

Grimjack at Touché 2022

Axiomatic Re-ranking and Query Reformulation

Jan Heinrich Reimer Johannes Huck Alexander Bondarenko

Martin Luther University Halle-Wittenberg
<https://webis.de>

September 8, 2022



Motivation

Task 2: Comparative Argument Retrieval

Retrieve argumentative passages to assist decisions between 2 objects.

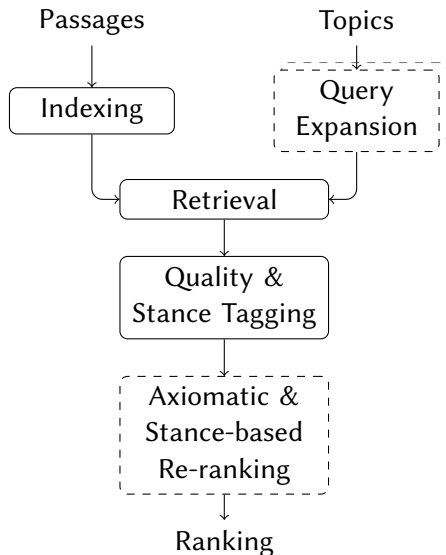
- ▶ Relevant
- ▶ High argument quality
- ▶ Sub task: Classify stance towards objects

Ideas and Questions

- ▶ Exploit argumentativeness through IR axioms [Bon+19; Bon+22a]
- ▶ Balance stances to prevent biased results [Che+21]
- ▶ Does T0++ zero-shot prompting work for argument retrieval? [San+21]

Pipeline & Tools

- ▶ Pyserini pipeline [Lin+21]
- ▶ Query expansion with synonyms: fastText, T0++
- ▶ Query reformulation: T0++ [San+21]
- ▶ Candidate retrieval: query likelihood (Dirichlet)
- ▶ Argument quality & stance: IBM Debater [Tol+19] or T0++ prompting
- ▶ Axiomatic re-ranking: `ir_axioms` [Bon+22a]



Submitted Runs

1. Query Likelihood Baseline

Dirichlet smoothing, stance from IBM Debater (threshold 0.125)

2. Argument Axioms

KWIKSORT re-ranking with 7 argumentative IR axioms

3. Stance-based Re-ranking with Argument Axioms

balance stance exposure (pro A vs. pro B)

4. All You Need is T0

expand & generate queries, estimate argument quality & stance by prompting T0++

5. Argumentative Stance-based Re-ranking with T0

expand & generate queries with T0++ and fastText, stance from IBM Debater, axiomatic and stance-based re-ranking

Results

Retrieval

- ▶ Dirichlet baseline worse than BM25 baseline [Bon+22b]
- ▶ T0++ query expansion decreases nDCG@5
- ▶ stance-based re-ranking (balancing) can slightly increase nDCG@5
- ▶ re-ranking can't compensate the bad Dirichlet retrieval performance

Table: Relevance

Run	nDCG@5
Captain Levi best [Ran+22]	0.758
<i>Puss in Boots</i> BM25 [Bon+22b]	0.469
Grimjack stance + axiom. re-rank	0.422
Grimjack axiom. re-rank	0.376
Grimjack baseline	0.376
Grimjack stance + axiom + T0	0.349
Grimjack T0	0.345

Table: Quality

Run	nDCG@5
Aldo Nadi best [Aba+22]	0.774
<i>Puss in Boots</i> BM25 [Bon+22b]	0.476
Grimjack stance + axiom. re-rank	0.403
Grimjack stance + axiom + T0	0.365
Grimjack axioms	0.363
Grimjack baseline	0.363
Grimjack T0	0.344


Results

Stance Detection

- ▶ T0++ stance prediction achieves highest macro-averaged F_1 -score
- ▶ different number of ground-truth labels per team limits comparability
- ▶ with only top-5 (i.e., all approaches have ground-truth labels up to that depth): T0++ falls behind Team Levi's approaches
- ▶ unclear how to account for sampling bias

Run	All		Top-5	
	F_1	N	F_1	N
Grimjack T0	0.313	1208	0.235	250
Captain Levi best [Ran+22]	0.301	1688	0.359	250
Grimjack axioms	0.207	1282	0.180	250
Grimjack baseline	0.207	1282	0.180	250
Grimjack stance + axiom. re-rank	0.207	1282	0.175	250
Grimjack stance + axiom + T0	0.199	1180	0.168	250
<i>Puss In Boots always NO</i> [Bon+22b]	0.158	1328	0.159	250

