

Touché @ CLEF 2022

Shared Tasks on Argument Retrieval



Alexander Bondarenko

Maik Fröbe

Johannes Kiesel

Shahbaz Syed

Timon Gurcke

Meriem Beloucif

Alexander Panchenko

Chris Biemann

Benno Stein

Henning Wachsmuth

Martin Potthast

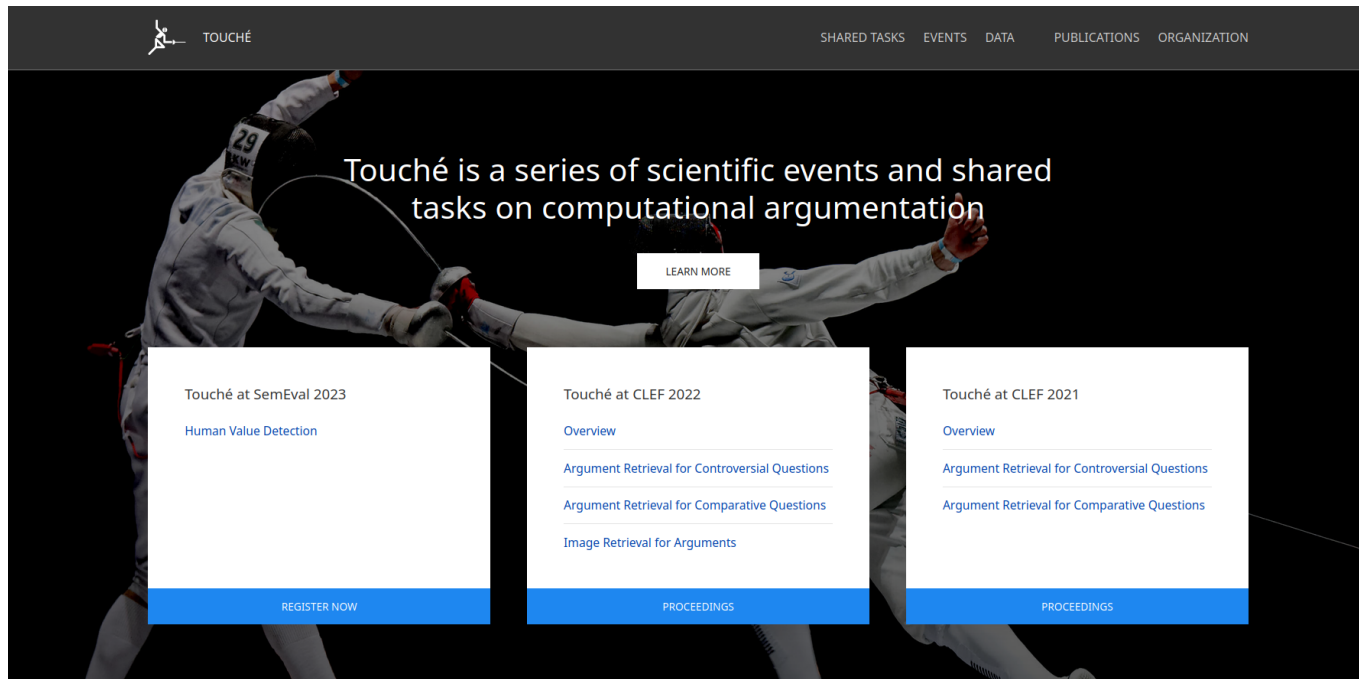
Matthias Hagen

[touche.webis.de]

Touché: Argument Retrieval

Goals

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



Task 1: Supporting conversations on controversial topics

- ❑ Scenario: Users search for arguments on controversial topics
- ❑ Task: Retrieve and rank pairs of sentences, analyze quality

Task 1: Supporting conversations on controversial topics

- ❑ Scenario: Users search for arguments on controversial topics
- ❑ Task: Retrieve and rank pairs of sentences, analyze quality

Task 2: Answering comparative questions with arguments

- ❑ Scenario: Users face personal decisions from everyday life
- ❑ Task: Retrieve and rank arguments, analyze quality, detect the stance

Task 1: Supporting conversations on controversial topics

- ❑ Scenario: Users search for arguments on controversial topics
- ❑ Task: Retrieve and rank pairs of sentences, analyze quality

Task 2: Answering comparative questions with arguments

- ❑ Scenario: Users face personal decisions from everyday life
- ❑ Task: Retrieve and rank arguments, analyze quality, detect the stance

