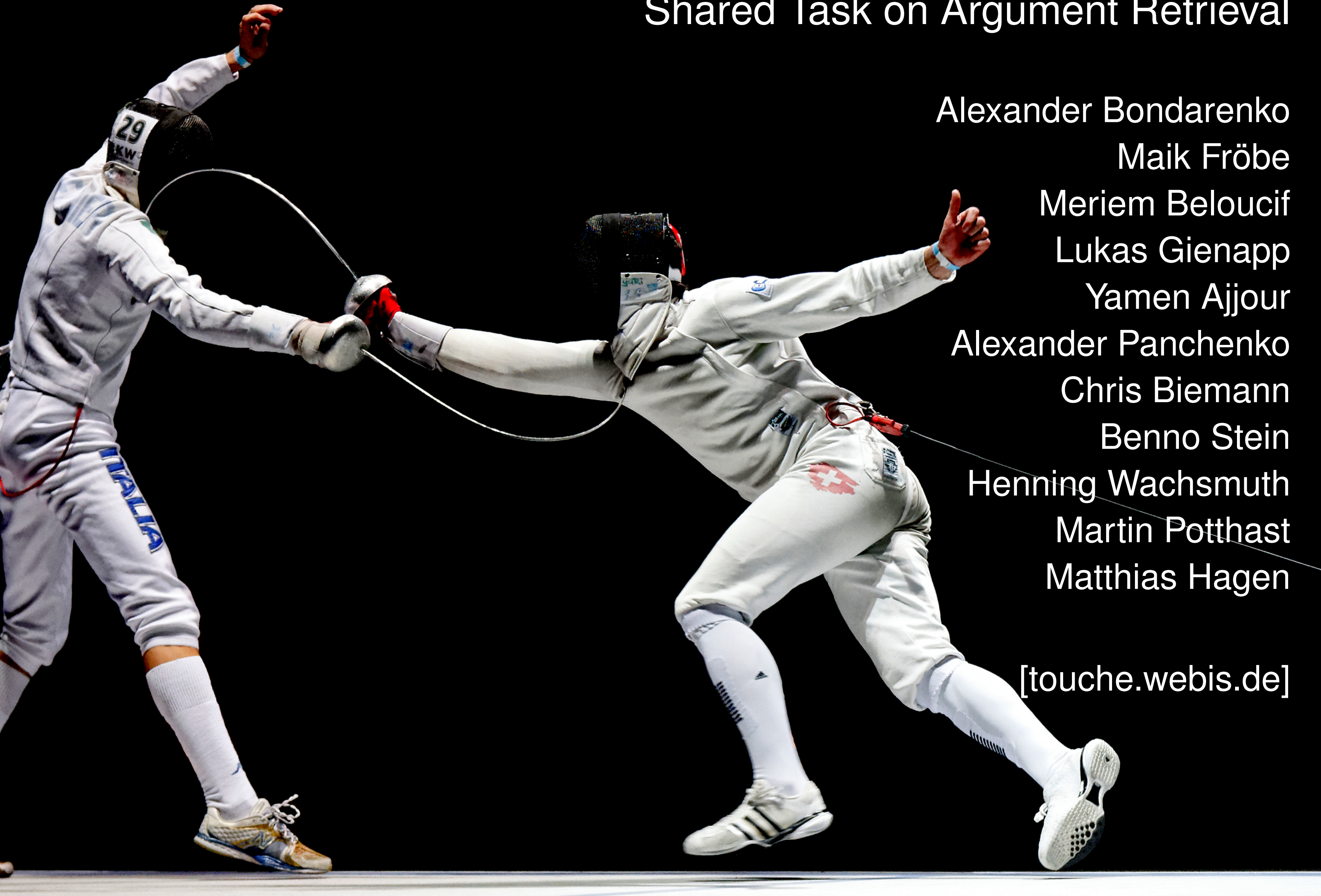


Touché @ CLEF

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]

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]

Time

Document Retrieval

Answer Passage Retrieval

Sentence Retrieval

Passages as Features

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

Task 1: Supporting debates on controversial topics

- 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]

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

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.*

Example **pro** argument:

One reason that I believe that **climate change is real** is the increase in global temperature and the shrinking of the Arctic ice. This is shown on this website [link].

Touché: Argument Retrieval

Document Collection

Task 1: Supporting debates on controversial topics

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

Touché: Argument Retrieval

Statistics 2020

❑ Registrations:	21 teams (incl. for both tasks)
❑ Nicknames:	Real or fictional fencers / swordsmen (e.g., Zorro)
❑ Submissions:	13 participating teams
❑ Approaches:	30 valid runs were evaluated
❑ Baseline:	DirichletLM (Lucene Implementation)
❑ Evaluation:	5,262 manual relevance judgments (nDCG@5)

Touché: Argument Retrieval

Evaluation

Argument retrieval: How good are the results?

