TOUCHÉ 2020

Session 1: Conversational Argument Retrieval

Moderator: Lukas Gienapp

Keynote:

Argument Retrieval in Project Debater

Yufang Hou, IBM Research Europe, Dublin

[webpage]

Touché: Argument Retrieval

Shared Task

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)

- 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

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]
- Two versions, differing in size; up to participants which one to use

Touché: Argument Retrieval

Statistics

TOUCHÉ 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

TOUCHÉ 2020

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 $\in [0, ..., 4]$ (low to high)

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

Touché: Argument Retrieval

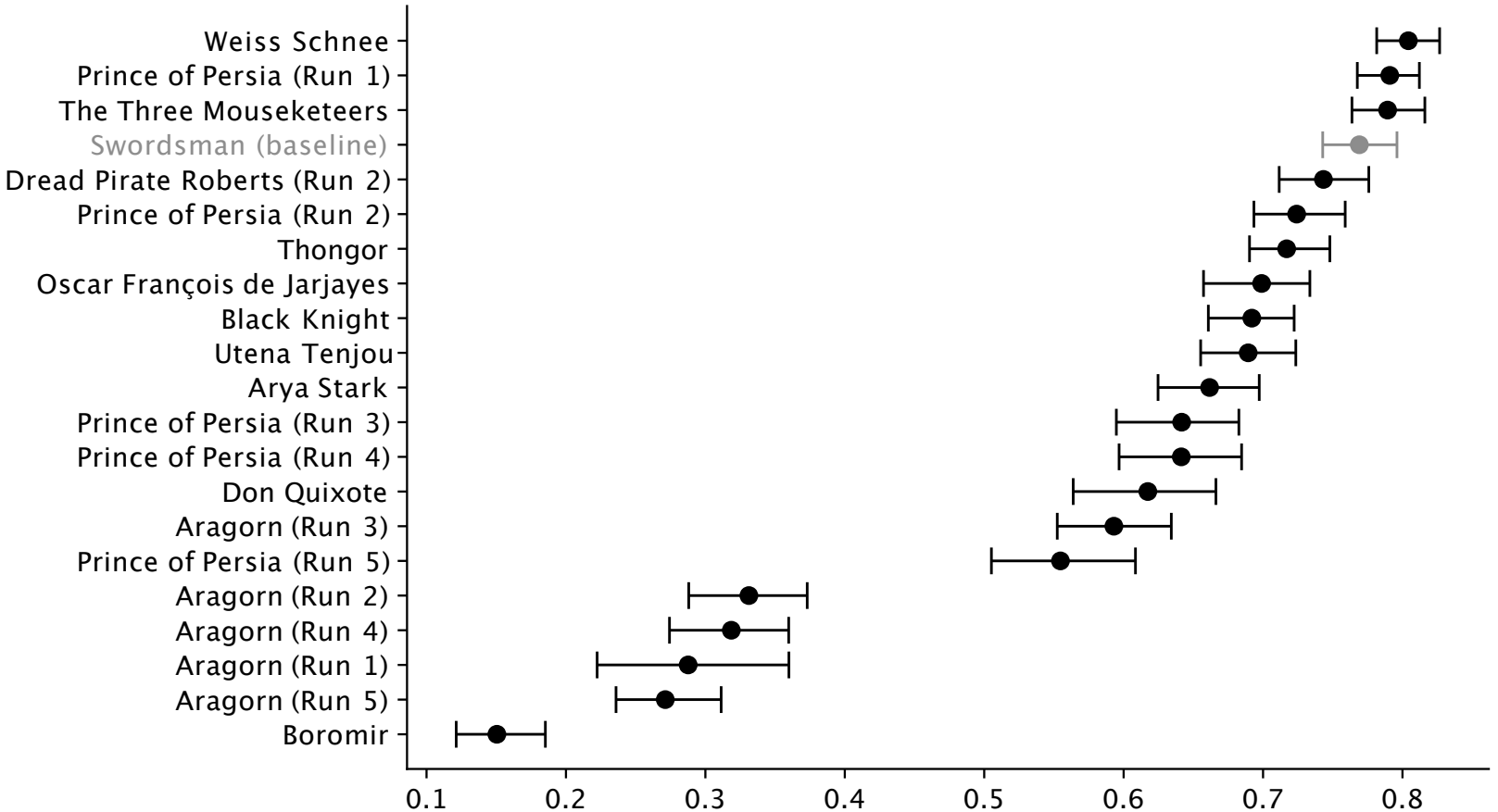
Task 1 Strategy Overview

TOUCHÉ 2020

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

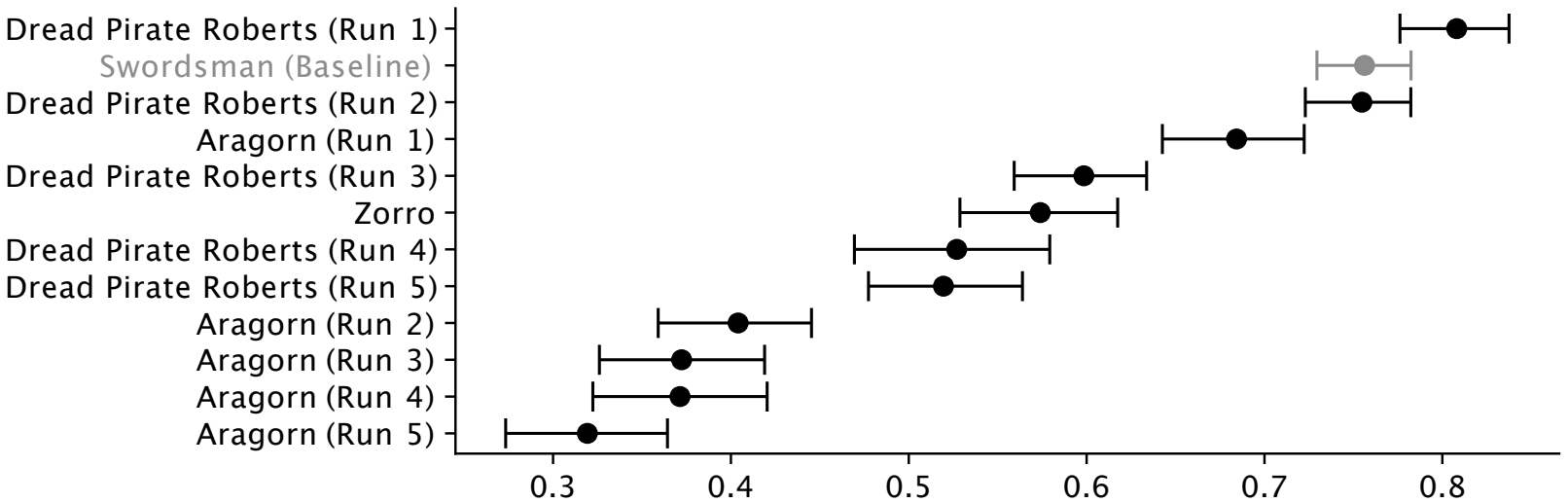
Task 1 Results



Mean nDCG@5 and 95% confidence intervals args.me version 1.

Touché: Argument Retrieval

Task 1 Results



Mean nDCG@5 and 95% confidence intervals args.me version 2.

Touché: Argument Retrieval

Submitted Papers

TOUCHÉ 2020

Team	Paper
Pirate Roberts	Akiki, Potthast. Exploring Argument Retrieval with Transformers.
Weiss Schnee	Bundesmann, Christ, Richter. Creating an Argument Search Engine for Online Debates.
Oscar Francois	Staudte & Lange. SentArg: A Hybrid Doc2Vec/DPH Model with Sentiment Analysis Refinement.
Don Quixote	Dumani & Schenkel. Ranking Arguments by Combining Claim Similarity and Arg. Quality Dimensions.
Aragorn	Entezari & Völske. Argument Retrieval Using Deep Neural Ranking Models
Zorro	Shahshahani & Kamps. University of Amsterdam at CLEF 2020
Baseline	Lucene Implementation of DirichletLM [Zhai & Lafferty 2004] Good results in pilot study [Potthast et al. 2019]

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

Summary

TOUCHÉ 2020

- All participants index themselves; most participants use the smaller version of the args.me corpus
- “Simple” argumentation-agnostic baselines perform well
- Very little labeled training data available for neural approaches
- “Best” so far: general retrieval model + domain-specific embedding/transformer-based query expansion; quality as (re)ranking feature

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.
- Zhai,Lafferty. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. ACM TOIS, 22(2), 2004.