Touché @ CLEF 2021
Shared Tasks 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]

**Document Retrieval**
- Answer Passage Retrieval
- Sentence Retrieval
- QA Factoid Retrieval
- Passages as Features
- Snippet 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]

**Document Retrieval**
- Answer Passage Retrieval
- Sentence Retrieval
- QA Factoid Retrieval
- Passages as Features
- Snippet 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**

**Time**
Task 1: Supporting conversations on controversial topics

- **Scenario:** Users search for arguments on controversial topics
- **Task:** Retrieve and rank “strong” pro/con arguments on the topic
- **Data:** 400,000 “arguments” (short text passages) [args.me]
Task 1: Supporting conversations on controversial topics
- Scenario: Users search for arguments on controversial topics
- Task: Retrieve and rank “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 and rank arguments for “Is X better than Y for Z?”
- Data: ClueWeb12 or ChatNoir [chatnoir.eu]
Task 1: Supporting conversations on controversial topics
- Scenario: Users search for arguments on controversial topics
- Task: Retrieve and rank “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 and rank 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]
### Statistics

- **Registrations:** 36 teams (vs. 28 teams last year)
- **Nicknames:** Real or fictional fencers / swordsmen (e.g., Zorro)
- **Submissions:** 27 participating teams (vs. 17 last year)
- **Approaches:** 88 valid runs were evaluated (vs. 41 last year)
- **Baselines:** DirichletLM and BM25F-based ChatNoir [chatnoir.eu]
- **Evaluation:** 5,787 manual relevance and quality judgments (nDCG@5)

<table>
<thead>
<tr>
<th>Country</th>
<th>Registrations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>16</td>
</tr>
<tr>
<td>Italy</td>
<td>10</td>
</tr>
<tr>
<td>Mexico</td>
<td>2</td>
</tr>
<tr>
<td>US</td>
<td>2</td>
</tr>
<tr>
<td>Canada</td>
<td>1</td>
</tr>
<tr>
<td>India</td>
<td>1</td>
</tr>
<tr>
<td>Netherlands</td>
<td>1</td>
</tr>
<tr>
<td>Nigeria</td>
<td>1</td>
</tr>
<tr>
<td>Russia</td>
<td>1</td>
</tr>
<tr>
<td>Tunisia</td>
<td>1</td>
</tr>
</tbody>
</table>

Number of registrations
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 10:30-10:40 | Touché 2021 Welcome  
*Alexander Bondarenko* |
| 10:40-10:50 | Overview of Task 1 on Argument Retrieval for Controversial Questions  
*Lukas Gienapp* |
| 10:50-11:05 | Exploring Argument Retrieval for Controversial Questions Using Retrieve and Re-rank Pipelines  
*[paper]*  
*Raunak Agarwal, Andrei Koniaev, Robin Schaefer* |
| 11:05-11:20 | Exploring Document Expansion for Argument Retrieval  
*[paper]*  
*Alina Mailach, Denise Arnold, Stefan Eysoldt, Simon Kleine* |
| 11:20-11:35 | Team Skeletor at Touché 2021: Argument Retrieval and Visualization for Controversial Questions  
*[paper]*  
*Kevin Ros, Carl Edwards, Heng Ji, ChengXiang Zhai* |
| 11:35-12:00 | Lightning talks  
*Touché Task 1 participants* |
| 12:00-13:00 | Lunch break |
# Session 2: Argument Retrieval for Comparative Questions

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
</table>
| 13:00-13:30 | **Keynote:** Theory-based Argument Quality for Advanced Argument Retrieval: Opportunities and Challenges  
Anne Lauscher                                           |
| 13:30-13:40 | Overview of Task 2 on Argument Retrieval for Comparative Questions  
Alexander Bondarenko                                     |
| 13:40-13:50 | DistilBERT-based Argumentation Retrieval for Answering Comparative Questions  
Alaa Alhamzeh, Mohamed Bouhaouel, Előd Egyed-Zsigmond, Jelena Mitrović |
| 13:50-14:00 | Retrieving Comparative Arguments using Ensemble Methods and Neural Information Retrieval  
Viktoriia Chekalina, Alexander Panchenko                     |
| 14:00-14:10 | Lightning talks  
*Touche Task 2 participants*                                    |
| 14:10-14:30 | Panel discussion and closing remarks                                                               |