Conclusion

 `heinrichreimer/grimjack`

- ▶ T0 approaches rather unsuccessful for retrieval
- ▶ unclear evaluation wrt. stance classification
- ▶ balancing pro A and pro B arguments helps
- ▶ can't distinguish neutral from no stance

Future Work

- ▶ Dirichlet vs. BM25
- ▶ reproduce & evaluate stance prediction on independent test dataset

Conclusion

- 🐙 heinrichreimer/grimjack
 - ▶ T0 approaches rather unsuccessful for retrieval
 - ▶ unclear evaluation wrt. stance classification
 - ▶ balancing pro A and pro B arguments helps
 - ▶ can't distinguish neutral from no stance

Future Work

- ▶ Dirichlet vs. BM25
- ▶ reproduce & evaluate stance prediction on independent test dataset

Thank you!

Thanks to the SIGIR for waiving my registration fee.

References

- 📖 Aba, Maria et al. (Sept. 2022). “SEUPD@CLEF: Team Kueri on Argument Retrieval for Comparative Questions”. In: *Working Notes Papers of the CLEF 2022 Evaluation Labs*. Ed. by Guglielmo Faggioli et al. CEUR Workshop Proceedings.
- 📖 Bondarenko, Alexander et al. (2018). “Webis at TREC 2018: Common Core Track”. In: *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018*. Ed. by Ellen M. Voorhees et al. Vol. 500-331. NIST Special Publication. National Institute of Standards and Technology (NIST).
- 📖 Bondarenko, Alexander et al. (2019). “Webis at TREC 2019: Decision Track”. In: *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019*. Ed. by Ellen M. Voorhees et al. Vol. 1250. NIST Special Publication. National Institute of Standards and Technology (NIST).
- 📖 Bondarenko, Alexander et al. (July 2022a). “Axiomatic Retrieval Experimentation with `ir_axioms`”. In: *45th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2022)*. ACM.
- 📖 Bondarenko, Alexander et al. (Sept. 2022b). “Overview of Touché 2022: Argument Retrieval”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022)*. Lecture Notes in Computer Science. Berlin Heidelberg New York: Springer.

References (cont.)

- Cherumanal, Sachin Pathiyan et al. (2021). “Evaluating Fairness in Argument Retrieval”. In: *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*. Ed. by Gianluca Demartini et al. ACM, pp. 3363–3367.
- Lin, Jimmy et al. (2021). “Pyserini: An Easy-to-Use Python Toolkit to Support Replicable IR Research with Sparse and Dense Representations”. In: *CoRR* abs/2102.10073. arXiv: 2102.10073.
- Rana, Ashish et al. (Sept. 2022). “LeviRANK: Limited Query Expansion with Voting Integration for Document Retrieval and Ranking”. In: *Working Notes Papers of the CLEF 2022 Evaluation Labs*. Ed. by Guglielmo Faggioli et al. CEUR Workshop Proceedings.
- Sanh, Victor et al. (2021). “Multitask Prompted Training Enables Zero-Shot Task Generalization”. In: *CoRR* abs/2110.08207. arXiv: 2110.08207.
- Toledo, Assaf et al. (2019). “Automatic Argument Quality Assessment - New Datasets and Methods”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui et al. Association for Computational Linguistics, pp. 5624–5634.

Axiomatic Re-ranking with `ir_axioms` [Bon+22a]

- ▶ KWIKSORT: re-rank based on pairwise axiomatic preferences
- ▶ majority vote: only if $\geq 50\%$ axioms agree, change ranking

Name	Description
ArgUC [Bon+18]	Prefer more argumentative units.
QTArg [Bon+18]	Prefer more query terms in argumentative units.
QTPArg [Bon+18]	Prefer earlier query terms in argumentative units.
CompArg	Prefer more comparative objects in argumentative units.
CompPArg	Prefer earlier comparative objects in argumentative units.
aSLDoc [Bon+19]	Prefer passages with 12–20 words per 15–18 sentence.
ArgQ	Prefer higher argument quality.

Query Reformulation with T0++ [San+21]

- ▶ **prompt:** <text>. Extract a natural search query from this description.
- ▶ **with** <text> being the topic description (D) or narrative (N)

Topic	Field	Generated query
Train or plane? Which is the better choice?	D	Travel
	N	What are the benefits of trains over planes for inter-continental travel?
Should I major in philosophy or psychology?	D	What is the difference between philosophy and psychology?
	N	What are the benefits of a major in English or history?