TOUCHE
2023
Touché @ CLEF 2023
Argument and Causal Retrieval

Alexander Bondarenko
Maik Fröbe
Johannes Kiesel
Ferdinand Schlatt
Valentin Barriere
Brian Ravenet
Léo Hemamou
Simon Luck
Jan Heinrich Reimer
Benno Stein
Martin Potthast
Matthias Hagen

touche.webis.de
Task 1: Argument Retrieval for Controversial Questions
- Retrieve relevant and high-quality argumentative documents, detect stance
Task 1: Argument Retrieval for Controversial Questions
- Retrieve relevant and high-quality argumentative documents, detect stance

Task 2: Evidence Retrieval for Causal Questions
- Retrieve and rank causality-related documents and detect causal stance
Task 1: Argument Retrieval for Controversial Questions
   - Retrieve relevant and high-quality argumentative documents, detect stance

Task 2: Evidence Retrieval for Causal Questions
   - Retrieve and rank causality-related documents and detect causal stance

Task 3: Image Retrieval for Arguments
   - Retrieve images for each stance (pro / con) that support that stance
Task 1: Argument Retrieval for Controversial Questions
- Retrieve relevant and high-quality argumentative documents, detect stance

Task 2: Evidence Retrieval for Causal Questions
- Retrieve and rank causality-related documents and detect causal stance

Task 3: Image Retrieval for Arguments
- Retrieve images for each stance (pro / con) that support that stance

Task 4: Multilingual and Multi-target Stance Classification
- Detect the stance of a comment on a proposal
Registrations: 41 teams (vs. 58 teams last year)
Nicknames: Real or fictional fencers / swordfighters (e.g., Zorro)
Submissions: 7 participating teams (vs. 23 last year)
Approaches: 30 valid runs were evaluated (vs. 84 last year)
Judgments: 1,500 web documents, 700 images, 25,000 comments
# Touché: Argument and Causal Retrieval Workshop Program

**Thursday, September 21. Touché: Argument and Causal Retrieval Workshop**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:30-11:35</td>
<td>Welcome</td>
</tr>
<tr>
<td>11:35-11:45</td>
<td><strong>Session 1: Argument Retrieval for Controversial Questions</strong></td>
</tr>
<tr>
<td>11:45-12:00</td>
<td>Overview of Task 1 on Argument Retrieval for Controversial Questions (Alexander Bondarenko) [paper]</td>
</tr>
<tr>
<td>11:45-12:00</td>
<td>Argument Quality Prediction for Ranking Documents (Moritz Plenz) [paper]</td>
</tr>
<tr>
<td>12:00-12:10</td>
<td><strong>Session 2: Evidence Retrieval for Causal Questions</strong></td>
</tr>
<tr>
<td>12:10-12:20</td>
<td>Overview of Task 2 on Evidence Retrieval for Causal Questions (Alexander Bondarenko) [paper]</td>
</tr>
<tr>
<td>12:10-12:20</td>
<td>Evidence Retrieval for Causal Questions Using Query Expansion and Reranking</td>
</tr>
<tr>
<td>12:20-12:30</td>
<td><strong>Session 3: Image Retrieval for Arguments</strong></td>
</tr>
<tr>
<td>12:30-12:45</td>
<td>Overview of Task 3 on Image Retrieval for Arguments (Johannes Kiesel) [paper]</td>
</tr>
<tr>
<td>12:30-12:45</td>
<td>Matching Images and Keywords with CLIP (Fatihah Ulya Hakiem)</td>
</tr>
<tr>
<td>12:45-13:00</td>
<td>Comparing Image Generation, Stance Detection and Feature Matching for Image Retrieval for Arguments (Sarah Bachinger, Maximilian Enderling, and Max Möbius)</td>
</tr>
<tr>
<td>13:00-14:00</td>
<td>Lunch</td>
</tr>
</tbody>
</table>
## Session 4: Multilingual Multi-Target Stance Classification

<table>
<thead>
<tr>
<th>Time</th>
<th>Session Title</th>
<th>Presenter/Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00-14:10</td>
<td>Overview of Task 4 on Multilingual Multi-Target Stance Classification (Valentin Barriere)</td>
<td>[paper]</td>
</tr>
<tr>
<td>14:10-14:25</td>
<td>Intra-Multilingual Multi-Target Stance Classification using BERT (Karla Schaefer)</td>
<td></td>
</tr>
</tbody>
</table>