[paper]
## Session 2: Argument Retrieval for Comparative Questions

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Presenter(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13:00-13:30</td>
<td><strong>Keynote:</strong> Theory-based Argument Quality for Advanced Argument Retrieval: Opportunities and Challenges</td>
<td>Anne Lauscher</td>
</tr>
<tr>
<td>13:30-13:40</td>
<td>Overview of Task 2 on Argument Retrieval for Comparative Questions</td>
<td>Alexander Bondarenko</td>
</tr>
<tr>
<td>13:50-14:00</td>
<td>Retrieving Comparative Arguments using Ensemble Methods and Neural Information Retrieval [paper]</td>
<td>Viktoriia Chekalina, Alexander Panchenko</td>
</tr>
<tr>
<td>14:00-14:10</td>
<td>Lightning talks</td>
<td>Touché Task 2 participants</td>
</tr>
<tr>
<td>14:10-14:30</td>
<td>Panel discussion and closing remarks</td>
<td></td>
</tr>
</tbody>
</table>

Spoiler: Touché will run again at CLEF 2022
Session 1: Argument Retrieval for Controversial Questions

Moderator: Lukas Gienapp
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
Example topic for Task 1:

<table>
<thead>
<tr>
<th>Title</th>
<th>Should hate speech be penalized more?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Given the increasing amount of online hate speech, a user questions the necessity and legitimacy of taking legislative action to punish or inhibit hate speech.</td>
</tr>
<tr>
<td>Narrative</td>
<td>Highly relevant arguments include those that take a stance in favor of or opposed to stronger legislation and penalization of hate speech and that offer valid reasons for either stance. Relevant arguments talk about [...]</td>
</tr>
</tbody>
</table>
Touché: Argument Retrieval

Data

Document Collection

- 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]
- Newest/largest version of the corpus (∼ 400,000 structured arguments)

Additional Data

- 50 topics from last year
- relevance judgements from last year
Touché: Argument Retrieval

Statistics

- Submissions: 21 participating teams (up from 13)
- Nicknames: Real or fictional fencers / swordsmen (e.g., Zorro)
- Approaches: 55 valid runs were evaluated (up from 30)
- Baseline: DirichletLM (Lucene Implementation), args.me
- Topics: 50 new topics
- Evaluation: 3,711 manual relevance and quality judgments (nDCG@5)
Argument retrieval: How good are the results?

- Evaluation w.r.t. argument *relevance* and argument *quality* (new this year)
- Top-5 pooling, removing duplicates
- 3,711 unique arguments (text passages)
- Annotation by eight graduate and undergraduate student volunteers, computer science background
  - Different from last year: crowdsourcing difficult for argument quality
  - Pilot study agreement similar to previous expert studies, follow-up discussion to ensure uniform annotations
- nDCG@5 for relevance and quality
**Touché: Argument Retrieval**

**Evaluation**

**Relevance:** How well do arguments fit the topic?

- **Spam**
- **Not relevant**
- **Relevant**
- **Highly relevant**

**Topic:** *Should hate speech be penalized more?*

- 🙌👍: Takes stance for/against stronger legislation on hate speech
- 👍: About hate speech, but not legal aspects
- 👎: Argumentative, but not about hate speech
- ⚠️: Not an argument
Rhetorical quality: How well-written is an argument?

Gender is a social construct because we are told when we are first born by a dude what gender but if he didn’t tell us that we wouldn’t have a gender its only cause he told us that gender that we are that gender.

(Topic: Is gender a social construct?)

Cancel culture gives a voice to disenfranchised or less powerful people. It is a way to acknowledge that you don’t have to have the power to change structural inequality. Even if they don’t have the power to change all of public sentiment, for many individuals, it is the first time they do have a voice in public discourse.

