TOUCHÉ 2020

# Session 2: Comparative Argument Retrieval

Moderator: Alexander Bondarenko

# Touché: Argument Retrieval

## Session 2: Comparative Argument Retrieval

TOUCHÉ 2020

Keynote:

## Debate Technology for Empowering the Public: Insights and Avenues

Annette Hautli-Janisz, University of Konstanz

[webpage]

# Touché: Argument Retrieval

## Shared Task

TOUCHÉ
2020

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 accessible via ChatNoir API [chatnoir.eu]

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

# 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

## Statistics

TOUCHÉ 2020

- Registrations: 18 teams (incl. for both tasks)
- Nicknames: Real or fictional fencers / swordsmen (e.g., Katana)
- Submissions: 5 participating teams
- Approaches: 11 valid runs were evaluated
- Baseline: BM25F-based ChatNoir [chatnoir.eu]
- Evaluation: 1,783 manual relevance judgments (nDCG@5)

# Touché: Argument Retrieval

## Evaluation

Classical (TREC style) IR relevance judgments

| | | |
|---|---|---|
| 👎 | 👍 | 👍👍 |
| Not relevant | Relevant | Highly relevant |

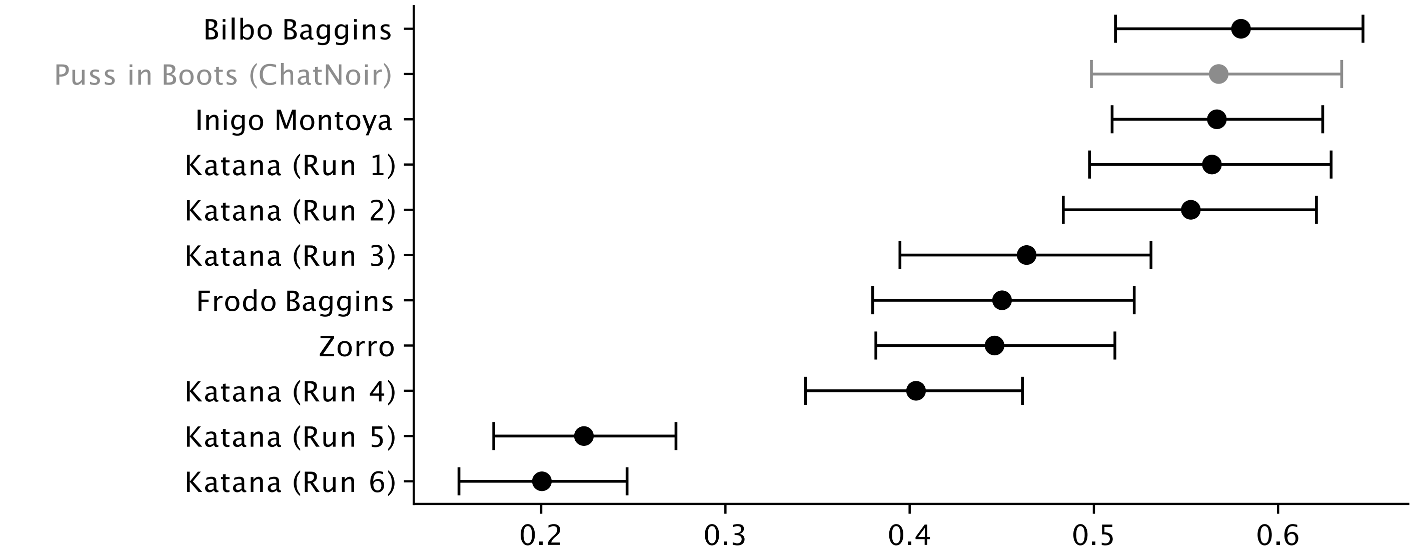
Argument retrieval: How good are the arguments?

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

# Touché: Argument Retrieval

## Task 2 Results

TOUCHÉ 2020



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

## Submitted Papers

| Team | Paper |
|---|---|
| Bilbo Baggins | Abye, Sager, Triebel.<br>An Open-Domain Web Search Engine for Answering Comparative Questions. |
| Inigo Montoya | Huck.<br>Development of a Search Engine to Answer Comparative Queries. |
| Katana | Chekalina & Panchenko.<br>Retrieving Comparative Arguments using Deep Pre-trained Language Models and NLU. |
| Frodo Baggins | Sievers.<br>Question answering for comparative questions with GPT-2 |
| Zorro | Shahshahani & Kamps.<br>University of Amsterdam at CLEF 2020 |
| Baseline | Bevendorff, Stein, Hagen, Potthast (ECIR 2018).<br>Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. |

# Touché: Argument Retrieval

## Task 2 Results

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

- All approaches re-rank ChatNoir results
- “Simple” argumentation-agnostic baselines perform well
- Top-4 runs are classical feature engineering approaches
- No training data was available for neural approaches
- “Best” so far: query expansion, argument quality, comparative features

# Touché: Argument Retrieval

## Related Publications

TOUCHÉ 2020

- 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.