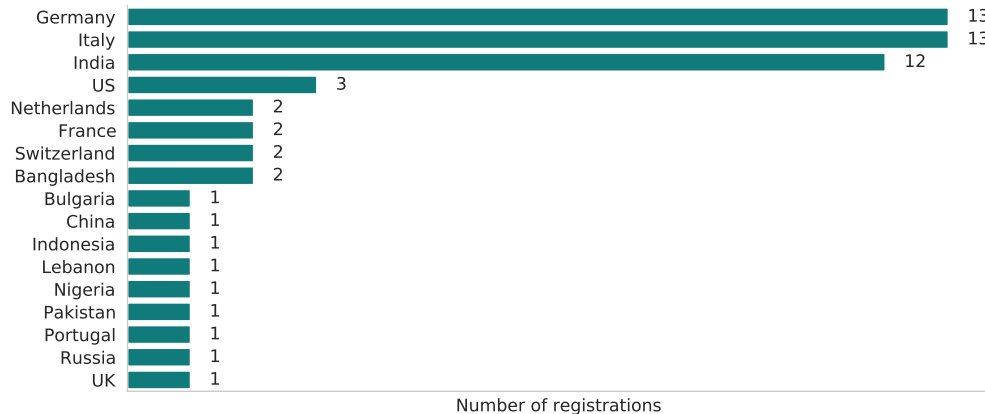
Task 3: Image retrieval for arguments

- ❑ Scenario: Users search for visual support for arguments
- ❑ Task: Retrieve images for each stance (pro/con) that support that stance

Touché: Argument Retrieval

Statistics

- ❑ Registrations: 58 teams (vs. 29 teams last year)
- ❑ Nicknames: Real or fictional fencers / swordsmen (e.g., Zorro)
- ❑ Submissions: 23 participating teams (vs. 27 last year)
- ❑ Approaches: 84 valid runs were evaluated (vs. 88 last year)
- ❑ Judgments: 15 644 manual judgments(sentences, passages, images)



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: Retrieving and analyzing an argument gist

- ❑ Scenario: Users search for arguments on socially important topics
 - ❑ Goal: Help to find the overview of different opinions / arguments
 - ❑ Task: Retrieve and rank pairs of sentences, analyze quality
 - ❑ Data: Approx. 5.7 million sentences (premises and claims).
-
- ❑ Run submissions similar to “classical” TREC tracks
 - ❑ Software submissions via TIRA [\[tira.io\]](https://tira.io)

Example topic for Task 1:

Title	<i>Should hate speech be penalized more?</i>
Description	<i>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.</i>
Narrative	<i>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 [...]</i>

- ❑ Submissions: 10 participating teams
- ❑ Nicknames: Real or fictional fencers / swordsmen (e.g., Daario Naharis)



- ❑ Approaches: 43 valid runs were evaluated
- ❑ Baseline: Graph-based sentence pair extraction [Alshomary et al. 2020]
- ❑ Topics: 50 search topics
- ❑ Evaluation: 6 930 manual relevance, coherence; and quality judgments

- ❑ Almost all teams outperformed the baseline
- ❑ Relevance and coherence evaluation indicates promising results (improvement over the baseline)
- ❑ The retrieved sentence pairs have a good quality (are argumentative)
- ❑ Finding coherent pairs of sentences is challenging
- ❑ Trends among submissions:
 - Deploying “classical” retrieval models with parameter optimization
 - Frequent focus on transformer based ML models to find coherent pairs

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, analyze quality, detect stance
- ❑ Data: Approx. 1 million passages from ClueWeb12

- ❑ Run submissions similar to “classical” TREC tracks
- ❑ Software submissions via TIRA [\[tira.io\]](https://tira.io)

- ❑ Registrations: 10 teams (46 teams: for task 2 + other tasks)
- ❑ Nicknames: Real or fictional fencers / swordsmen (e.g., Katana)



- ❑ Submissions: 7 participating teams (vs. 6 last year)
- ❑ Approaches: 25 valid runs were evaluated (vs. 19 last year)
- ❑ Baseline: BM25 / always 'no stance'
- ❑ Evaluation: 2 107 manual judgments: relevance, quality, stance (vs. 2 076 last year)

Example topic for Task 2:

Title	<i>Should I major in philosophy or psychology?</i>
Objects	<i>major in philosophy, psychology</i>
Description	<i>A soon-to-be high-school graduate finds themselves 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).</i>
Narrative	<i>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.</i>

- ❑ A few used relevance judgments from previous Touché
- ❑ Many labeled a sample of retrieved documents themselves
- ❑ Or relied on zero-shot approaches like T0++
- ❑ Using the docT5query-expanded document collection
- ❑ Main trend: transformer-based models (ColBERT, monoT5, duoT5)
- ❑ Stance: supervised classifiers (XGBoost, LSTM, RoBERTa, etc.)
- ❑ “Best” so far: retrieval / ranking pipelines that include argument mining methods and argument quality estimation

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: 24 000 web images with respective web documents