(Topic: Is cancel culture good for society?)

We labeled the quality regardless of relevance:

👍🏻 Proper language, good structure, good grammar, easy to follow
👍 Proper language but broken logic / hard to follow, or vice versa
👎 Profanity, hard to follow, hard to read, many grammar issues
## Touché: Argument Retrieval

### Results

(a) Highest relevance score per team

<table>
<thead>
<tr>
<th>Team</th>
<th>nDCG@5</th>
<th>Rele.</th>
<th>Qual.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dread Pirate Roberts</td>
<td>0.808</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Swordsman</strong></td>
<td>0.756</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Elrond*</td>
<td>0.720</td>
<td>0.809</td>
<td></td>
</tr>
<tr>
<td>Pippin Took*</td>
<td>0.705</td>
<td>0.798</td>
<td></td>
</tr>
<tr>
<td>Robin Hood</td>
<td>0.691</td>
<td>0.756</td>
<td></td>
</tr>
<tr>
<td>Asterix*</td>
<td>0.681</td>
<td>0.802</td>
<td></td>
</tr>
<tr>
<td>Mercutio</td>
<td>0.678</td>
<td>0.804</td>
<td></td>
</tr>
<tr>
<td>::</td>
<td>::</td>
<td>::</td>
<td></td>
</tr>
<tr>
<td><strong>Swordsman</strong></td>
<td>0.626</td>
<td>0.796</td>
<td></td>
</tr>
</tbody>
</table>

(* different runs (systems) from the same team; baseline DirichletLM ranking is shown in bold; highest results from 2020 are in gray (no quality).
## Touché: Argument Retrieval

### Results

#### (a) Highest relevance score per team

<table>
<thead>
<tr>
<th>Team</th>
<th>nDCG@5</th>
<th>Rele.</th>
<th>Qual.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dread Pirate Roberts</td>
<td>0.808</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Swordsman</td>
<td>0.756</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Elrond*</td>
<td>0.720</td>
<td>0.809</td>
<td></td>
</tr>
<tr>
<td>Pippin Took*</td>
<td>0.705</td>
<td>0.798</td>
<td></td>
</tr>
<tr>
<td>Robin Hood</td>
<td>0.691</td>
<td>0.756</td>
<td></td>
</tr>
<tr>
<td>Asterix*</td>
<td>0.681</td>
<td>0.802</td>
<td></td>
</tr>
<tr>
<td>Mercutio</td>
<td>0.678</td>
<td>0.804</td>
<td></td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td></td>
</tr>
<tr>
<td>Swordsman</td>
<td>0.626</td>
<td>0.796</td>
<td></td>
</tr>
</tbody>
</table>

(*) different runs (systems) from the same team; baseline DirichletLM ranking is shown in bold; highest results from 2020 are in gray (no quality).

#### (b) Highest quality score per team

<table>
<thead>
<tr>
<th>Team</th>
<th>nDCG@5</th>
<th>Qual.</th>
<th>Rele.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heimdall</td>
<td>0.841</td>
<td>0.639</td>
<td></td>
</tr>
<tr>
<td>Skeletor</td>
<td>0.827</td>
<td>0.666</td>
<td></td>
</tr>
<tr>
<td>Asterix*</td>
<td>0.818</td>
<td>0.663</td>
<td></td>
</tr>
<tr>
<td>Elrond*</td>
<td>0.817</td>
<td>0.681</td>
<td></td>
</tr>
<tr>
<td>Pippin Took*</td>
<td>0.814</td>
<td>0.683</td>
<td></td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td></td>
</tr>
<tr>
<td>Swordsman</td>
<td>0.796</td>
<td>0.626</td>
<td></td>
</tr>
</tbody>
</table>
Baseline scores lower, yet majority of teams outperforms vs. few last year
Quality evaluation shows promising results, improvements w.r.t. baseline
Two trends among submissions:
- Deploying “classical” retrieval models with parameter optimization
- Increased focus on ML for query expansion and assessing quality
Baseline scores lower, yet majority of teams outperforms vs. few last year