**Special Session**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session Title</th>
<th>Presenter/Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:25-14:45</td>
<td><strong>Best of Touché 2022</strong>: Neural Image Retrieval for Argumentation (Tobias Schreieder and Jan Braker)</td>
<td></td>
</tr>
<tr>
<td>14:45-15:00</td>
<td><strong>Closing</strong>: remarks, plenary discussion, future plans</td>
<td></td>
</tr>
</tbody>
</table>

Poster session takes place on **September 18** for all CLEF participants.
### Touché: Argument and Causal Retrieval

Workshop Program

**Session 4: Multilingual Multi-Target Stance Classification**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00-14:10</td>
<td>Overview of Task 4 on Multilingual Multi-Target Stance Classification (Valentin Barriere) [paper]</td>
</tr>
<tr>
<td>14:10-14:25</td>
<td>Intra-Multilingual Multi-Target Stance Classification using BERT (Karla Schaefer)</td>
</tr>
</tbody>
</table>

**Special Session**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:25-14:45</td>
<td>Best of Touché 2022: Neural Image Retrieval for Argumentation (Tobias Schreieder and Jan Braker)</td>
</tr>
<tr>
<td>14:45-15:00</td>
<td>Closing: remarks, plenary discussion, future plans</td>
</tr>
</tbody>
</table>

Poster session takes place on **September 18** for all CLEF participants

---

Spoiler: Touché will run again at CLEF 2024 (but with new tasks)

Submit your extended working notes to ECIR 2024
Session 1: Argument Retrieval for Controversial Questions

Moderator: Alexander Bondarenko
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
Task 1: Argument Retrieval for Controversial Questions

- Scenario: Users search for arguments on controversial topics
- Task: Retrieve and rank relevant and high-quality argument documents; identify the document stance
- Data: ClueWeb22-B (200 million documents); also available via [ChatNoir]

- Run submissions similar to “classical” TREC tracks
- Software submissions in TIRA [tira.io]
**Example topic for Task 1:**

<table>
<thead>
<tr>
<th>Title</th>
<th>Should teachers get tenure?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>A user has heard that some countries do give teachers tenure and others don’t. Interested in the reasoning for or against tenure, the user searches for positive and negative arguments [...]</td>
</tr>
<tr>
<td>Narrative</td>
<td>Highly relevant arguments make a clear statement about tenure for teachers in schools or universities. Relevant arguments consider tenure more generally, not specifically for teachers, or, instead of talking about tenure, consider the situation of teachers’ financial independence.</td>
</tr>
</tbody>
</table>
Touché: Argument and Causal Retrieval

Evaluation

Document relevance (nDCG@10):

👍👍 High relevant to the topic
👍👍 (Partially) relevant to the topic
👎👎 Everything else

Rhetorical argument quality (nDCG@10):

👍👍 Proper language, good structure, good grammar, easy to follow
👍👍 Proper language but broken logic / hard to follow, or vice versa
👎👎 Profanity, hard to follow, grammar issues / no arguments at all

Document stance (macro-avg. F1):

 сочета Pro, con, neutral, no stance
1 team (Renji Abarai) submitted 7 runs

Baseline (Puss in Boots): BM25F-based ChatNoir; Flan-T5 for stance

747 documents manually judged (relevance, argument quality, and stance)
Session 2: Evidence Retrieval for Causal Questions

Moderator: Alexander Bondarenko (on behalf of Ferdinand Schlatt)
Cause–Effect relationships:

- An integral part of human reasoning; an association of two ideas because of experiencing their regular conjunction [Khoo, 2002]

- A cause is an insufficient but necessary part of unnecessary but sufficient conditions for an effect (INUS) [Mackie, 1980]

<table>
<thead>
<tr>
<th></th>
<th>Fuel-soaked Rag</th>
<th>House Fire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sufficient condition</td>
<td>{Fuel-soaked rag, spark, wooden house, ...}</td>
<td></td>
</tr>
<tr>
<td>Unnecessary condition</td>
<td>Other possible conditions exist</td>
<td></td>
</tr>
<tr>
<td>Necessary part</td>
<td>Without the rag, no fire would happen</td>
<td></td>
</tr>
<tr>
<td>Insufficient part</td>
<td>Only the rag would not cause the fire</td>
<td></td>
</tr>
</tbody>
</table>
Task 2: Retrieving and analyzing evidence for causal claims

- **Scenario:** Users want to know if two events are causally related
- **Goal:** Help to find evidence for or against a causal claim
- **Task:** Retrieve and rank documents containing evidence identify the document stance
- **Data:** ClueWeb22-B (200 million documents); also available via [ChatNoir]