- Evaluation w.r.t. argument relevance
- Top-5 pooling
- 5,262 unique passages
- Amazon Mechanical Turk
- nDCG@5

Classical (TREC style) IR relevance judgments:

(1) Text is an argument ➜ relevance $\in [1, ..., 5]$ (low to high)

(2) Text is not an argument ➜ relevance $= -2$

Touché: Argument Retrieval

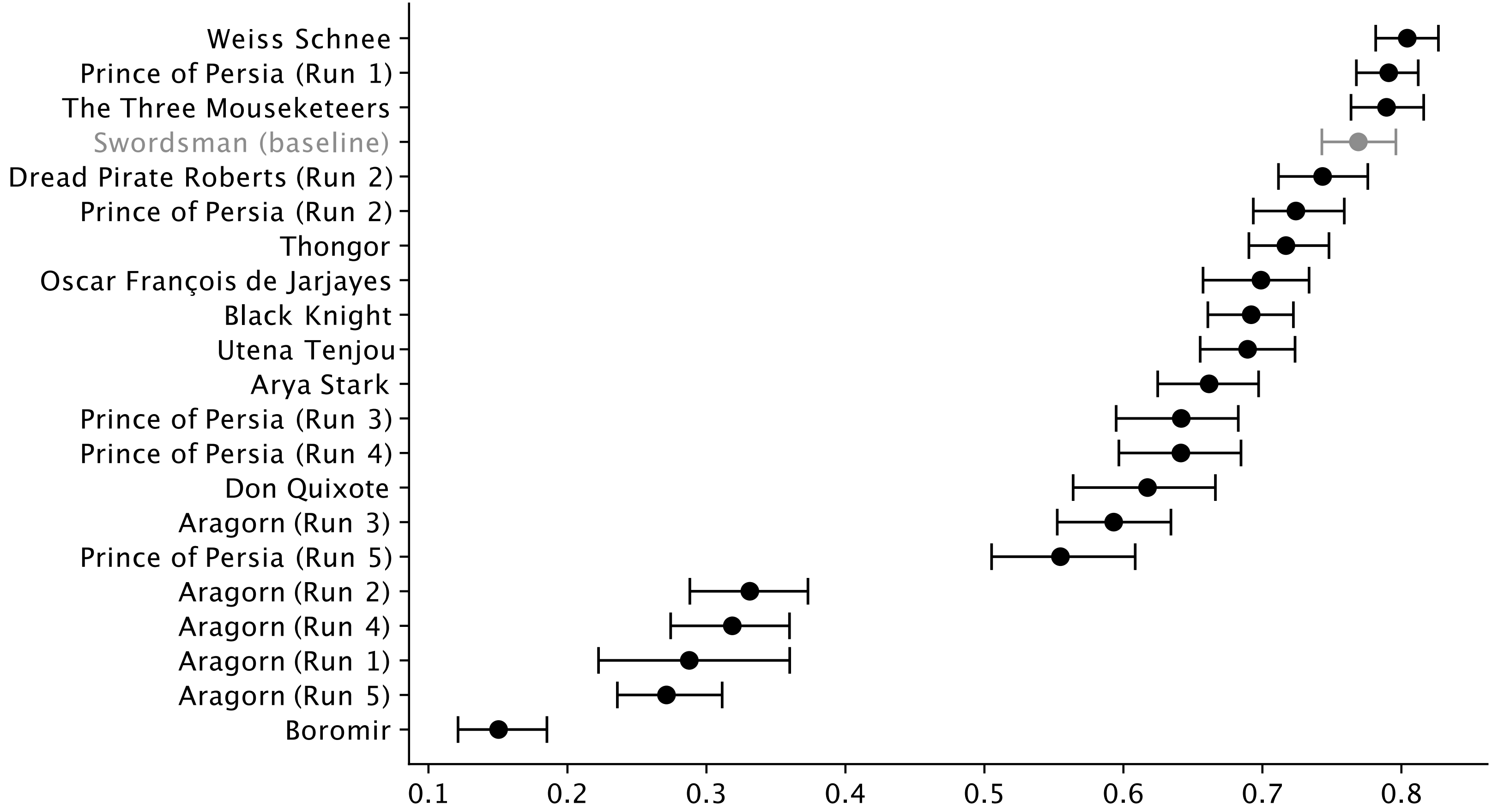
Task 1 Strategy Overview

TOUCHÉ 2021

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



Mean nDCG@5 and 95% confidence intervals.