Quality evaluation shows promising results, improvements w.r.t. baseline

Two trends among submissions:
- Deploying “classical” retrieval models with parameter optimization
- Increased focus on ML for query expansion and assessing quality

All approaches indexed the corpus themselves, no use of args.me API

All approaches used 2020 relev. judgments for training or parameter tuning

Extended a collection of relevance judgments, additionally argument quality judgments
# Touché: Argument Retrieval

## Submitted Papers

<table>
<thead>
<tr>
<th>Team</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asterix</td>
<td>Smerilli et al.: A Search Engine System for Touché Argument Retrieval task to answer Controversial Questions</td>
</tr>
<tr>
<td>Batman</td>
<td>Raimondi et al.: Step approach to information retrieval</td>
</tr>
<tr>
<td>Pirate Roberts</td>
<td>Akiki et al.: Learning to Rank Arguments with Feature Selection</td>
</tr>
<tr>
<td>G. Ishikawa</td>
<td>Carnelos et al.: Argument Retrieval for Controversial Questions</td>
</tr>
<tr>
<td>Heimdall</td>
<td>Gienapp.: Quality-aware Argument Retrieval with Topical Clustering</td>
</tr>
<tr>
<td><strong>Hua Mulan</strong></td>
<td>Mailach et al.: Exploring Document Expansion for Argument Retrieval</td>
</tr>
<tr>
<td>J.-P. Polnareff</td>
<td>Alecci et al.: Development of an IR System for Argument Search</td>
</tr>
<tr>
<td><strong>Macbeth</strong></td>
<td>Agarwal.: Exploring Argument Retrieval for Controversial Questions Using Retrieve and Re-rank Pipelines</td>
</tr>
<tr>
<td>Pippin Took</td>
<td>Togni.: Exploring Approaches for Touché Task 1</td>
</tr>
<tr>
<td>Shanks</td>
<td>Berno et al.: Shanks Touché Homework</td>
</tr>
<tr>
<td><strong>Skeletor</strong></td>
<td>Ros et al.: Argument Retrieval and Visualization</td>
</tr>
<tr>
<td>Yeagerists</td>
<td>Green et al.: Exploring BERT Synonyms and Quality Prediction for Argument Retrieval</td>
</tr>
<tr>
<td>Baseline</td>
<td>Lucene Implementation of DirichletLM [Zhai &amp; Lafferty 2004]</td>
</tr>
<tr>
<td></td>
<td>Good results in pilot study [Potthast et al. 2019]</td>
</tr>
<tr>
<td></td>
<td>args.me [Wachsmuth et al. 2017]</td>
</tr>
</tbody>
</table>

Marked in **bold** are featured talks.
Session 1: Participants’ paper presentations
# Touché: Argument Retrieval

**Session 2: Argument Retrieval for Comparative Questions**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
</table>
| 13:00-13:30| **Keynote:** Theory-based Argument Quality for Advanced Argument Retrieval: Opportunities and Challenges  
*Anne Lauscher* |
| 13:30-13:40| Overview of Task 2 on Argument Retrieval for Comparative Questions  
*Alejandro Bondarenko* |
| 13:40-13:50| DistilBERT-based Argumentation Retrieval for Answering Comparative Questions [*paper*]  
*Alaa Alhamzeh, Mohamed Bouhaouel, Előd Egyed-Zsigmond, Jelena Mitrović* |
| 13:50-14:00| Retrieving Comparative Arguments using Ensemble Methods and Neural Information Retrieval [*paper*]  
*Viktória Chekalina, Alexander Panchenko* |
| 14:00-14:10| Lightning talks  
*Touché Task 2 participants* |
| 14:10-14:30| Panel discussion and closing remarks |

**Moderator: Alexander Bondarenko**
Keynote:

Theory-based Argument Quality for Advanced Argument Retrieval: Opportunities and Challenges

