Touché @ CLEF
Shared Task 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

**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
Task 1: Supporting debates on controversial topics
- Scenario: Users search for arguments on controversial topics
- Task: Retrieve “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 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]
Touché: Argument Retrieval
Argument and Argumentation

Argument:

- A conclusion (claim) supported by premises (reasons) \[\text{Walton et al. 2008}\]
- Conveys a stance on a controversial topic \[\text{Freeley and Steinberg, 2009}\]

Conclusion  \( \text{Argumentation will be a key element of conversational agents.} \)

Premise 1  \( \text{Superficial conversation ("gossip") is not enough.} \)
Premise 2  \( \text{Users want to know the "Why" to make informed decisions.} \)

Argumentation:

- Usage of arguments to achieve persuasion, agreement, …
- Decision making and opinion formation processes
Example topic for Task 1:

<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Is climate change real?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Narrative</strong></td>
<td>Relevant arguments will support the given stance that climate change is real or attack a hoax side’s argument.</td>
</tr>
</tbody>
</table>

Example **pro** argument:

One reason that I believe that **climate change is real** is the increase in global temperature and the shrinking of the Arctic ice. This is shown on this website [link].
Task 1: Supporting debates on controversial topics

- Args.me corpus [Ajjour et al. 2019]
- Argument passages from debate portals: idebate.org, debate.org, ...
- Contains both, pro and con arguments
- Download or accessible via the API of args.me search engine [args.me]
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)
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 $\Rightarrow$ relevance $\in [1, ..., 5]$ (low to high)
2. Text is not an argument $\Rightarrow$ relevance $= -2$
## Touché: Argument Retrieval
### Task 1 Strategy Overview

<table>
<thead>
<tr>
<th>Team</th>
<th>Retrieval</th>
<th>Augmentation</th>
<th>(Re)ranking Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dread Pirate Roberts</td>
<td>DirichletLM/Similarity-based</td>
<td>Language modeling</td>
<td>—</td>
</tr>
<tr>
<td>Weiss Schnee</td>
<td>DPH</td>
<td>Embeddings</td>
<td>Quality</td>
</tr>
<tr>
<td>Prince of Persia</td>
<td>Multiple models</td>
<td>Synonyms</td>
<td>Sentiment</td>
</tr>
<tr>
<td>The Three Musketeers</td>
<td>DirichletLM</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Swordsman (Baseline)</td>
<td>DirichletLM</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Thongor</td>
<td>BM25/DirichletLM</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Oscar François de Jarjayes</td>
<td>DPH/Similarity-based</td>
<td>—</td>
<td>Sentiment</td>
</tr>
<tr>
<td>Black Knight</td>
<td>TF-IDF</td>
<td>Cluster-based</td>
<td>Stance, readability</td>
</tr>
<tr>
<td>Utena Tenjou</td>
<td>BM25</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Arya Stark</td>
<td>BM25</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Don Quixote</td>
<td>Divergence from Randomness</td>
<td>Cluster-based</td>
<td>Quality + Similarity</td>
</tr>
<tr>
<td>Boromir</td>
<td>Similarity-based</td>
<td>Topic modeling</td>
<td>Author credibility</td>
</tr>
<tr>
<td>Aragorn</td>
<td>BM25</td>
<td>—</td>
<td>Premise prediction</td>
</tr>
<tr>
<td>Zorro</td>
<td>BM25</td>
<td>—</td>
<td>Quality + NER</td>
</tr>
</tbody>
</table>
Mean nDCG@5 and 95% confidence intervals.
Easiest and hardest topics.

<table>
<thead>
<tr>
<th>Topic title</th>
<th>nDCG@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is Golf a Sport?</td>
<td>0.80</td>
</tr>
<tr>
<td>Should Churches Remain Tax-Exempt?</td>
<td>0.72</td>
</tr>
<tr>
<td>Should Everyone Get a Universal Basic Income?</td>
<td>0.69</td>
</tr>
<tr>
<td>Should birth control pills be available over the counter?</td>
<td>0.66</td>
</tr>
<tr>
<td>Is Human Activity Primarily Responsible for Global Climate Change?</td>
<td>0.63</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Should Student Loan Debt Be Easier to Discharge in Bankruptcy?</td>
<td>0.20</td>
</tr>
<tr>
<td>Should Social Security Be Privatized?</td>
<td>0.20</td>
</tr>
<tr>
<td>Is a College Education Worth It?</td>
<td>0.15</td>
</tr>
<tr>
<td>Should Felons Who Have Completed Their Sentence Be Allowed to Vote?</td>
<td>0.15</td>
</tr>
<tr>
<td>Should Adults Have the Right to Carry a Concealed Handgun?</td>
<td>0.07</td>
</tr>
<tr>
<td>Average across all topics</td>
<td>0.42</td>
</tr>
</tbody>
</table>
Touché: Argument Retrieval

Approaches

Baselines:
- BM25, DPH, TF-IDF, DirichletLM

Where to start:
- Tf-idf based models are good
- Statistical language models are better
- Argument quality matters
Touché: Argument Retrieval

Approaches

Winning submissions:

- Query expansion: WordNet synonyms / antonyms → GPT-2 generation
- Document representations using Transformer (e.g., BERT)
- Re-ranking based on argument quality prediction
- Re-ranking based on sentiment (neutral sentiment)
- Pseudo-relevance feedback

[Overview of Touché 2020: Argument Retrieval]

[Touche 2020 participant papers]

[Touche 2020 slides: talks, overview]

[Touche 2020 videos on YouTube]
Touché: Argument Retrieval
Outlook Touché 2021

- 50 search topics more [touche.webis.de]
- Deeper judgment pools
- Last year’s [topics and judgments] available for training
- Evaluate argument quality: e.g., well-written, logically cogent

[Computational Argumentation Quality Assessment in Natural Language]
References:


Argument Quality Datasets:

- [https://webis.de/data.html?q=quality](https://webis.de/data.html?q=quality)

Lecture Slides / Tutorials:

- [Argument Search]
- [Applications of Computational Argumentation]
- [Argument Retrieval]
References:


Argument Quality Datasets:

- [https://webis.de/data.html?q=quality](https://webis.de/data.html?q=quality)

Lecture Slides / Tutorials:

- [Argument Search]
- [Applications of Computational Argumentation]
- [Argument Retrieval]