Stance-Aware Re-Ranking for Non-factual Comparative Queries

Jan Heinrich Reimer and **Alexander Bondarenko** and **Maik Fröbe** and **Matthias Hagen**
Friedrich-Schiller-Universität Jena

Abstract

We propose a re-ranking approach to improve the retrieval effectiveness for non-factual comparative queries like 'Which city is better, London or Paris?' based on whether the results express a stance towards the comparison objects (London vs. Paris) or not. Applied to the 26 runs submitted to the Touché 2022 task on comparative argument retrieval, our stance-aware re-ranking significantly improves the retrieval effectiveness for all runs when perfect oracle-style stance labels are available. With our most effective practical stance detector based on GPT-3.5 (F_1 of 0.49 on four stance classes), our re-ranking still improves the effectiveness for all runs but only six improvements are significant. Artificially "deteriorating" the oracle-style labels, we further find that an F_1 of 0.90 for stance detection is necessary to significantly improve the retrieval effectiveness for the best run via stance-aware re-ranking.

1 Introduction

Argument retrieval is the task of identifying and ranking text passages or documents based on their topical relevance to an argumentative query and based on their argumentativeness (i.e., the presence and quality of arguments). Current argument search engines like args.me (Wachsmuth et al., 2017) or ArgumenText (Stab et al., 2018) mainly focus on retrieving pro and con arguments on socially relevant and potentially controversial topics like 'nuclear energy' or 'plastic bottles' but they do not directly target to find pros and cons for the different options in "everyday" non-factual comparisons like 'Which city is better, London or Paris?'.

Such information needs were in the focus of the comparative argument retrieval task at the Touché 2022 lab (Bondarenko et al., 2022b). Given a query with two comparison objects (e.g., the London vs. Paris example), the goal was to retrieve results that contain arguments for or against either object. Many participants of the task improved over a BM25 baseline (Robertson et al., 1994) by using neural (re-)ranking models like ColBERT (Khattab and Zaharia, 2020) or mono- and duoT5 (Pradeep et al., 2021), and by taking estimated argument quality into account. Still, none of the participants successfully exploited stance information for the ranking (i.e., whether a result expresses a stance on the comparison objects or not) even though stance detection was also offered as a subtask at Touché.

We close this gap and, as our first contribution, suggest a simple stance-aware re-ranking approach that can be applied to the retrieval results for any comparative query: rank documents that do not express a stance on the comparison objects below any documents that do. In an evaluation on the 26 runs submitted to the Touché 2022 task, we find that our re-ranking significantly improves the retrieval effectiveness of all runs when using the task's official ground truth stance labels (i.e., assuming a "perfect" oracle-style stance detector). When instead using the participants' stance predictions, hardly any run's effectiveness can be improved as the participants' stance detectors are not effective enough ($F_1 \leq 0.31$ on the four classes 'pro first object', 'pro second object', 'both equal', and 'no stance').

As our second contribution, we thus target a better practical stance detection effectiveness and compare three approaches: (1) a fine-tuned sentiment-prompted RoBERTa model (Liu et al., 2019), (2) a zero-shot stance detector based on a pre-trained Flan-T5 model (Chung et al., 2022), and (3) GPT-3.5 (Brown et al., 2020) with few-shot prompting. Among these, the GPT-3.5-based stance detector is the most effective with an F_1 of 0.49. Using the stances detected with GPT-3.5, our stance-aware re-ranking can again improve the retrieval effectiveness of all 26 runs but only 6 of the improvements (23%) are significant. In further experiments, we artificially perturb the ground truth stance labels to analyze what stance detection effectiveness is necessary to significantly improve

Proceedings of the 10th Workshop on Argument Mining, pages 45–51

the retrieval effectiveness of the best run via stance-aware re-ranking and find that an F_1 of 0.90 is required. Our code and data are publicly available.[1]

2 Re-Ranking Scenario: Touché 2022

Our re-ranking scenario is that of the Touché 2022 shared task on comparative argument retrieval. Given one of 50 non-factual comparative queries, relevant text passages from a collection of about one million passages should be retrieved and ranked, and (optionally) their stances be detected. For our experiments, we use the 26 runs (ranked lists of results) submitted to the task, as well as the relevance + quality assessments and the stance labels that the task organizers provided (Bondarenko et al., 2022b). In the task, the retrieval effectiveness of the submitted runs was evaluated using nDCG@5 (Järvelin and Kekäläinen, 2002) for topical relevance and for argument quality, and the stance detection effectiveness was evaluated using macro-avg. F_1 on the four stance classes.