Anne Lauscher, Bocconi University in Milan

[webpage]
Touché: Argument Retrieval

Shared Task

Task 2: Answering comparative questions with arguments

- **Scenario:** Users face personal decisions from everyday life
- **Goal:** Help to come to an informed decision on the comparison
- **Task:** Retrieve and rank 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]
Statistics

- Registrations: 13 teams, incl. for both tasks (vs. 18 last year)
- Nicknames: Real or fictional fencers / swordsmen (e.g., Katana)
- Submissions: 6 participating teams (vs. 5 last year)
- Approaches: 19 valid runs were evaluated (vs. 11 last year)
- Baseline: BM25F-based ChatNoir [chatnoir.eu]
- Evaluation: 2,076 manual relevance and quality judgments (nDCG@5) (vs. 1,783 last year)
Example topic for Task 2:

<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Should I major in philosophy or psychology?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>A soon-to-be high-school graduate finds themself at a crossroad in their life. [...] searching for information about the differences and similarities, advantages and disadvantages of majoring in either of them (e.g., with respect to career opportunities).</td>
</tr>
<tr>
<td><strong>Narrative</strong></td>
<td>Relevant documents will overview one of the two majors in terms of career prospects or developed new skills, or they will provide a list of reasons to major in one or the other. [...] Not relevant are study program and university advertisements or general descriptions of the disciplines that do not mention benefits, advantages, or pros/cons.</td>
</tr>
</tbody>
</table>
Argument retrieval: How good are web documents with arguments?

Classical (TREC-style) IR relevance judgments

Not relevant    Relevant    Highly relevant

Who is a better *pet*, a *cat* or a *dog*?

Comparing cats versus dogs as pets
Information about either cats or dogs as pets
Everything else: often ads
Touché: Argument Retrieval
Evaluation

Argument retrieval: How good are web documents with arguments?

Rhetorical quality: How well written?

扒扒扒 the best !!! Don't even try to argue with me. Yeah, ye-ah, yeah (Grrrr) I have always had cats, they are sooo cooool, and dogs just suck.

A cat’s independent nature generally helps them deal better than dogs with being left alone. Cats also tend to live longer than dogs, which is sometimes a consideration when searching for a lifelong furry companion.

We labeled the quality regardless of relevance

👏 Proper language, good structure, good grammar, easy to follow
👍 Proper language but broken logic / hard to follow, or vice versa
👎 Profanity, hard to follow, hard to read, many grammar issues
## Results

(a) Highest **relevance** score per team

<table>
<thead>
<tr>
<th>Team</th>
<th>nDCG@5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rele.</td>
</tr>
<tr>
<td>Bilbo Baggins</td>
<td>0.580</td>
</tr>
<tr>
<td>Puss in Boots</td>
<td>0.568</td>
</tr>
<tr>
<td>Katana*</td>
<td>0.489</td>
</tr>
<tr>
<td>Thor</td>
<td>0.478</td>
</tr>
<tr>
<td>Rayla*</td>
<td>0.473</td>
</tr>
<tr>
<td>Puss in Boots</td>
<td>0.422</td>
</tr>
</tbody>
</table>

(*) different runs (systems) from the same team; baseline **ChatNoir** ranking is shown in **bold**; highest results from 2020 are in **gray** (no quality).
### Results

**(a) Highest relevance score per team**

<table>
<thead>
<tr>
<th>Team</th>
<th>nDCG@5</th>
<th>Rele.</th>
<th>Qual.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilbo Baggins</td>
<td>0.580</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Puss in Boots</td>
<td>0.568</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Katana*</td>
<td>0.489</td>
<td>0.675</td>
<td></td>
</tr>
<tr>
<td>Thor</td>
<td>0.478</td>
<td>0.680</td>
<td></td>
</tr>
<tr>
<td>Rayla*</td>
<td>0.473</td>
<td>0.670</td>
<td></td>
</tr>
<tr>
<td>Puss in Boots</td>
<td>0.422</td>
<td>0.636</td>
<td></td>
</tr>
</tbody>
</table>