- Run submissions similar to “classical” TREC tracks
- **Software submissions in TIRA** [tira.io]
### Example topic for Task 2:

<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Could sun exposure cause hair loss?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cause</strong></td>
<td>sun exposure</td>
</tr>
<tr>
<td><strong>Effect</strong></td>
<td>hair loss</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>A user is wondering how to protect against hair loss and specifically, if an increased exposure to sunlight can cause hair loss.</td>
</tr>
<tr>
<td><strong>Narrative</strong></td>
<td>Highly relevant documents will provide information on a potential causal connection between exposure to sunlight and hair loss (medically: alopecia). This includes documents stating or giving evidence that the first is (or is not) a cause of the other. Documents stating that there is not enough evidence to decide either way are also highly relevant. […]</td>
</tr>
</tbody>
</table>
Touché: Argument and Causal Retrieval

Evaluation

Document relevance (nDCG@5):

👍👍 High relevant to the topic
👍 (Partially) relevant to the topic
/thumb down Everything else

Document stance (macro-avg. F1):

✅ Supporting Evidence
❌ Refuting Evidence
❓ Neutral Evidence
1 team (He-Man) submitted 3 runs

Baseline (Puss in Boots): BM25F-based ChatNoir; Flan-T5 for stance

718 documents manually judged (relevance and stance)
Evidence Retrieval for Causal Questions Using Query Expansion and Reranking

Aron Gaden, Niklas Rausch, Bruno Reinhold, and Lukas Zeit-Altpeter

Friedrich-Schiller-Universität Jena
Query: Could *sun exposure* cause *hair loss*?

First-stage retrieval with ChatNoir: (1) original and (2) expanded query

Dependency tree parsing to extract cause, effect, and causal phrase

Query expansion with synonyms from CauseNet

Query expansion with terms generated by ChatGPT
Touché: Argument and Causal Retrieval

Approach

- Query: Could *sun exposure cause hair loss*?
- First-stage retrieval with ChatNoir: (1) original and (2) expanded query
- Dependency tree parsing to extract cause, effect, and causal phrase
- Query expansion with synonyms from CauseNet
- Query expansion with terms generated by ChatGPT

- Re-ranking using a position bias
- Dependency tree parsing: cause, effect, and causal phrase (in documents)
- Documents containing the causal relationship from the original query earlier in the document are ranked higher
# Touché: Argument and Causal Retrieval

## Results

<table>
<thead>
<tr>
<th>Team</th>
<th>Run Tag</th>
<th>nDCG@5 Relevance</th>
<th>F1 macro Stance</th>
</tr>
</thead>
<tbody>
<tr>
<td>He-Man</td>
<td>no_expansion_rerank</td>
<td>0.657†</td>
<td>–</td>
</tr>
<tr>
<td><strong>Puss in Boots</strong></td>
<td><strong>ChatNoir</strong></td>
<td>0.585</td>
<td>0.256</td>
</tr>
<tr>
<td>He-Man</td>
<td>gpt_expansion_rerank</td>
<td>0.374</td>
<td>–</td>
</tr>
<tr>
<td>He-Man</td>
<td>causenet_expansion_rerank</td>
<td>0.268</td>
<td>–</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Team</th>
<th>Run Tag</th>
<th>nDCG@5 Relevance</th>
<th>F1 macro Stance</th>
</tr>
</thead>
<tbody>
<tr>
<td>He-Man</td>
<td>no_expansion_rerank</td>
<td>0.657†</td>
<td>–</td>
</tr>
<tr>
<td><strong>Puss in Boots</strong></td>
<td><strong>ChatNoir</strong></td>
<td>0.585</td>
<td>0.256</td>
</tr>
<tr>
<td>He-Man</td>
<td>gpt_expansion_rerank</td>
<td>0.374</td>
<td>–</td>
</tr>
<tr>
<td>He-Man</td>
<td>causenet_expansion_rerank</td>
<td>0.268</td>
<td>–</td>
</tr>
</tbody>
</table>

- Simple yet effective approach
- A high-precision but low-recall solution
- **Error:** (drinking wine, blood urine) \(\rightarrow\) (eating food, diarrhea)
- Room for future research
Session 3: Image Retrieval for Arguments

Moderator: Johannes Kiesel
Task 3: Image retrieval for arguments

- **Scenario:** Users search for images to corroborate their argumentation
- **Task:** Retrieve and rank images to support or attack a given stance
- **Data:** 56,000 web images with respective web documents and Google Cloud Vision data