- ❑ Run submissions similar to “classical” TREC tracks
- ❑ Software submissions via TIRA [\[tira.io\]](https://tira.io)

- ❑ Submissions: 3 participating teams (+ baseline)



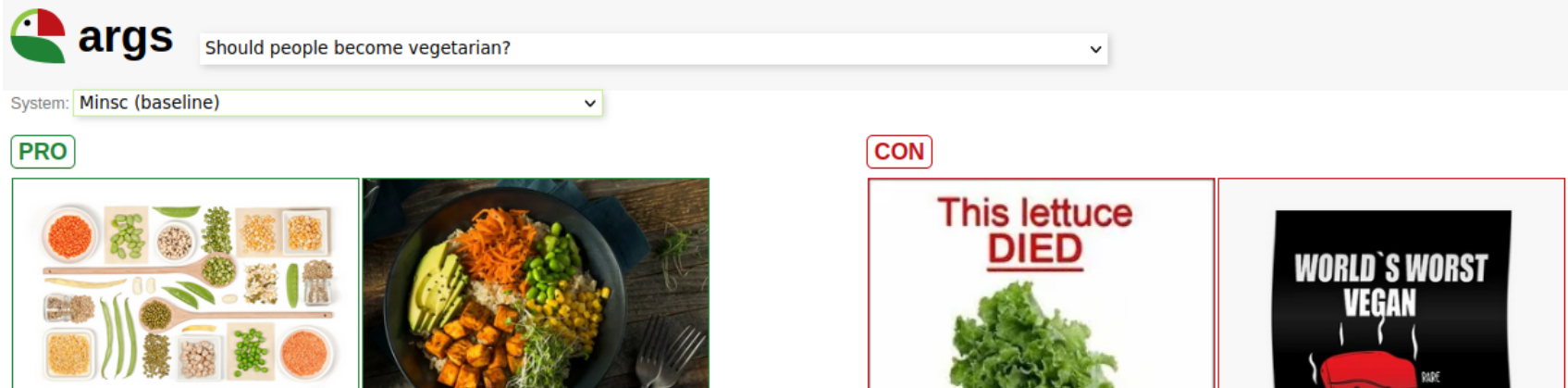
- ❑ Approaches: 12 valid runs were evaluated (+ baseline)
- ❑ Baseline: Google image search with query suffix “good” or “anti”
- ❑ Evaluation: 7 000 images-topic pairs judged manually (MTurk, MACE)

-
- Santiago Cabrera as Aramis in “The Musketeers”
 - Sean Bean as Boromir in “The Lord of the Rings”
 - Jester image by @deantna (on Pinterest).
 - Minsc (and Boo) by u/Kazuliski (on Reddit)

Touché: Argument Retrieval

Summary

- ❑ All three teams employed text extraction from images (Tesseract OCR)
- ❑ Two teams (including best-performing) used also web page text
- ❑ All teams used sentiment or emotion features: from text, colors, or faces
- ❑ “Best”: Retrieval on image + page text; stance/sentiment classifier with BERT
- ❑ Baseline beaten: Google image search with appending “good” or “anti” to topic
- ❑ Visually compare the results: <https://images.args.me>



The screenshot shows the 'args' website interface. At the top, there is a search bar with the query 'Should people become vegetarian?' and a dropdown arrow. Below the search bar, the system used is 'Minsc (baseline)'. The results are divided into two columns: 'PRO' and 'CON'. The 'PRO' column shows two images: a collection of various plant-based ingredients in small containers, and a bowl of a colorful vegetarian salad. The 'CON' column shows two images: a head of lettuce with the text 'This lettuce DIED' overlaid, and a black t-shirt with the text 'WORLD'S WORST VEGAN' and a red chili pepper graphic.

Touché: Argument Retrieval

Workshop Program



Thursday, September 8.

09:00-09:05 **Welcome**

09:05-09:45 **Session 1: Argument Retrieval for Controversial Questions**

09:05-09:15 Overview of Task 1 on Argument Retrieval for Controversial Questions (Shahbaz Syed) [\[paper\]](#)

09:15-09:25 Team Bruce Banner at Touché 2022: Argument retrieval for controversial questions (Bernardo Moreira) [\[paper\]](#)

09:25-09:35 The Pearl Retriever: Two-Stage Retrieval for Pairs of Argumentative Sentences (Sebastian Schmidt) [\[paper\]](#)

09:35-09:45 Team INTSEG on Argument Retrieval for Controversial Questions (Paria Tahan) [\[paper\]](#)

09:45-10:00 **Best of Touché 2021:** Query Expansion, Argument Mining and Document Scoring for an Efficient Question Answering System (Alaa Alhamzeh)

10:00-10:30 **Keynote:** [Ranking Arguments and Argumentative Documents: Case Studies and Challenges](#) (Andrea Galassi)