**(b) Highest quality score per team**

<table>
<thead>
<tr>
<th>Team</th>
<th>nDCG@5</th>
<th>Qual.</th>
<th>Rele.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilbo Baggins</td>
<td>–</td>
<td>0.580</td>
<td></td>
</tr>
<tr>
<td>Puss in Boots</td>
<td>–</td>
<td>0.568</td>
<td></td>
</tr>
<tr>
<td>Rayla*</td>
<td>0.688</td>
<td>0.466</td>
<td></td>
</tr>
<tr>
<td>Katana*</td>
<td>0.684</td>
<td>0.460</td>
<td></td>
</tr>
<tr>
<td>Thor</td>
<td>0.680</td>
<td>0.478</td>
<td></td>
</tr>
<tr>
<td>Puss in Boots</td>
<td>0.636</td>
<td>0.422</td>
<td></td>
</tr>
</tbody>
</table>

(*) different runs (systems) from the same team; baseline **ChatNoir** ranking is shown in bold; highest results from 2020 are in gray (no quality).
### Results

#### (a) Highest relevance score per team

<table>
<thead>
<tr>
<th>Team</th>
<th>nDCG@5</th>
<th>Rele.</th>
<th>Qual.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilbo Baggins</td>
<td>0.580</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Puss in Boots</td>
<td>0.568</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Katana*</td>
<td>0.489</td>
<td>0.675</td>
<td></td>
</tr>
<tr>
<td>Thor</td>
<td>0.478</td>
<td>0.680</td>
<td></td>
</tr>
<tr>
<td>Rayla*</td>
<td>0.473</td>
<td>0.670</td>
<td></td>
</tr>
<tr>
<td>Puss in Boots</td>
<td>0.422</td>
<td>0.636</td>
<td></td>
</tr>
</tbody>
</table>

(*) different runs (systems) from the same team; baseline ChatNoir ranking is shown in bold; highest results from 2020 are in gray (no quality).

#### (b) Highest quality score per team

<table>
<thead>
<tr>
<th>Team</th>
<th>nDCG@5</th>
<th>Qual.</th>
<th>Rele.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilbo Baggins</td>
<td>–</td>
<td>0.580</td>
<td></td>
</tr>
<tr>
<td>Puss in Boots</td>
<td>–</td>
<td>0.568</td>
<td></td>
</tr>
<tr>
<td>Rayla*</td>
<td>0.688</td>
<td>0.466</td>
<td></td>
</tr>
<tr>
<td>Katana*</td>
<td>0.684</td>
<td>0.460</td>
<td></td>
</tr>
<tr>
<td>Thor</td>
<td>0.680</td>
<td>0.478</td>
<td></td>
</tr>
<tr>
<td>Puss in Boots</td>
<td>0.636</td>
<td>0.422</td>
<td></td>
</tr>
</tbody>
</table>

### Team Representation

<table>
<thead>
<tr>
<th>Team</th>
<th>Represent.</th>
<th>Query processing</th>
<th>(Re-)Ranking features</th>
<th>(Re-)Ranking method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilbo Baggins</td>
<td>Bag of words</td>
<td>N. entities, comp. aspects</td>
<td>Credibility, support, ...</td>
<td>Mean sum of scores</td>
</tr>
<tr>
<td>Puss in Boots</td>
<td>Bag of words</td>
<td>—</td>
<td>SpamRank</td>
<td>BM25F</td>
</tr>
<tr>
<td>Katana</td>
<td>Bag of words, BERT</td>
<td>—</td>
<td>Comparativeness, relevance, tf-idf</td>
<td>R. Forests, XGBoost, LightGBM, BERT</td>
</tr>
<tr>
<td>Rayla</td>
<td>SBERT</td>
<td>Stop words, lemmas, synonyms/antonyms</td>
<td>Relevance, Page-, SpamRank, argument support (DistilBERT)</td>
<td>Weighted features linear combination</td>
</tr>
<tr>
<td>Thor</td>
<td>Bag of words</td>
<td>Punctuation, synonyms</td>
<td>Premises &amp; claims with TARGER</td>
<td>Re-index &amp; query w. BM25</td>
</tr>
</tbody>
</table>
Touché: Argument Retrieval