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

- **Submissions:** 3 participating teams (+ baseline)
- **Approaches:** 12 valid runs were evaluated (+ baseline)
- **Baseline:** Re-implementation of Aramis approach
- **Evaluation:** 7,000 images-topic pairs judged manually

- Matthew Lewis as Neville Longbottom in “Harry Potter”
- George Takei as Hikaru Sulu in “Star Trek”
- Patrick Stewart as Jean-Luc Picard in “Star Trek”
- Minsc (and Boo) by u/Kazuliski (on Reddit)
<table>
<thead>
<tr>
<th>Team</th>
<th>Run</th>
<th>Precision@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neville Longbottom</td>
<td>clip_chatgpt_args.raw</td>
<td>0.785 0.338 0.222</td>
</tr>
<tr>
<td>Hikaru Sulu</td>
<td>Keywords</td>
<td>0.664 0.350 0.185</td>
</tr>
<tr>
<td>Jean-Luc Picard</td>
<td>No stance detection</td>
<td>0.523 0.292 0.162</td>
</tr>
<tr>
<td>Minsc</td>
<td>Baseline (Aramis)</td>
<td>0.376 0.194 0.102</td>
</tr>
<tr>
<td>Boromir</td>
<td>On 2022 data</td>
<td>0.878 0.768 0.425</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Team</th>
<th>Run</th>
<th>Precision@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neville Longbottom</td>
<td>clip_chatgpt_args.raw</td>
<td>0.785 0.338 0.222</td>
</tr>
<tr>
<td>Hikaru Sulu</td>
<td>Keywords</td>
<td>0.664 0.350 0.185</td>
</tr>
<tr>
<td>Jean-Luc Picard</td>
<td>No stance detection</td>
<td>0.523 0.292 0.162</td>
</tr>
<tr>
<td>Minsc</td>
<td>Baseline (Aramis)</td>
<td>0.376 0.194 0.102</td>
</tr>
<tr>
<td>Boromir</td>
<td>On 2022 data</td>
<td>0.878 0.768 0.425</td>
</tr>
</tbody>
</table>

### Neville Longbottom

- ChatGPT for generating arguments for topic + stance
- ChatGPT for generating image descriptions for arguments
- CLIP for ranking images by similarity to descriptions
- Experimented with re-ranking using description for other stance or IBM’s debater pro-con score
Session 3: Participants’ paper presentations
Session 4: Multilingual and Multi-target Stance Classification

Moderator: Valentin Barriere
Task 4: Multilingual and Multi-target Stance Classification

- **Scenario:** Stakeholders want to get an overview about citizens’ opinions on an important societal topic
- **Task:** Detect the stance of a comment towards a proposal
- **Data:** 4200 proposals and 20000 comments focused on various topics written in 26 different languages
Task 4: Multilingual and Multi-target Stance Classification

- **Scenario:** Stakeholders want to get an overview about citizens’ opinions on an important societal topic.
- **Task:** Detect the stance of a comment towards a proposal.
- **Data:** 4,200 proposals and 20,000 comments focused on various topics written in 26 different languages.
Data

Example data instance for Task 4:

<table>
<thead>
<tr>
<th>Title</th>
<th>Topic</th>
<th>Proposal</th>
<th>Comment</th>
<th>Stance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focus on anti-aging and longevity research</td>
<td>Health</td>
<td>The EU has presented their green paper on aging, and correctly named the aging . . .</td>
<td>The idea of prevention being better than a cure is nothing new or revolutionary. Rejuvenation . . .</td>
<td>In favor</td>
</tr>
<tr>
<td>Encourage people eat less meat</td>
<td>Climate change</td>
<td>I think it would be great that everyone gets a meat card. You take the card to the store . . .</td>
<td>La valeur nutritionnelle de la viande reste un argument très fort en faveur de la consommation . . .</td>
<td>Against</td>
</tr>
</tbody>
</table>
Touché: Argument and Causal Retrieval
Evaluation

- Subtask 1: Cross-debate classification
- Subtask 2: All-data-available classification
- Baselines: a) always predict the majority class ‘in favor’ (Cavalier Simple) b) multilingual masked language model XLM-R (Cavalier)
- Participants: 2 teams, 8 runs

**Touché: Argument and Causal Retrieval**

**Evaluation**

- **Subtask 1:** Cross-debate classification
- **Subtask 2:** All-data-available classification
- **Baselines:**
  - a) always predict the majority class ‘in favor’ (Cavalier Simple)
  - b) multilingual masked language model XLM-R (Cavalier)
- **Participants:** 2 teams, 8 runs