10:30-11:00 Coffee break

[\[touche.webis.de\]](https://touche.webis.de)

Touché: Argument Retrieval

Workshop Program



11:00-11:40 Session 2: Argument Retrieval for Comparative Questions

11:00-11:10 Overview of Task 2 on Argument Retrieval for Comparative Questions (Alexander Bondarenko) [\[paper\]](#)

11:10-11:20 Grimjack at Touché 2022: Axiomatic re-ranking and query reformulation (Jan Heinrich Reimer) [\[paper\]](#)

11:20-11:30 LeviRank: Limited query expansion with voting integration for document retrieval and Ranking (Ashish Rana) [\[paper\]](#)

11:30-11:35 Stacked model based argument extraction and stance detection using embedded LSTM model (Pavani Rajula) [\[paper\]](#)

11:35-11:40 Retrieving comparative arguments using deep language models (Viktoriia Chekalina) [\[paper\]](#)

11:40-12:10 Session 3: Image Retrieval for Arguments

11:40-11:50 Overview of Task 3 on Image Retrieval for Arguments (Johannes Kiesel) [\[paper\]](#)

11:50-12:00 Aramis at Touché 2022: Argument detection in pictures using machine learning (Jan Braker) [\[paper\]](#)

12:00-12:10 Boromir at Touché 2022: Combining natural language processing and machine learning techniques for image retrieval for arguments (Miriam Louise Carnot) [\[paper\]](#)

12:10-12:30 **Closing:** remarks, plenary discussion, future plans

[\[touche.webis.de\]](https://touche.webis.de)

Task 1: Argument Retrieval for Controversial Questions

- ❑ Scenario: Users search for arguments on controversial topics
- ❑ Task: Retrieve and rank relevant and high-quality documents
identify the document stance
- ❑ Data: ClueWeb22 (10 billion web documents); will be indexed in [\[ChatNoir\]](#)

Title	<i>Should teachers get tenure?</i>
Claim	<i>Teachers should get tenure</i>
Description	<i>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. [...]</i>
Narrative	<i>Highly relevant documents clearly focus on tenure for teachers in schools or universities. Relevant documents consider tenure more generally, not specifically for teachers, or [...]</i>

Task 2: Causal Retrieval

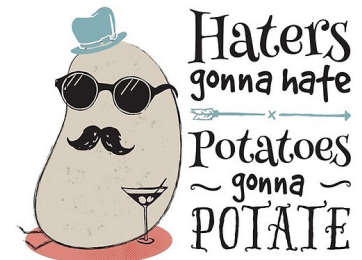
- ❑ Scenario: Support users that search for answers to causal questions
- ❑ Task: Retrieve and rank causality-related relevant documents and detect if the document supports or refutes the causal statement
- ❑ Data: ClueWeb22 (10 billion web documents); will be indexed in [\[ChatNoir\]](#)

Title	<i>Can broccoli cause constipation?</i>
Claim	<i>Broccoli causes constipation</i>
Description	<i>A young parent has a child experiencing constipation after eating some broccoli for dinner and is wondering whether broccoli could cause constipation [...]</i>
Narrative	<i>Relevant documents will discuss if broccoli and other high fiber foods can cause or ease constipation [...]</i>

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: > 30 000 web images with respective web documents

Should hate speech be banned?



Task 4: Intra-Multilingual and Multi-target Stance Classification

- ❑ Scenario: Users want to form an opinion 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 from Online Participatory Democracy Platform

Title	Topic	Proposal	Comment	Stance
Focus on Anti-Aging and Longevity research	Health	The EU has presented their green paper on aging, and correctly named the aging ...	The idea of prevention being better than a cure is nothing new or revolutionary. Rejuvenation ...	Pro
Impose an IQ or arithmetic-logic test to immigrants	Migration	We should impose an IQ test or at least several cognitive tests making sure immigrants have ...	On ne peut pas trier les migrants par un simple score sur les capacités cognitives. Certains fuient la guerre et vous ...	Against

Task 4: Intra-Multilingual and Multi-target Stance Classification

- ❑ Scenario: Users want to form an opinion 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 from Online Participatory Democracy Platform

Title	Topic	Proposal	Comment	Stance
Focus on Anti-Aging and Longevity research	Health	The EU has presented their green paper on aging, and correctly named the aging ...	The idea of prevention being better than a cure is nothing new or revolutionary. Rejuvenation ...	Pro
Impose an IQ or arithmetic-logic test to immigrants	Migration	We should impose an IQ test or at least several cognitive tests making sure immigrants have ...	On ne peut pas trier les migrants par un simple score sur les capacités cognitives. Certains fuient la guerre et vous ...	Against

— *thank you!*