Summary

- All approaches re-ranked ChatNoir results
- All approaches used relevance judgments from 2020
- Majority improved over baseline vs. few last year
- Extended a collection of relevance judgments
- Additionally argument quality judgments
- “Best” so far: query processing and expansion, comparative and argumentative (incl. argument quality) features, neural but also BM25
<table>
<thead>
<tr>
<th>Team</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack Sparrow</td>
<td>Thi Kim Hanh Luu, Jan-Niklas Weder. Argument Retrieval for Comparative Questions Based on Independent Features [paper]</td>
</tr>
<tr>
<td>Katana</td>
<td>Viktoria Chekalina, Alexander Panchenko. Retrieving Comparative Arguments using Ensemble Methods and Neural Information Retrieval [paper]</td>
</tr>
<tr>
<td>Rayla</td>
<td>Alaa Alhamzeh, Mohamed Bouhaouel, Előd Egyed-Zsigmond, Jelena Mitrović. DistilBERT-based Argumentation Retrieval for Answering Comparative Questions [paper]</td>
</tr>
<tr>
<td>Lab overview</td>
<td>Bondarenko et al. Overview of Touché 2021: Argument Retrieval [paper] [CLEF 2021 Working Notes]</td>
</tr>
</tbody>
</table>
Session 2: Participants’ paper presentations
Session 3: Panel discussion and closing remarks

Moderator: Alexander Bondarenko
Touché: Argument Retrieval

Statistics over two years

- Registrations: 64 teams
- Submissions: 44 participating teams
- Approaches: 129 valid runs were evaluated
- Evaluation: 12,832 manual relevance judgments (5,787 additional quality)

Number of registrations:
- Germany: 33
- Italy: 11
- India: 3
- US: 3
- France: 2
- Mexico: 2
- Netherlands: 2
- Russia: 2
- Canada: 1
- China: 1
- Nigeria: 1
- Pakistan: 1
- Switzerland: 1
- Tunisia: 1

Number of registrations
Touché: Argument Retrieval

Summary

- Platform for argument retrieval researchers  [touche.webis.de]
- Argument relevance / quality corpora / rankings
- Tools for submission and evaluation  [tira.io]

- All (almost) used labeled data from 2020
- Majority improved over baselines
24 participant working notes published in proceedings [CEUR-WS.org]

- Dumani, Schenkel. 

- Nilles, Dumani, Schenkel. 

- Thakur, Reimers, Rücklé, Srivastava, Gurevych. 

- Cherumanal, Spina, Scholer, Croft. 
  Evaluating Fairness in Argument Retrieval. CIKM 2021.
Task 1: Argument Retrieval for Controversial Questions

- **Scenario:** Users search for argument gist on controversial topics

- **Task:** Retrieve and rank sentences (main claim and premise) that convey key points pertinent to the controversial topic

- **Data:** 400,000 “arguments” (short text passages) [args.me]
Task 1: Argument Retrieval for Controversial Questions

- **Scenario:** Users search for argument gist on controversial topics
- **Task:** Retrieve and rank sentences (main claim and premise) that convey key points pertinent to the controversial topic
- **Data:** 400,000 “arguments” (short text passages) [args.me]

Task 2: Argument Retrieval for Comparative Questions

- **Scenario:** Support users to come to informed decisions
- **Task:** Retrieve relevant argumentative passages for compared objects and detect their respective stances w.r.t the objects
- **Data:** > 1 million text passages (from web documents)
Task 3: Image Retrieval for Arguments

- **Scenario:** Users search for images to corroborate their argumentation.
- **Task:** Retrieve and rank images that can be used to support or attack a given stance.
- **Data:** > 10,000 web images with respective web documents.

*Should hate speech be banned?*
Free discussion

thank you!
Touché: Argument Retrieval

Related Publications