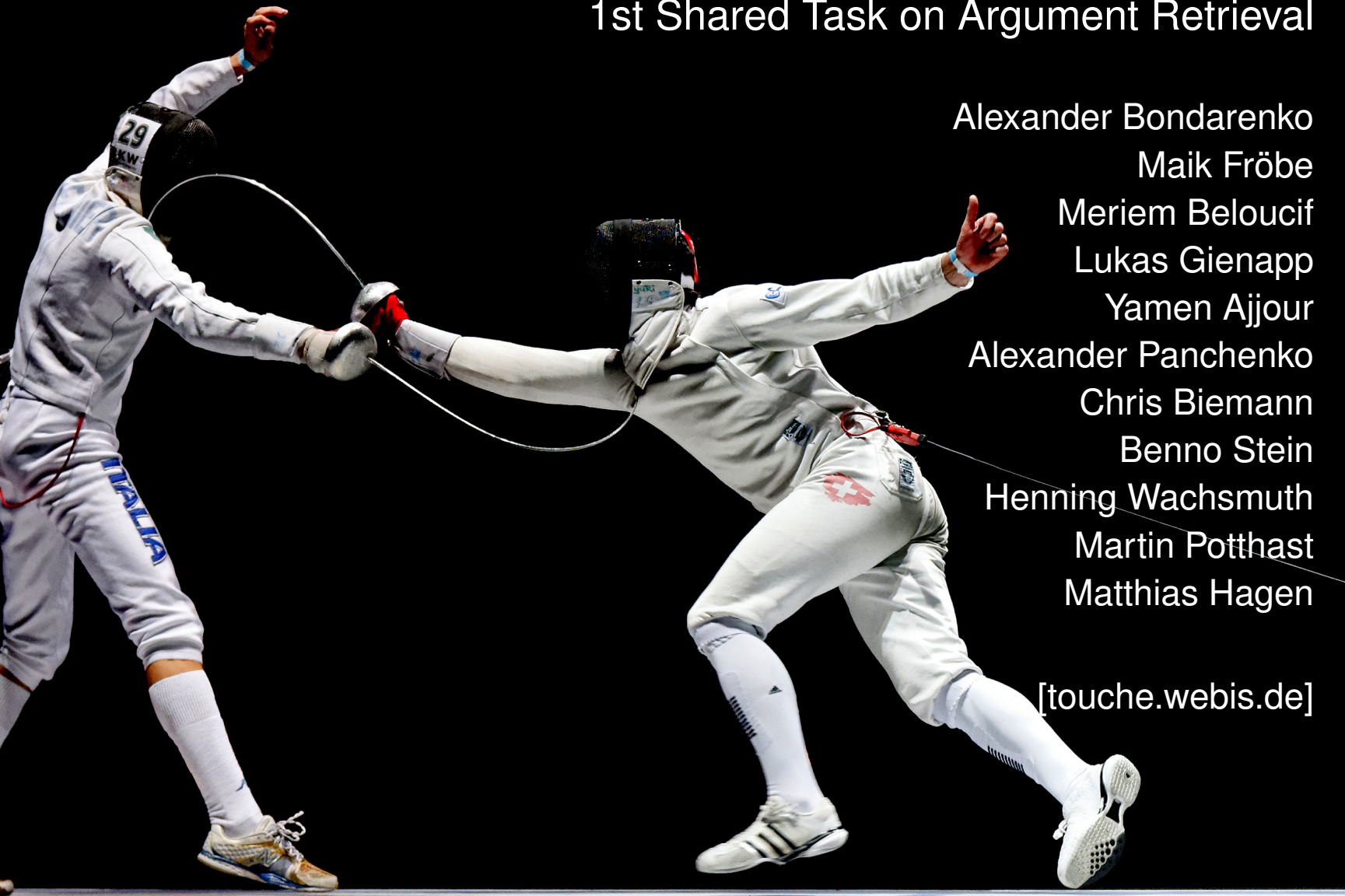
Touché @ CLEF 2020

1st Shared Task on Argument Retrieval



Alexander Bondarenko
Maik Fröbe
Meriem Beloucif
Lukas Gienapp
Yamen Ajjour
Alexander Panchenko
Chris Biemann
Benno Stein
Henning Wachsmuth
Martin Potthast
Matthias Hagen

[touche.webis.de]

A Timeline [Croft 2019]