3 Stance-Aware Re-Ranking

Interestingly, none of the Touché participants successfully used stance information in their retrieval approaches. This is somewhat surprising as, intuitively, helpful retrieval results for non-factual comparative queries should express some stance towards the comparison objects (either favoring one of the objects or stating that both are equal). Our suggested re-ranking approach thus simply moves all results that do not express a stance to the end of a ranking (i.e., below any result that expresses a stance), while preserving the relative order of the documents that express a stance. Table 1 shows a respective example for a top-5 re-ranking. We have implemented this stance-aware re-ranking approach in the PyTerrier framework (Macdonald et al., 2021) as a module that expects a ranking and stances for the individual results as inputs.

4 Initial Re-Ranking Experiments

In our initial experiments, we re-rank the top-5 results of each of the 26 runs submitted to Touché based on the task's ground truth stance labels (i.e., assuming "perfect" oracle-style stance detection) or based on the participants' detected stances. Following the Touché setup, we report nDCG@5 scores for relevance and for quality and refer to the runs by their team names (e.g., Aldo Nadi or Captain L.).

[1] Code and data: github.com/webis-de/ArgMining-23

Table 1: Example of our stance-aware re-ranking. Results with no stance ('$\perp$') are moved below all results with a stance ('$O_{1/2}$': pro first / second object; '$=$': both equal) that keep their original relative ordering.

Approach	Rank				
	1	2	3	4	5
Original run	$\perp$	O_1	$\perp$	$=$	O_2
Our re-ranking	O_1	$=$	O_2	$\perp$	$\perp$

4.1 Oracle-Style Stances

To demonstrate the potential of our stance-aware re-ranking, we first re-rank based on "perfect" stances from the Touché ground truth. The results in column 'Oracle' of Table 2 show that our re-ranking then significantly improves almost all nDCG@5 scores—only the improvement of the quality score of the quality-wise best run (Aldo Nadi A) is not significant. Interestingly, the scores of the oracle-style re-ranking often are close to a run's hypothetical optimal top-5 re-ranking (column 'Opt.').

Comparing a run's rank in the original leaderboard (column '#' in 'Touché') to the potential rank if the oracle-style re-ranking was applied to only that run ('#' in 'Oracle'; 'Δ' indicates the rank change), one can, for instance, observe that the relevance-wise top-3 runs each could reach rank 1.

4.2 Touché Participants' Detected Stances

When we re-rank based on the participants' detected stances, the effectiveness of hardly any run can be improved (column 'Orig.' in Table 2); some even get worse (e.g., Captain L. B). Compared to the oracle scenario, the participants' stance detection is not effective enough ($F_1 \leq 0.31$). We thus aim to improve the practical stance detection.

5 Improving the Stance Detection

Targeting better practical stance detection, we compare three approaches: (1) a fine-tuned sentiment-prompted RoBERTa model (Liu et al., 2019), (2) a zero-shot stance detector based on a pre-trained Flan-T5 model (Chung et al., 2022), and (3) GPT-3.5 (Brown et al., 2020) with few-shot prompting. Following Touché, we use macro-avg. F_1 to compare the detection effectiveness (class distribution: 'pro first object' 19%, 'pro second object' 13%, 'both equal' 20%, 'no stance' 48%).

For the RoBERTa-based detector, we fine-tune a RoBERTa model using the sentiment-prompting

Table 2: Effectiveness (as nDCG@5) of the runs submitted to the Touché 2022 task on comparative argument retrieval (referred to by their Touché team names) and with our stance-aware re-ranking; originally submitted ('Touché'), best achievable when re-ranking the top-5 ('Opt.'), and after stance-aware re-ranking of the top-5 with: ground truth stance labels ('Oracle'), simulated labels with an F_1 of about 0.75 ('Simul.'), stance detected with GPT-3.5, Flan-T5, or RoBERTa, and with a team's original detection approach (two teams did not detect stance; grayed out). The '#' columns denote an approach's rank in the task leaderboard; for re-rankings, these columns give the rank the re-ranking would have achieved if all other runs would stay as submitted to Touché. Differences in effectiveness or rank compared to the originally submitted run are shown in the 'Δ' columns; statistically significant effectiveness differences are bold-faced (Student's t-test, $\alpha = 0.05$, Bonferroni correction for the multiple tests).

