

Grimjack at Touché 2022

Advanced IR, Winter Semester 2021/22

Johannes Huck Jan Heinrich Reimer

Martin Luther University Halle-Wittenberg

February 8, 2022



Task at hand

- ▶ Task 2 of Touché: Argument Retrieval
- ▶ Argument Retrieval for Comparative Questions
- ▶ Task: Retrieve relevant passages to answer comparative questions and detect their stance w.r.t the objects
- ▶ Data: > 1 million text passages

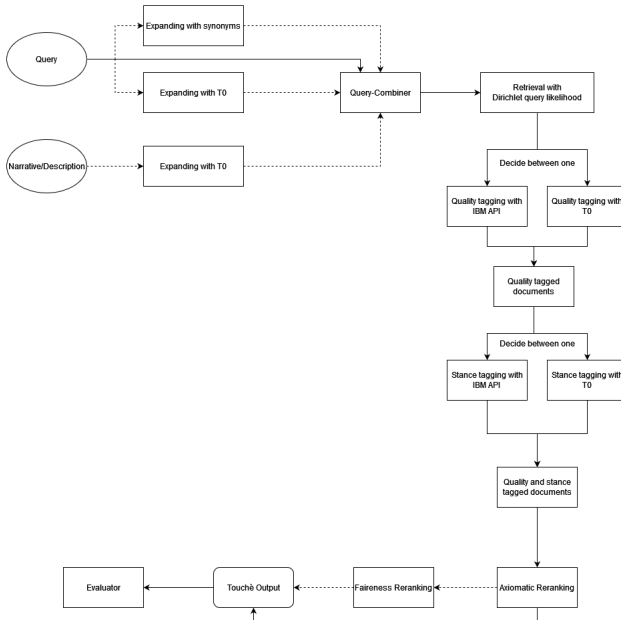


https://mobile.twitter.com/webis_de/status/1468529926026534913?cxt=HHwWgoC97fyLouEoAAAA

General approach

- ▶ Programmed in Python
 - ▶ Easy to use
 - ▶ High readability
 - ▶ Many IR libraries available
- ▶ Three modules: Search, Run file and Evaluate
- ▶ Pipeline consists of
 - ▶ Query-Expander and Query-Combiner
 - ▶ Initial Retrieval
 - ▶ Argument quality and stance tagging
 - ▶ Reranking
- ▶ Indexing and initial retrieval via pyserini [Lin+21]

Pipeline



Query-Expander and Query-Combiner

- ▶ Expanding queries with synonyms of comparative objects
- ▶ Two Different approaches
 - ▶ Based on embeddings with glove
 - ▶ Based on language model T0 [San+21]
 - ▶ We ask: What are synonyms of the word <token> ?
- ▶ With T0 also new queries from narrative and description
- ▶ We ask: <text> Extract a natural search query from this description.
- ▶ Combining all new queries with OR
- ▶ Retrieving ranked list of passages with this new query

Argument quality tagging

- ▶ Extracting arguments with TARGER [Che+19]
- ▶ For each argument we want to know the quality w.r.t. the topic
- ▶ Two different approaches
 - ▶ Based IBM Debater API [Tol+19]
 - ▶ Based on T0
 - ▶ We ask: <sentence> How would you rate the readability and consistency in this sentence? very good, good, bad, very bad
- ▶ IBM Debater API returns a score between 0 and 1
- ▶ 0 means lowest quality and 1 highest quality

Example

Arg: Cars should only provide assisted driving, not complete autonomy

Topic: We should further explore the development of autonomous vehicles

Score: 0.7256

Argument stance tagging

- ▶ Next we want to know the stance w.r.t. the topic
- ▶ Two different approaches
 - ▶ Based on IBM Debater API [Bar+17]
 - ▶ Based on T0
 - ▶ We ask: <sentence> Is this sentence pro/against <comparative_object>? yes or no
- ▶ It is also possible to expand with sentiments
- ▶ Both approaches only work for single target stance
- ▶ Calculating the multi target stance
 - ▶ Calculate the difference between objects
 - ▶ Use a threshold
 - ▶ Convert T0s output into a numerical representation

Axiomatic Reranking

- ▶ Compute preferences between documents (\triangleq axioms)
- ▶ Multiple axioms vote against the original ranking
- ▶ Rerank with KwikSort [Hag+16]

Argumentative Axioms

ArgUC Prefer more argumentative units [Bon+18]

QTArg Prefer more query terms in argumentative units [Bon+18]

QTPArg Prefer earlier query terms in argumentative units [Bon+18]

aSL Prefer sentences with 12–20 words [Bon+21]

CompArg Prefer more comparative objects in argumentative units

CompPArg Prefer earlier comparative objects in argumentative units

ArgQ Prefer higher argument quality

Fairness Reranking

- ▶ Idea: prefer subjective arguments over neutral arguments but guarantee fair exposure for each stance (pro/con)
- ▶ Alternating stance
 - ▶ Three filtered lists by stance: first, second, neutral/other
 - ▶ Alternately select from first/second list
 - ▶ Fallback to neutral list if first/second list is empty
- ▶ Balanced top- k stance
 - ▶ Count number of documents pro first or pro second in top- k ranking
 - ▶ If first - second > 1 :
Move last pro first document from top- k ranking
after the first pro second document after top- k ranking

Evaluation

- ▶ Use Qrels from Touché 2020/2021
- ▶ Problem: recent years used whole documents, we're retrieving passages
- ▶ If a document is relevant all of its passages are relevant
- ▶ Map passage IDs to document IDs, only 13 % have judgements
- ▶ Underestimates performance

Final Remarks

- ▶ Approach is very flexible
- ▶ We investigate influence of components w.r.t the retrieval score
- ▶ Stance classification may be better with RoBERTa approach
- ▶ We cannot distinguish between neutral and no stance
- ▶ We investigate how reranking influences the retrieval score
- ▶ T0 solves a lot of IR tasks
- ▶ Is it possible to only use T0 for retrieval?

Thank you!

References

- 📖 Bar-Haim, Roy et al. (2017). “Stance Classification of Context-Dependent Claims”. In: *EACL*.
- 📖 Bondarenko, Alexander et al. (Nov. 2018). “Webis at TREC 2018: Common Core Track”. In: *27th International Text Retrieval Conference (TREC 2018)*. Ed. by Ellen M. Voorhees et al. NIST Special Publication. National Institute of Standards and Technology (NIST).
- 📖 Bondarenko, Alexander et al. (2021). “Axiomatic Re-Ranking for Argument Retrieval”. In.
- 📖 Chernodub, Artem N. et al. (2019). “TARGER: Neural Argument Mining at Your Fingertips”. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*. Ed. by Marta R. Costa-jussà et al. Association for Computational Linguistics, pp. 195–200.
- 📖 Hagen, Matthias et al. (Oct. 2016). “Axiomatic Result Re-Ranking”. In: *25th ACM International Conference on Information and Knowledge Management (CIKM 2016)*. ACM, pp. 721–730.
- 📖 Lin, Jimmy J. et al. (2021). “Pyserini: An Easy-to-Use Python Toolkit to Support Replicable IR Research with Sparse and Dense Representations”. In: *ArXiv* abs/2102.10073.

References (cont.)

- 📖 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: *EMNLP*.