TOUCHÉ 2020

Time

Document Retrieval

Answer Passage Retrieval

Passages as Features

Sentence Retrieval

Snippet Retrieval

QA Factoid Retrieval

CQA or Non-Factoid QA

Conversational Answer Retrieval

Answer Passage Retrieval Revisited

Response Retrieval/Generation

Question Answering/Machine Comprehension

Complex Answer Retrieval
(Passages as Summaries)

A Timeline [Croft 2019]

TOUCHÉ 2020

Time

Document Retrieval

Answer Passage Retrieval

Passages as Features

Sentence Retrieval

Snippet Retrieval

QA Factoid Retrieval

CQA or Non-Factoid QA

Conversational Answer Retrieval

Answer Passage Retrieval Revisited

Response Retrieval/Generation

Question Answering/Machine Comprehension

Complex Answer Retrieval
(Passages as Summaries)

Argument Retrieval

Touché: Argument Retrieval

Shared Tasks

TOUCHÉ 2020

Task 1: Supporting argumentative conversations

- Scenario: Users search for arguments on controversial topics
- Task: Retrieve “strong” pro/con arguments on the topic
- Data: 400,000 “arguments” (short text passages) [args.me]

Task 2: Answering comparative questions with arguments

- Scenario: Users face personal decisions from everyday life
- Task: Retrieve arguments for “Is X better than Y for Z?”
- Data: ClueWeb12 or ChatNoir [chatnoir.eu]

- Run submissions similar to “classical” TREC tracks
- Software submissions via TIRA [tira.io]

Touché: Argument Retrieval

Argument and Argumentation

TOUCHÉ 2020

Argument:

- A conclusion (claim) supported by premises (reasons) [Walton et al. 2008]
- Conveys a stance on a controversial topic [Freeley and Steinberg, 2009]

Conclusion *Argumentation will be a key element of conversational agents.*

Premise 1 *Superficial conversation ("gossip") is not enough.*

Premise 2 *Users want to know the "Why" to make informed decisions.*

Argumentation:

- Usage of arguments to achieve persuasion, agreement, . . .
- Decision making and opinion formation processes

Touché: Argument Retrieval

Topics

TOUCHÉ 2020

Example topic for Task 1:

Title	*Is climate change real?*
Description	*You read an opinion piece on how climate change is a hoax and disagree. Now you are looking for arguments supporting the claim that climate change is in fact real.*
Narrative	*Relevant arguments will support the given stance that climate change is real or attack a hoax side's argument.*

Touché: Argument Retrieval

Topics

TOUCHÉ 2020

Example topic for Task 2:

Title	*Which is better, a laptop or a desktop?*
Description	*A user wants to buy a new PC but has no prior preferences. [. . .] This can range from situations like frequent traveling where a mobile device is to be favored to situations of a rather "stationary" gaming desktop PC.*
Narrative	*Highly relevant documents will describe what the major similarities and dissimilarities of laptops and desktops [. . .] A comparison of the technical and architectural characteristics without a personal opinion, recommendation or pros/cons is not relevant.*

Touché: Argument Retrieval

Document Collections

TOUCHÉ 2020

Task 1: Supporting argumentative conversations

- Args.me corpus [Ajjour et al. 2019]
- Argument passages from debate portals: idebate.org, debate.org, . . .
- Download or accessible via the API of args.me search engine [args.me]

Task 2: Answering comparative questions with arguments

- ClueWeb12: accessible via the ChatNoir API [chatnoir.eu]

Touché: Argument Retrieval

Statistics

- Registrations: 28 teams
- Nicknames: Real or fictional fencers / swordsmen (e.g., Zorro)
- Submissions: 17 participating teams
- Approaches: 41 valid runs were evaluated
- Baselines: DirichletLM and BM25F-based ChatNoir [chatnoir.eu]
- Evaluation: 7,045 manual relevance judgments (nDCG@5)

Touché: Argument Retrieval

Evaluation

TOUCHÉ 2020

Classical (TREC style) IR relevance judgments

Not relevant

Relevant

Highly relevant

Argument retrieval: How good are the arguments?