Run	Touché		Opt.	Oracle (F_1=1.00)			Simul. ($F_1\approx$0.75)			GPT-3.5 (F_1=0.49)			Flan-T5 (F_1=0.39)			RoBERTa (F_1=0.34)			Orig. ($F_1\leq$0.31)		
	Score	#	Score	Score	Δ	# Δ	Score	Δ	# Δ	Score	Δ	# Δ	Score	Δ	# Δ	Score	Δ	# Δ	Score	Δ	# Δ
Topical relevance																					
Captain L. B	0.76	1	0.81	**0.79** (+0.03)	1	(0)	0.78 (+0.02)	1	(0)	0.78 (+0.02)	1	(0)	0.76 (±0.00)	1	(0)	0.77 (+0.01)	1	(0)	0.75 (−0.01)	2	(↓1)
Captain L. A	0.76	2	0.82	**0.79** (+0.03)	1	(↑1)	0.77 (+0.01)	1	(↑1)	0.78 (+0.02)	1	(↑1)	0.76 (±0.00)	1	(↑1)	0.76 (±0.00)	1	(↑1)	0.75 (−0.01)	3	(↓1)
Captain L. D	0.75	3	0.81	**0.78** (+0.03)	1	(↑2)	0.77 (+0.02)	1	(↑2)	0.77 (+0.02)	1	(↑2)	0.76 (+0.01)	1	(↑2)	0.75 (±0.00)	3	(0)	0.75 (±0.00)	3	(0)
Captain L. E	0.73	4	0.79	**0.75** (+0.02)	4	(0)	0.74 (+0.01)	4	(0)	0.74 (+0.01)	4	(0)	0.72 (−0.01)	4	(0)	0.73 (±0.00)	4	(0)	0.73 (±0.00)	4	(0)
Captain L. C	0.72	5	0.78	**0.75** (+0.03)	4	(↑1)	0.75 (+0.03)	4	(↑1)	0.74 (+0.02)	4	(↑1)	0.72 (±0.00)	5	(0)	0.73 (+0.01)	5	(0)	0.72 (±0.00)	5	(0)
Aldo Nadi E	0.71	6	0.77	**0.74** (+0.03)	4	(↑2)	0.73 (+0.02)	4	(↑2)	0.72 (+0.01)	6	(0)	0.71 (±0.00)	6	(0)	0.71 (±0.00)	6	(0)	0.71 (±0.00)	6	(0)
Aldo Nadi A	0.70	7	0.75	**0.73** (+0.03)	4	(↑3)	0.72 (+0.02)	6	(↑1)	0.70 (±0.00)	7	(0)	0.71 (+0.01)	6	(↑1)	0.70 (±0.00)	7	(0)	0.70 (±0.00)	7	(0)
Aldo Nadi D	0.67	8	0.73	**0.70** (+0.03)	7	(↑1)	0.69 (+0.02)	8	(0)	0.68 (+0.01)	8	(0)	0.68 (+0.01)	8	(0)	0.68 (+0.01)	8	(0)	0.67 (±0.00)	8	(0)
Aldo Nadi C	0.64	9	0.70	**0.67** (+0.03)	8	(↑1)	**0.67** (+0.03)	9	(0)	0.65 (+0.01)	9	(0)	0.64 (±0.00)	9	(0)	0.63 (−0.01)	9	(0)	0.64 (±0.00)	9	(0)
Katana A	0.62	10	0.69	**0.65** (+0.03)	9	(↑1)	**0.64** (+0.02)	9	(↑1)	**0.65** (+0.03)	9	(↑1)	0.63 (+0.01)	10	(0)	0.62 (±0.00)	10	(0)	0.62 (±0.00)	10	(0)
Katana C	0.60	11	0.67	**0.64** (+0.04)	9	(↑2)	**0.63** (+0.03)	10	(↑1)	**0.62** (+0.02)	10	(↑1)	0.61 (+0.01)	11	(0)	0.60 (±0.00)	11	(0)	0.60 (±0.00)	11	(0)
Captain T. A	0.57	12	0.64	**0.61** (+0.04)	11	(↑1)	**0.61** (+0.04)	11	(↑1)	0.59 (+0.02)	12	(0)	0.58 (+0.01)	12	(0)	0.58 (+0.01)	12	(0)	0.57 (±0.00)	12	(0)
Captain T. B	0.57	13	0.64	**0.61** (+0.04)	11	(↑2)	0.59 (+0.02)	12	(↑1)	0.58 (+0.01)	12	(↑1)	0.57 (±0.00)	13	(0)	0.57 (±0.00)	13	(0)	0.57 (±0.00)	13	(0)
Captain T. C	0.56	14	0.62	**0.59** (+0.03)	12	(↑2)	0.58 (+0.02)	12	(↑2)	0.57 (+0.01)	13	(↑1)	0.57 (+0.01)	14	(0)	0.56 (±0.00)	15	(↑1)	0.56 (±0.00)	14	(0)
Katana B	0.56	15	0.63	**0.60** (+0.04)	11	(↑4)	**0.59** (+0.03)	12	(↑3)	0.58 (+0.02)	12	(↑3)	0.56 (±0.00)	16	(↓1)	0.56 (±0.00)	15	(0)	0.56 (±0.00)	15	(0)
Captain T. E	0.56	16	0.64	**0.60** (+0.04)	12	(↑4)	**0.59** (+0.03)	12	(↑4)	0.58 (+0.02)	12	(↑4)	0.57 (+0.01)	14	(↑2)	0.56 (±0.00)	16	(0)	0.56 (±0.00)	16	(0)
Aldo Nadi B	0.55	17	0.61	**0.58** (+0.03)	12	(↑5)	**0.57** (+0.02)	14	(↑3)	0.56 (+0.01)	16	(↑1)	0.55 (±0.00)	17	(0)	0.55 (±0.00)	17	(0)	0.55 (±0.00)	17	(0)
