum$	207.28 m	1.70 m	341.17 m	4.82 m

Analysis of Anchor Text for Web Search [Eiron and McCurley, SIGIR'03]

Reproduction 1: Do anchor texts closely resemble query length?

	IBM Dataset	MS MARCO Dataset (V1)
Documents	2.95 M	3.21 M
Anchor Texts	2.57 M	81.24 M
Queries	1.27 M	18.82 M
Authors and Searchers	IBM employees	Diverse





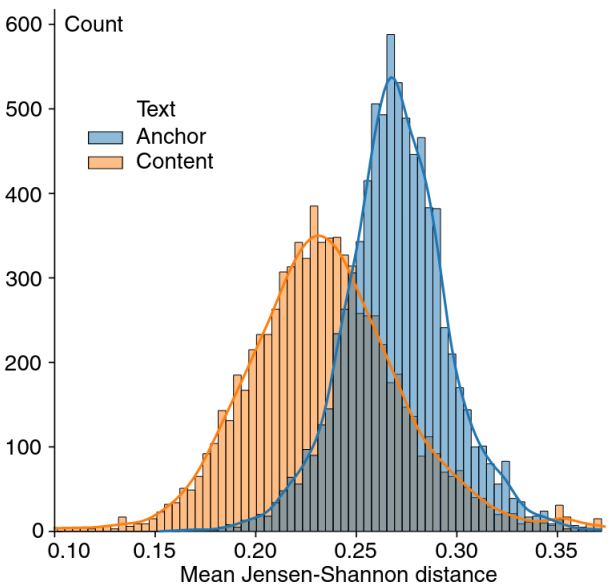
- Jensen-Shannon distances for reproduction:
  - ✗ 0.14 on IBM vs. 0.28 on MS MARCO
  - ✓ 0.10 for Anchor@IBM vs Anchor@MS MARCO

Analysis of Anchor Text for Web Search [Eiron and McCurley, SIGIR'03]

Search Result Homogeneity on the IBM Dataset (14 queries)

- Queries against the document content tended to retrieve results for every possible meaning of the query terms
- Queries against the anchor text retrieved more homogeneous results

Reproduction 2: Search Result Homogeneity on MS MARCO (7,962 queries)



`kelly clarkson fan club`

Index: Anchor	**Mean JS distance:** 0.32
1	Homepage of Kelly Clarkson
2	Wikipedia article on Kelly Clarkson
3	Latest News on Kelly Clarkson
⋮	
98	Statistics on movie star Grace Kelly
99	Vacancy by the Kelly Services company
100	Login Page to Facebook

Index: Content	**Mean JS distance:** 0.13
1	News on Kelly Clarkson's career
2	Wikipedia article on a Kelly Clarkson single
3	IMDb biography of Kelly Clarkson
⋮	
98	News article on a Kelly Clarkson album
99	Review of Kelly Clarkson at American Idol
100	Article on weight gain of Kelly Clarkson

Effective Site Finding Using Anchor Information [Craswell et al., SIGIR'01]

Reproduction: Anchor text helps for navigational queries pointing to random pages

- We construct 100 navigational queries pointing to random entry pages
  - Heuristic to find entry-pages: URL-path empty or index.html
    [Westerveld et al., TREC'01]
  - From 92,562 candidates we select 100 pages at random
  - Manually formulated queries searchers would use to find the pages

Retrieval system		**Random**	
Text/Features	Model	MRR	Recall@10
Anchor	BM25@2018	0.70	0.82
	BM25@2016–2021	**0.74**	**0.89**
Content	BM25@Content	0.21	0.36
	DeepCT@Anchor	**0.43**	**0.58**
	DeepCT@Train	0.27	0.44
	MonoT5	0.39	0.53
ORCAS	BM25	**0.60**	**0.70**
Content + ORCAS + Anchor	λ-MART	0.48	0.63

Effective Site Finding Using Anchor Information [Craswell et al., SIGIR'01]

Reproduction: Anchor text helps for navigational queries pointing to popular pages

- We construct 100 navigational queries pointing to popular entry pages
  - Randomly selected with domains in the Alexa top-1000 of 2018
  - Manually formulated queries

Retrieval system		**Popular**	
Text/Features	Model	MRR	Recall@10
Anchor	BM25@2018	0.54	0.81
	BM25@2016–2021	**0.55**	**0.84**
Content	BM25@Content	0.02	0.03
	DeepCT@Anchor	**0.03**	**0.08**
	DeepCT@Train	0.02	0.05
	MonoT5	0.02	0.05
ORCAS	BM25	**0.28**	**0.43**
Content + ORCAS + Anchor	λ-MART	0.08	0.18

Effectiveness for Informational Queries

nDCG@10 on TREC-DL Topics

Retrieval systems		**DL 2019**	**DL 2020**
Text/Features	Model		
Anchor	BM25@2018-13	0.35	0.27
	BM25@2016–2021	**0.41**	**0.34**
Content	BM25@Content	0.51	0.53
	DeepCT@Anchor	0.53	0.55
	DeepCT@Train	0.54	0.51
	MonoT5	**0.68**	0.62
ORCAS	BM25	0.45	0.36
Content + ORCAS + Anchor	λ-MART	**0.59**	**0.57**

Conclusions

Summary

- Anchor text is effective for navigational queries
- Transformer-based approaches are less effective for navigational queries than anchor text-oriented BM25 retrieval
- Decision tree using basic query understanding
  - Informational query $\Rightarrow$ Transformer
  - Navigational query pointing to random page $\Rightarrow$ Anchor text
  - Navigational query pointing to popular page $\Rightarrow$ Click logs
- Dataset available in ir_datasets

Conclusions

Summary

- Anchor text is effective for navigational queries
- Transformer-based approaches are less effective for navigational queries than anchor text-oriented BM25 retrieval
- Decision tree using basic query understanding
  - Informational query $\Rightarrow$ Transformer
  - Navigational query pointing to random page $\Rightarrow$ Anchor text
  - Navigational query pointing to popular page $\Rightarrow$ Click logs
- Dataset available in ir_datasets

Future work

- Mining of parallel texts between query logs and anchor texts
- Selection of subsets of the training data from MS MARCO
  - Exclude navigational training examples covered by anchor text

Conclusions

Summary

- Anchor text is effective for navigational queries
- Transformer-based approaches are less effective for navigational queries than anchor text-oriented BM25 retrieval
- Decision tree using basic query understanding
  - Informational query $\Rightarrow$ Transformer
  - Navigational query pointing to random page $\Rightarrow$ Anchor text
  - Navigational query pointing to popular page $\Rightarrow$ Click logs
- Dataset available in ir_datasets

Future work

- Mining of parallel texts between query logs and anchor texts
- Selection of subsets of the training data from MS MARCO
  - Exclude navigational training examples covered by anchor text

thank you!