<table>
<thead>
<tr>
<th>Team</th>
<th>F1 macro</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>en</td>
<td>fr</td>
</tr>
<tr>
<td><strong>Subtask 1: Cross-debate classification</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cavalier</td>
<td>59.4</td>
<td>54.9</td>
</tr>
<tr>
<td>Queen of Swords</td>
<td>44.8</td>
<td>41.3</td>
</tr>
<tr>
<td>Cavalier Simple</td>
<td>24.4</td>
<td>24.2</td>
</tr>
<tr>
<td><strong>Subtask 2: All-data-available classification</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cavalier</td>
<td>57.2</td>
<td>54.6</td>
</tr>
<tr>
<td>Silver Surfer</td>
<td>36.7</td>
<td>33.9</td>
</tr>
<tr>
<td>Queen of Swords</td>
<td>35.1</td>
<td>31.5</td>
</tr>
</tbody>
</table>

© touche.webis.de 2023
Touché: Argument and Causal Retrieval

Session 4: Participant’s paper presentation
Special Session

Moderator: Léo Hemamou

- **Best of Touché 2022**: Neural Image Retrieval for Argumentation (Tobias Schreieder and Jan Braker)
Panel discussion, closing remarks, future plans

Moderators: Alexander Bondarenko and Johannes Kiesel
Touché: Argument and Causal Retrieval
Statistics over 4 Years

- Registrations: 163 teams (avg. 41 per year)
- Submissions: 74 participating teams (avg. 19 per year)
- Approaches: 243 valid runs were evaluated (avg. 61 per year)
- Evaluation: > 30,000 manual judgments
Touché: Argument and Causal Retrieval
Statistics over 4 Years

- Registrations: 163 teams (avg. 41 per year)
- Submissions: 74 participating teams (avg. 19 per year)
- Approaches: 243 valid runs were evaluated (avg. 61 per year)
- Evaluation: > 30,000 manual judgments

- Tasks: Argument Retrieval for Controversial Questions
  Argument Retrieval for Comparative Questions
  Image Retrieval for Arguments
  Evidence Retrieval for Causal Questions
  Multilingual Multi-Target Stance Classification
Touché: Argument and Causal Retrieval

Summary

- Platform for argument and causal retrieval and analysis [touche.webis.de]
- Relevance / quality / stance corpora and runs
- Tools for submission and evaluation [tira.io]
Task 1: Human Value Detection (ValueEval)

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Nicolas Handke, Nicolas Stefanovitch, Bertrand De Longueville Mario Scharfbillig, Henning Wachsmuth, Benno Stein

- Scenario: Users want to find different views (expressed by values) in texts
- Task: Given a text, detect for each sentence
  - Subtask 1: which human values it refers to and
  - Subtask 2: whether it signals (partial) attainment or constraint of the value
- Data: > 3000 news+manifestos, 8 languages, 400 to 800 words each

Example:
The *budget* for last year’s government policies on defence *went out of control*.

Value (Subtask 1): Power: Resources
Attainment (Subtask 2): (Partially) constrained
Task 2: Multilingual Ideology and Power Identification in Parliamentary Debates
Çağrı Çöltekin, Nikola Ljubešić, Katja Meden, Tomaž Erjavec, Vaidas Morkevičius, Matyáš Kopp

- **Scenario:** To better understand how political ideology the position of the speaker affects parliamentary debates

- **Task:** Given a transcribed speech in some language, detect
  - **Subtask 1:** the ideology of the speaker’s party
  - **Subtask 2:** whether the speaker belongs to a governing party (coalition)

- **Data:** Speech samples from multiple national/regional parliaments from the ParlaMint project, and their automatic translations to English

Dataset: https://www.clarin.eu/parlamint
Task 3: Image Retrieval/Generation for Arguments (joint task with ImageCLEF)
Maximilian Heinrich, Johannes Kiesel, Martin Potthast, Benno Stein

- **Scenario:** Users want to better convey arguments (with images)
- **Task:** Retrieve/generate images to reinforce an argument's premise
- **Data:** > 10 000 web images and Stable Diffusion API

**Example:**

**Claim:** Legislation to impose restrictive photo ID requirements has the potential to block millions of American voters

**Premise:** Indiana’s photo ID law barred twelve retired nuns from voting

**Submission:**

- Retrieved from dataset
- Generated “text-image”
- Generated via Stable Diffusion
Open discussion

thank you!