Captain T. D	0.54	18	0.61	**0.58** (+0.04)	12	(↑6)	0.56 (+0.02)	16	(↑2)	0.56 (+0.02)	16	(↑2)	0.54 (±0.00)	18	(0)	0.54 (±0.00)	18	(0)	0.54 (±0.00)	18	(0)
Olivier A. A	0.48	19	0.57	**0.55** (+0.07)	17	(↑2)	**0.53** (+0.05)	19	(0)	**0.52** (+0.04)	19	(0)	0.50 (+0.02)	19	(0)	**0.51** (+0.03)	19	(0)	0.49 (+0.01)	19	(0)
Puss in B. A	0.47	20	0.55	**0.52** (+0.05)	19	(↑1)	**0.50** (+0.03)	19	(↑1)	0.49 (+0.02)	19	(↑1)	0.47 (±0.00)	20	(0)	0.47 (±0.00)	20	(0)	0.47 (±0.00)	20	(0)
Grimjack E	0.42	21	0.48	**0.46** (+0.04)	21	(0)	**0.45** (+0.03)	21	(0)	0.44 (+0.02)	21	(0)	0.42 (±0.00)	21	(0)	0.43 (+0.01)	21	(0)	0.42 (±0.00)	21	(0)
Grimjack C	0.38	22	0.46	**0.44** (+0.06)	21	(↑1)	**0.41** (+0.03)	22	(0)	**0.41** (+0.03)	22	(0)	0.39 (+0.01)	22	(0)	0.40 (+0.02)	22	(0)	0.38 (±0.00)	22	(0)
Grimjack B	0.38	23	0.46	**0.44** (+0.06)	21	(↑2)	**0.42** (+0.04)	22	(↑1)	**0.41** (+0.03)	22	(↑1)	0.39 (+0.01)	22	(↑1)	0.40 (+0.02)	22	(↑1)	0.38 (±0.00)	23	(0)
Grimjack D	0.35	24	0.41	**0.38** (+0.03)	22	(↑2)	0.36 (+0.01)	24	(0)	0.36 (+0.01)	24	(0)	0.35 (±0.00)	24	(0)	0.35 (±0.00)	24	(0)	0.35 (±0.00)	24	(0)
Grimjack A	0.34	25	0.43	**0.40** (+0.06)	22	(↑3)	**0.38** (+0.04)	22	(↑3)	**0.38** (+0.04)	22	(↑3)	0.37 (+0.03)	24	(↑1)	0.37 (+0.03)	24	(↑1)	0.34 (±0.00)	25	(0)
Asuna A	0.26	26	0.34	**0.32** (+0.06)	26	(0)	**0.30** (+0.04)	26	(0)	0.28 (+0.02)	26	(0)	0.27 (+0.01)	26	(0)	0.27 (+0.01)	26	(0)	0.26 (±0.00)	26	(0)
Argument quality																					
Aldo Nadi A	0.77	1	0.83	0.80 (+0.03)	1	(0)	0.78 (+0.01)	1	(0)	0.78 (+0.01)	1	(0)	0.78 (+0.01)	1	(0)	0.78 (+0.01)	1	(0)	0.77 (±0.00)	1	(0)
Aldo Nadi C	0.76	2	0.81	**0.79** (+0.03)	1	(↑1)	0.78 (+0.02)	1	(↑1)	0.77 (+0.01)	2	(0)	0.76 (±0.00)	2	(0)	0.76 (±0.00)	2	(0)	0.76 (±0.00)	2	(0)
Aldo Nadi E	0.75	3	0.80	**0.77** (+0.02)	2	(↑1)	**0.77** (+0.02)	2	(↑1)	0.75 (±0.00)	3	(0)	0.74 (−0.01)	3	(0)	0.75 (±0.00)	3	(0)	0.75 (±0.00)	3	(0)
Captain L. B	0.74	4	0.82	**0.77** (+0.03)	2	(↑2)	**0.77** (+0.03)	2	(↑2)	**0.77** (+0.03)	2	(↑2)	0.76 (+0.02)	3	(↑1)	0.75 (+0.01)	3	(↑1)	0.74 (±0.00)	5	(↓1)
Captain L. A	0.74	5	0.82	**0.77** (+0.03)	2	(↑3)	**0.77** (+0.03)	2	(↑3)	**0.77** (+0.03)	2	(↑3)	**0.76** (+0.02)	3	(↑2)	0.75 (+0.01)	3	(↑2)	0.74 (±0.00)	5	(0)
Captain L. D	0.73	6	0.79	**0.75** (+0.02)	4	(↑2)	0.74 (+0.01)	4	(↑2)	**0.75** (+0.02)	3	(↑3)	**0.75** (+0.02)	4	(↑2)	0.73 (±0.00)	6	(0)	0.73 (±0.00)	6	(0)
Captain L. E	0.71	7	0.76	**0.73** (+0.02)	7	(0)	0.72 (+0.01)	7	(0)	0.72 (+0.01)	7	(0)	0.72 (+0.01)	7	(0)	0.71 (±0.00)	7	(0)	0.71 (±0.00)	7	(0)
Captain L. C	0.70	8	0.77	**0.72** (+0.02)	7	(↑1)	0.71 (+0.01)	7	(↑1)	**0.72** (+0.02)	7	(↑1)	0.70 (±0.00)	8	(0)	0.70 (±0.00)	8	(0)	0.70 (±0.00)	8	(0)