Task 1

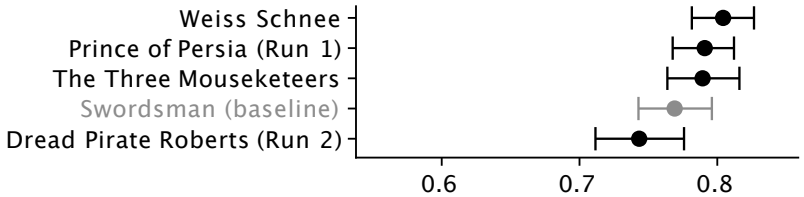
- Argument relevance
- Top-5 pooling
- 5,262 unique passages
- Amazon Mechanical Turk
- nDCG@5

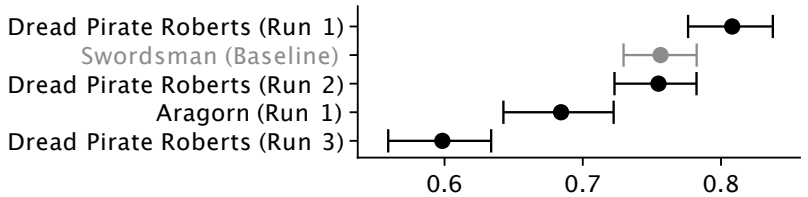
Task 2

- Document relevance
- Top-5 pooling
- 1,783 unique documents
- Volunteers
- nDCG@5

Touché: Argument Retrieval

Task 1 Results





Mean nDCG@5, top 5 runs, args.me version 1 and 2.

Team	Retrieval	Augmentation	(Re)ranking Feature
Dread Pirate Roberts	DirichletLM/Similarity-based	Language modeling	—
Weiss Schnee	DPH	Embeddings	Quality
Prince of Persia	Multiple models	Synonyms	Sentiment
The Three Mouseketeers	DirichletLM	—	—
Swordsman (Baseline)	DirichletLM	—	—
Thongor	BM25/DirichletLM	—	—
Oscar François de Jarjayes	DPH/Similarity-based	—	Sentiment
Black Knight	TF-IDF	Cluster-based	Stance, readability
Utena Tenjou	BM25	—	—
Arya Stark	BM25	—	—
Don Quixote	Divergence from Randomness	Cluster-based	Quality + Similarity
Boromir	Similarity-based	Topic modeling	Author credibility
Aragorn	BM25	—	Premise prediction
Zorro	BM25	—	Quality + NER

Touché: Argument Retrieval

Task 1 Results

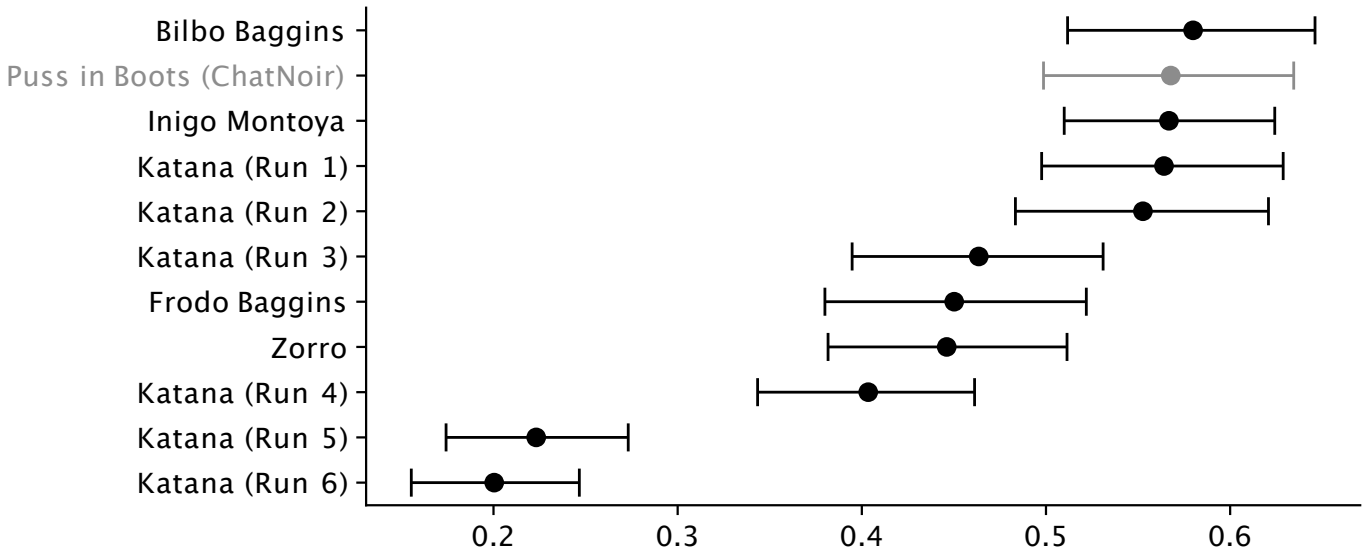
TOUCHÉ 2020

Easiest and hardest topics.

