Touché @ CLEF 2020
1st 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 [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
Task 1: Supporting argumentative conversations
- 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]
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>Is climate change real?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</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>Narrative</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 topic for Task 2:

<table>
<thead>
<tr>
<th>Title</th>
<th>Which is better, a laptop or a desktop?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>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.</td>
</tr>
<tr>
<td>Narrative</td>
<td>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.</td>
</tr>
</tbody>
</table>
Task 1: Supporting argumentative conversations

- 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]

Task 2: Answering comparative questions with arguments

- ClueWeb12: accessible via the ChatNoir API [chatnoir.eu]
Touché: Argument Retrieval

Statistics

- Registrations: 28 teams
- Nicknames: Real or fictional fencers / swordsmen (e.g., Zorro)
- Submissions: 17 participating teams
- Approaches: 41 valid runs were evaluated
- Baselines: DirichletLM and BM25F-based ChatNoir [chatnoir.eu]
- Evaluation: 7,045 manual relevance judgments (nDCG@5)
Argument retrieval: How good are the arguments?

Task 1
- Argument relevance
- Top-5 pooling
- 5,262 unique passages
- Amazon Mechanical Turk
- nDCG@5

Task 2
- Document relevance
- Top-5 pooling
- 1,783 unique documents
- Volunteers
- nDCG@5
Mean nDCG@5, top 5 runs, args.me version 1 and 2.

<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 Mouseketeers</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>
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>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><strong>Average across all topics</strong></td>
<td><strong>0.42</strong></td>
</tr>
</tbody>
</table>
Task 2 Results

Mean nDCG@5 and 95% confidence intervals.

<table>
<thead>
<tr>
<th>Team</th>
<th>Representation</th>
<th>Query processing</th>
<th>(Re-)Ranking features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilbo Baggins</td>
<td>Bag of words</td>
<td>Named entities, comp. aspects</td>
<td>Credibility, support</td>
</tr>
<tr>
<td>Puss in Boots</td>
<td>Bag of words</td>
<td>—</td>
<td>BM25F, SpamRank</td>
</tr>
<tr>
<td>Inigo Montoya</td>
<td>Bag of words</td>
<td>Tokens &amp; logic. OR</td>
<td>Argum. units (TARGER)</td>
</tr>
<tr>
<td>Katana</td>
<td>Diff. language models</td>
<td>Diff. language models</td>
<td>Comparativeness score</td>
</tr>
<tr>
<td>Frodo Baggins</td>
<td>Bag of words</td>
<td>GloVe nearest neighbors</td>
<td>Simil. with gen. documents (GPT-2)</td>
</tr>
<tr>
<td>Zorro</td>
<td>Bag of words</td>
<td>—</td>
<td>PageRank, argumentativeness</td>
</tr>
</tbody>
</table>
Easiest and hardest topics.

<table>
<thead>
<tr>
<th>Topic title</th>
<th>nDCG@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which is better, a laptop or a desktop?</td>
<td>0.84</td>
</tr>
<tr>
<td>What is better for the environment, a real or a fake Christmas tree?</td>
<td>0.80</td>
</tr>
<tr>
<td>Which is better, Pepsi or Coke?</td>
<td>0.70</td>
</tr>
<tr>
<td>What is better: ASP or PHP?</td>
<td>0.70</td>
</tr>
<tr>
<td>Which is better, Linux or Microsoft?</td>
<td>0.70</td>
</tr>
<tr>
<td>Which city is more expensive to live in: San Francisco or New York?</td>
<td>0.18</td>
</tr>
<tr>
<td>Which smartphone has a better battery life: Xperia or iPhone?</td>
<td>0.17</td>
</tr>
<tr>
<td>What is better: to use a brush or a sponge?</td>
<td>0.16</td>
</tr>
<tr>
<td>What is the longest river in the U.S.?</td>
<td>0.10</td>
</tr>
<tr>
<td>What are the advantages and disadvantages of PHP over Python and vice versa?</td>
<td>0.10</td>
</tr>
<tr>
<td>Average across all topics</td>
<td>0.46</td>
</tr>
</tbody>
</table>
Touché: Argument Retrieval

Summary

- Platform for argument retrieval researchers
- Argument relevance / quality corpora
- Tools for submission and evaluation

- “Simple” argumentation-agnostic baselines perform well
- “Best” so far: query expansion, argument quality, comparative features
- Workshop Touché today at 15:00

Bondarenko et al. Overview of Touché 2020: Argument Retrieval
[https://webis.de/publications.html?q=stein2020v]
Touché: Argument Retrieval
Outlook Touché 2021

- 50 search topics more
- Deeper judgment pools
- This year’s topics and judgments available for training
- Evaluate argument quality dimensions:
  e.g., well-written, logically cogent, good support [Wachsmuth, et al. 2017]
- 50 search topics more
- Deeper judgment pools
- This year’s topics and judgments available for training
- Evaluate argument quality dimensions:
  e.g., well-written, logically cogent, good support [Wachsmuth, et al. 2017]

thank you!
Touché: Argument Retrieval

Related Publications