Aldo Nadi D	0.66	9	0.73	**0.69** (+0.03)	9	(0)	0.68 (+0.02)	9	(0)	0.68 (+0.02)	9	(0)	0.67 (+0.01)	9	(0)	0.68 (+0.02)	9	(0)	0.66 (±0.00)	9	(0)
Katana C	0.64	10	0.71	**0.67** (+0.03)	9	(↑1)	**0.68** (+0.04)	9	(↑1)	0.66 (+0.02)	9	(↑1)	0.66 (+0.02)	10	(0)	0.65 (+0.01)	10	(0)	0.64 (±0.00)	10	(0)
Katana A	0.64	11	0.72	**0.68** (+0.04)	9	(↑2)	**0.67** (+0.03)	9	(↑2)	**0.67** (+0.03)	9	(↑2)	0.66 (+0.02)	10	(↑1)	0.65 (+0.01)	10	(↑1)	0.64 (±0.00)	11	(0)
Katana B	0.64	12	0.70	**0.67** (+0.03)	9	(↑3)	**0.66** (+0.02)	9	(↑3)	0.66 (+0.02)	10	(↑2)	0.64 (±0.00)	12	(0)	0.64 (±0.00)	11	(↑1)	0.64 (±0.00)	12	(0)
Captain T. E	0.60	13	0.67	**0.63** (+0.03)	13	(0)	**0.62** (+0.02)	13	(0)	**0.62** (+0.02)	13	(0)	0.61 (+0.01)	13	(0)	0.60 (±0.00)	13	(0)	0.60 (±0.00)	13	(0)
Captain T. B	0.59	14	0.65	**0.62** (+0.03)	13	(↑1)	0.61 (+0.02)	13	(↑1)	0.61 (+0.02)	13	(↑1)	0.60 (+0.01)	13	(↑1)	0.59 (±0.00)	14	(0)	0.59 (±0.00)	14	(0)
Captain T. A	0.59	15	0.65	**0.62** (+0.03)	13	(↑2)	**0.62** (+0.03)	13	(↑2)	0.61 (+0.02)	13	(↑2)	0.60 (+0.01)	14	(↑1)	0.59 (±0.00)	15	(0)	0.59 (±0.00)	15	(0)
Captain T. C	0.58	16	0.64	**0.61** (+0.03)	13	(↑3)	0.60 (+0.02)	13	(↑3)	0.60 (+0.02)	13	(↑3)	0.59 (+0.01)	15	(↑1)	0.58 (±0.00)	16	(0)	0.58 (±0.00)	16	(0)
Olivier A. A	0.57	17	0.65	**0.62** (+0.05)	13	(↑4)	**0.61** (+0.04)	13	(↑4)	**0.61** (+0.04)	13	(↑4)	0.59 (+0.02)	16	(↑1)	0.59 (+0.02)	15	(↑2)	0.58 (+0.01)	17	(0)
Aldo Nadi B	0.57	18	0.63	**0.60** (+0.03)	13	(↑5)	0.58 (+0.01)	16	(↑2)	0.58 (+0.01)	17	(↑1)	0.58 (+0.01)	17	(↑1)	0.58 (+0.01)	16	(↑2)	0.57 (±0.00)	18	(0)
Captain T. D	0.57	19	0.65	**0.61** (+0.04)	13	(↑6)	**0.60** (+0.03)	13	(↑6)	**0.59** (+0.02)	15	(↑4)	0.58 (+0.01)	17	(↑2)	0.57 (±0.00)	17	(↑2)	0.57 (±0.00)	19	(0)
Puss in B. A	0.48	20	0.54	**0.51** (+0.02)	20	(0)	0.49 (+0.01)	20	(0)	0.49 (+0.01)	20	(0)	0.49 (+0.01)	20	(0)	0.50 (+0.02)	20	(0)	0.48 (±0.00)	20	(0)
Grimjack E	0.40	21	0.47	**0.44** (+0.04)	21	(0)	0.42 (+0.02)	21	(0)	**0.42** (+0.02)	21	(0)	0.41 (+0.01)	21	(0)	**0.43** (+0.03)	21	(0)	0.40 (±0.00)	21	(0)
Grimjack D	0.37	22	0.42	**0.39** (+0.02)	22	(0)	0.38 (+0.01)	22	(0)	0.38 (+0.01)	22	(0)	0.38 (+0.01)	22	(0)	0.37 (±0.00)	22	(0)	0.37 (±0.00)	22	(0)
Grimjack C	0.36	23	0.44	**0.41** (+0.05)	21	(↑2)	**0.40** (+0.04)	22	(↑1)	**0.39** (+0.03)	22	(↑1)	**0.39** (+0.03)	22	(↑1)	**0.40** (+0.04)	22	(↑1)	0.36 (±0.00)	23	(0)
Grimjack B	0.36	24	0.44	**0.41** (+0.05)	21	(↑3)	**0.40** (+0.04)	22	(↑2)	**0.39** (+0.03)	22	(↑2)	**0.39** (+0.03)	22	(↑2)	**0.40** (+0.04)	22	(↑2)	0.36 (±0.00)	24	(0)
Grimjack A	0.34	25	0.42	**0.39** (+0.05)	22	(↑3)	**0.38** (+0.04)	22	(↑3)	**0.38** (+0.04)	22	(↑3)	**0.38** (+0.04)	22	(↑3)	0.36 (+0.02)	25	(0)	0.34 (±0.00)	25	(0)
Asuna A	0.33	26	0.42	**0.38** (+0.05)	22	(↑4)	**0.37** (+0.04)	22	(↑4)	0.35 (+0.02)	25	(↑1)	0.35 (+0.02)	25	(↑1)	0.35 (+0.02)	25	(↑1)	0.33 (±0.00)	26	(0)