Topic title	nDCG@5
Is Golf a Sport?	0.80
Should Churches Remain Tax-Exempt?	0.72
Should Everyone Get a Universal Basic Income?	0.69
Should birth control pills be available over the counter?	0.66
Is Human Activity Primarily Responsible for Global Climate Change?	0.63
...	...
Should Student Loan Debt Be Easier to Discharge in Bankruptcy?	0.20
Should Social Security Be Privatized?	0.20
Is a College Education Worth It?	0.15
Should Felons Who Have Completed Their Sentence Be Allowed to Vote?	0.15
Should Adults Have the Right to Carry a Concealed Handgun?	0.07
Average across all topics	0.42

Touché: Argument Retrieval

Task 2 Results



Mean nDCG@5 and 95% confidence intervals.

Team	Representation	Query processing	(Re-)Ranking features
Bilbo Baggins	Bag of words	Named entities, comp. aspects	Credibility, support
Puss in Boots	Bag of words	—	BM25F, SpamRank
Inigo Montoya	Bag of words	Tokens & logic. OR	Argum. units (TARGER)
Katana	Diff. language models	Diff. language models	Comparativeness score
Frodo Baggins	Bag of words	GloVe nearest neighbors	Simil. with gen. documents (GPT-2)
Zorro	Bag of words	—	PageRank, argumentativeness

Touché: Argument Retrieval

Task 2 Results

TOUCHÉ 2020

Easiest and hardest topics.

Topic title	nDCG@5
Which is better, a laptop or a desktop?	0.84
What is better for the environment, a real or a fake Christmas tree?	0.80
Which is better, Pepsi or Coke?	0.70
What is better: ASP or PHP?	0.70
Which is better, Linux or Microsoft?	0.70
...	...
Which city is more expensive to live in: San Francisco or New York?	0.18
Which smartphone has a better battery life: Xperia or iPhone?	0.17
What is better: to use a brush or a sponge?	0.16
What is the longest river in the U.S.?	0.10
What are the advantages and disadvantages of PHP over Python and vice versa?	0.10
Average across all topics	0.46

Touché: Argument Retrieval

Summary

TOUCHÉ 2020

- Platform for argument retrieval researchers
- Argument relevance / quality corpora
- Tools for submission and evaluation

- “Simple” argumentation-agnostic baselines perform well
- “Best” so far: query expansion, argument quality, comparative features
- Workshop Touché today at 15:00

Bondarenko et al. Overview of Touché 2020: Argument Retrieval [https://webis.de/publications.html?q=stein2020v]

Touché: Argument Retrieval

Outlook Touché 2021

TOUCHÉ 2021

- 50 search topics more
- Deeper judgment pools
- This year's topics and judgments available for training
- Evaluate argument quality dimensions:
  e.g., well-written, logically cogent, good support [Wachsmuth, et al. 2017]

Touché: Argument Retrieval

Outlook Touché 2021

TOUCHÉ 2021

- 50 search topics more
- Deeper judgment pools
- This year's topics and judgments available for training
- Evaluate argument quality dimensions:
  e.g., well-written, logically cogent, good support [Wachsmuth, et al. 2017]

Touché: Argument Retrieval

Related Publications

- Ajjour, Wachsmuth, Kiesel, Potthast, Hagen, Stein. Data Acquisition for Argument Search: The args.me corpus. Proceedings of KI 2019.
- Bevendorff, Stein, Hagen, Potthas. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. Proceedings of ECIR 2018.
- Braunstain, Kurland, Carmel, Szpektor, Shtok. Supporting Human Answers for Advice-Seeking Questions in CQA Sites. Proceedings of ECIR 2016.
- Croft. The Relevance of Answers. Keynote at CLEF 2019. https://ciir.cs.umass.edu/downloads/clef2019/CLEF_2019_Croft.pdf
- Freely and Steinberg. Argumentation and Debate: Critical Thinking for Reasoned Decision Making (12th ed.). Boston, MA: Wadsworth Cengage Learning, 2009.
- Potthast, Gienapp, Euchner, Heilenkötter, Weidmann, Wachsmuth, Stein, Hagen. Argument Search: Assessing Argument Relevance. Proceedings of SIGIR 2019.
- Wachsmuth, Naderi, Hou, Bilu, Prabhakaran, Alberdingk Thijm, Hirst, Stein. Computational Argumentation Quality Assessment in Natural Language. Proceedings of EACL 2017.
- Walton, Reed, Macagno. Argumentation Schemes. Cambridge: Cambridge University Press, 2008.