Touché: Argument Retrieval

Task 1 Results

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

Approaches

Baselines:

- BM25, DPH, TF-IDF, **DirichletLM**

Where to start:

- Tf-idf based models are good
- Statistical language models are better
- Argument quality matters

Touché: Argument Retrieval

Approaches

Winning submissions:

- Query expansion: WordNet synonyms / antonyms → GPT-2 generation
- Document representations using Transformer (e.g., BERT)
- Re-ranking based on argument quality prediction
- Re-ranking based on sentiment (neutral sentiment)
- Pseudo-relevance feedback

[Overview of Touché 2020: Argument Retrieval]

[Touché 2020 participant papers]

[Touché 2020 slides: talks, overview]

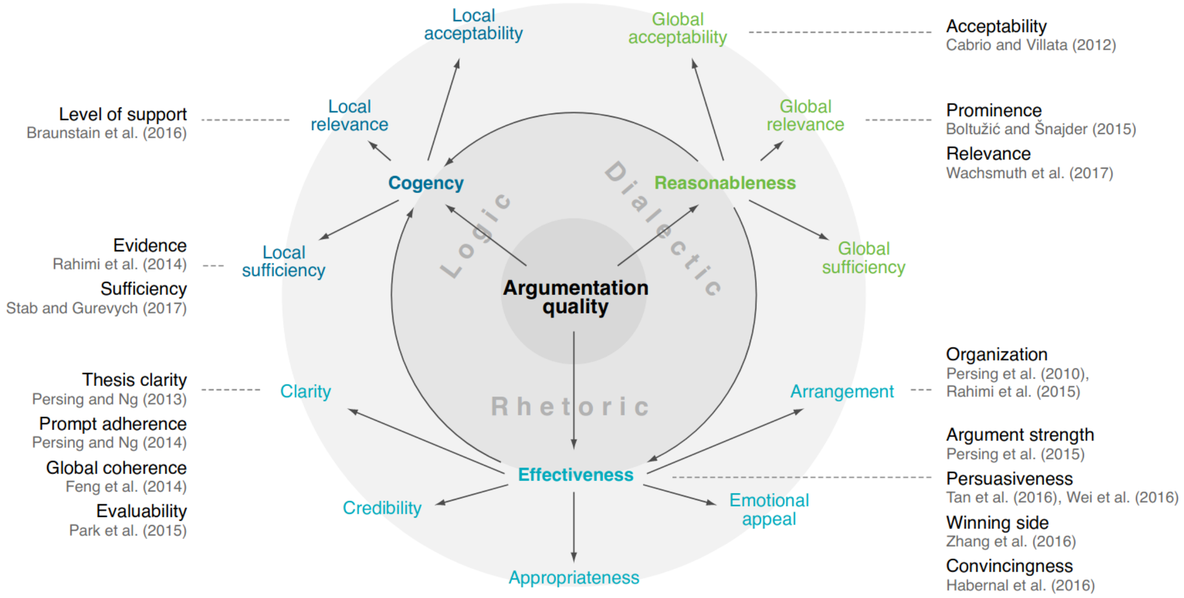
[Touché 2020 videos on YouTube]

Touché: Argument Retrieval

Outlook Touché 2021

- ❑ 50 search topics more [touche.webis.de]
- ❑ Deeper judgment pools
- ❑ Last year’s [topics and judgments] available for training
- ❑ Evaluate argument quality: e.g., well-written, logically cogent

  [Computational Argumentation Quality Assessment in Natural Language]



Touché: Argument Retrieval

References:

- Ajjour et al. Data Acquisition for Argument Search: The args.me corpus. Proc. of KI 2019.
- 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.
- Wachsmuth et al. Computational Argumentation Quality Assessment in Natural Language. Proc. of EACL 2017.
- Walton et al. Argumentation Schemes. Cambridge: Cambridge University Press, 2008.

Argument Quality Datasets:

- [https://webis.de/data.html?q=quality]

Lecture Slides / Tutorials:

- [Argument Search]
- [Applications of Computational Argumentation]
- [Argument Retrieval]

Touché: Argument Retrieval

References:

- Ajjour et al. Data Acquisition for Argument Search: The args.me corpus. Proc. of KI 2019.
- 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.
- Wachsmuth et al. Computational Argumentation Quality Assessment in Natural Language. Proc. of EACL 2017.
- Walton et al. Argumentation Schemes. Cambridge: Cambridge University Press, 2008.

Argument Quality Datasets:

- [https://webis.de/data.html?q=quality]

Lecture Slides / Tutorials:

- [Argument Search]
- [Applications of Computational Argumentation]
- [Argument Retrieval]

thank you!