idea and data of Bondarenko et al. (2022a). For the Flan-T5-based detector, we let Flan-T5 predict stances for each sentence in a passage (to avoid truncation at 512 tokens) using 4 zero-shot prompts (one per comparison object and pro/con) and then aggregate the stances (prompts and aggregation: Appendix A). Finally, for the GPT-3.5-based detector, we few-shot prompt GPT-3.5[2] with four examples (one per stance) that consist of a comparative query, two comparison objects, a text passage, and a stance + short explanation (prompt: Appendix B).

Using GPT-3.5-based stances (with an F_1 of 0.49, it is our most effective practical stance detector), our re-ranking approach can improve all nDCG@5 scores, but only 6 of the relevance-wise (23%) and 12 of the quality-wise improvements (46%) are significant (column 'GPT-3.5' in Table 2). The relevance-wise top-3 runs each would reach rank 1 after re-ranking, while the quality-wise best run cannot be "dethroned". The Flan-T5-based stances (F_1 of 0.39) also suffice to move the relevance-wise top-3 runs to rank 1 (column 'Flan-T5' in Table 2), while for the RoBERTa-based stances (F_1 of 0.34) only the relevance-wise second run could make it to the top (column 'RoBERTa' in Table 2).

6 Testing Limits with Simulated Stances

To analyze the (potential) impact of stance detectors that are more effective than our currently most effective practical approach (GPT-3.5-based; F_1 of 0.49), we gradually artificially deteriorate the ground truth stances as follows. From the passages with ground truth stance labels, we iteratively randomly select one without replacement and sample a stance label from the ground truth label distribution (O_1: 19%, O_2: 13%, $=$: 20%, $\perp$: 48%; a sampled label for a passage could be the same as in the ground truth) until the F_1 of the perturbed ground truth falls below a desired stopping threshold. Using this process, we simulate "stance detectors" with F_1 scores of 0.95, 0.9, 0.85, ..., 0.25, 0.2. For each threshold, we run the process ten times with different random seeds to obtain ten perturbed ground truths per target F_1 score. The ten perturbed ground truths are then each used to re-rank a run's retrieval results and the resulting ten nDCG@5 scores are averaged—to somewhat smooth out possible randomization effects.

[2]Accessed via its API on January 19, 2023; default parameters (model: text-davinci-003, temp.: 0.0, max tokens: 64, top-p: 1.0, frequency penalty: 0.0, presence penalty: 0.0).

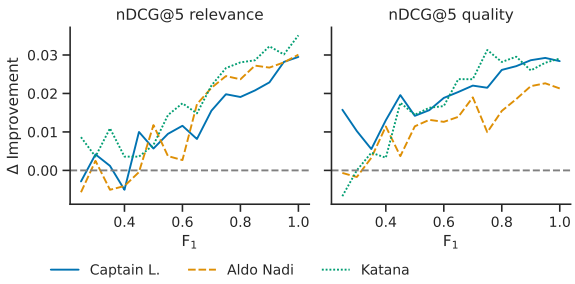

Figure 1: Effectiveness improvements of the top-3 teams' best runs when re-ranked with stance labels of the simulated target F_1 scores. For each target F_1 score, the improvement is averaged over the re-rankings with the ten simulated ground truths of that F_1 score. The 16 actually discrete improvement values per run are connected as line plots for a better visual discriminability.

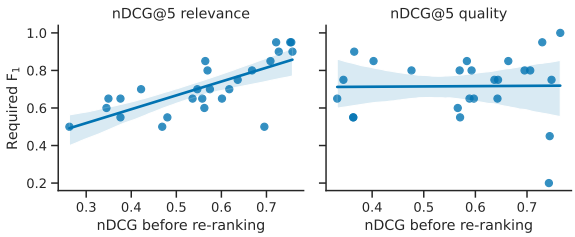

Figure 2: Minimum simulated stance detection F_1 scores for which the stance-aware re-ranking significantly improves an original Touché run (given by their nDCG@5 score before re-ranking). For each F_1 score, the improvement is averaged over the re-rankings with the ten simulated ground truths of that F_1 score.

As an example, column 'Simul.' in Table 2 shows the effects of a simulated stance detection with an F_1 of 0.75—midst of the perfect oracle and our currently best practical detector (GPT-3.5-based). One can observe that the $F_1 = 0.75$-based re-ranking improves the effectiveness scores of all runs, as is the case with GPT-3.5-based stances, and that a few more of the differences are significant—none of the relevance-wise top-8, though.

To clarify whether there is a relationship between stance detection F_1 and retrieval effectiveness improvement, Figure 1 shows the effectiveness scores when re-ranking the top-3 teams' best runs with the perturbed ground truths of different target stance detection F_1 scores. One can clearly observe that an increasing stance detection F_1 yields increased retrieval effectiveness improvements (relevance and quality; trends similar for other runs and teams).

The minimally needed stance detection F_1 so that the respective stance-aware re-ranking significantly improves an original run is shown in Fig-

ure 2 (runs given by their initial nDCG@5 scores). As for the relevance-wise improvements, one can observe a clear trend that runs with a better initial effectiveness require better stance detection to yield significant improvements. For the relevance-wise best runs, even almost perfect stance detection F_1 scores of 0.9 or 0.95 are needed to yield significant relevance-wise improvements.

As for the quality-wise improvements, no clear trend is observable. Two "outliers" of runs with a good initial effectiveness only require some rather low stance detection F_1 for significant improvements, but many runs with quite different initial quality-wise effectiveness require pretty high F_1 scores. Interestingly, the quality-wise best run Aldo Nadi A can never be significantly improved, even with perfect oracle-style stance labels.

7 Conclusion

We have proposed a simple stance-aware re-ranking approach for non-factual comparative queries that just moves results that do not express a stance on the comparison objects below any results that do. For all 26 runs submitted to the Touché 2022 task on comparative argument retrieval, our re-ranking can significantly improve the retrieval effectiveness when using the official Touché stance labels (i.e., assuming a "perfect" oracle-style stance detector). Then again, re-ranking based on the stances detected by the task participants ($F_1 \leq 0.31$) hardly improves any run. We thus experimented with other stance detectors to achieve better practical stance effectiveness. Using our most effective detector (GPT-3.5-based; F_1 of 0.49), the re-ranking can again improve the retrieval effectiveness for all 26 runs but only 6 of the relevance-wise and 12 of the quality-wise improvements are significant. In a final experiment with controlled perturbation of the ground truth stances, we found that better stance detection effectiveness tends to yield better re-ranking effectiveness and that a stance detection F_1 of 0.90 is necessary to significantly improve the relevance-wise most effective run.

Substantially improving the practical stance detection effectiveness thus is an interesting direction for future work that could also be the basis for a diversified result presentation: splitting the results into three separate lists for 'pro first object', 'pro second object', and 'both equal'. Besides, our re-ranking approach does not yet consider any potential confidence scores of a stance detection model and also no potentially predicted stance "magnitude". Developing stance detectors that assign a confidence or stance magnitude might actually be helpful to further improve the stance-aware re-ranking (e.g., to rank results with high-confidence stances above the ones with low confidence).

Acknowledgments

This work has been partially supported by the DFG (German Research Foundation) through the project "ACQuA 2.0: Answering Comparative Questions with Arguments" (project 376430233) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999).

References

Alexander Bondarenko, Yamen Ajjour, Valentin Dittmar, Niklas Homann, Pavel Braslavski, and Matthias Hagen. 2022a. Towards understanding and answering comparative questions. In *Proceedings of WSDM 2022*, pages 66–74.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022b. Overview of Touché 2022: Argument retrieval. In *Proceedings of CLEF 2022*, pages 311–336.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS 2020*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. arXiv 2210.11416.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of SIGIR 2020*, pages 39–48.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv 1907.11692.

Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. 2021. PyTerrier: Declarative experimentation in Python from BM25 to dense retrieval. In *Proceedings of CIKM 2021*, pages 4526–4533.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. arXiv 2101.05667.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of TREC 1994*, pages 109–126.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for arguments in heterogeneous sources. In *Proceedings of NAACL-HLT 2018*, pages 21–25.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the Web. In *Proceedings of ArgMining@EMNLP 2017*, pages 49–59.

A Flan-T5 Prompts and Aggregation

Positive prompt:

<sentence>

Is this sentence pro <object O_x>? yes or no

Negative prompt:

<sentence>

Is this sentence against <object O_x>? yes or no

On these prompts, Flan-T5 usually generates some longer answer text. We derive object stance scores st_{O_x} for the objects O_1 and O_2 based on whether the outputs contain some "trigger" terms like yes, no, pro, or con (left table below). Afterwards, we map the object stance scores to sentence stance scores st_s (right table below).

A passage's stance is the average of all contained sentences' stances (ignoring sentences without stance) mapped to: > 0 'pro first obj.', < 0 'pro second obj.', 0 'both equal', $\perp$ 'no stance'.

Flan-T5 output contains		**Stance**
Pos. prompt	Neg. prompt	st_{O_x}
$(\text{yes} \vee \text{pro}) \wedge \neg\text{no}$	$(\text{yes} \vee \text{con}) \wedge \neg\text{no}$	0
$(\text{yes} \vee \text{pro}) \wedge \neg\text{no}$	$(\neg\text{yes} \wedge \neg\text{con}) \vee \text{no}$	1
$(\neg\text{yes} \wedge \neg\text{pro}) \vee \text{no}$	$(\text{yes} \vee \text{con}) \wedge \neg\text{no}$	0
$(\neg\text{yes} \wedge \neg\text{pro}) \vee \text{no}$	$(\neg\text{yes} \wedge \neg\text{con}) \vee \text{no}$	$\perp$

Sentence Stance		
st_{O_1}	st_{O_2}	st_s
$\perp$	$\perp$	$\perp$
a	a	0
a	$\perp$	a
$\perp$	a	$-a$
a	b	$a - b$

B GPT-3.5 Prompt (Few-Shot)

You will be shown a text passage that compares two objects. Decide if the passage provides arguments pro first object, pro second object, both equal, or no stance is given. First, we start with examples and definitions. Please read them carefully.

Question: Apple vs Microsoft: Which is better?
Answer passage: I switched from PC to Mac about 2 years ago, after becoming familiar with Macs using my sister's computer. I will NEVER go back to PCs. I also like that Macs are simplified for basic things such as photos, music, internet, and e-mail. Truthfully, the only programs I have issues with are Microsoft applications like Word and IE. I think Apple's superiority comes from the fact that Macs are inherently more stable systems.
First object: Apple, second object: Microsoft.
Explanation: The answer provides a strong pro argument (opinion) for MAC (which is referred to as Apple). Note, that the text passage may not use the same object names as the question, e.g., it can contain synonyms or abbreviations or just mention only one object. Stance: pro first object.

Question: Is it better to dual-boot or run a VM?
Answer passage: Dual boot is a waste of time. I describe it to people as the 5-minute alt-tab. [...] I avoid dual boot like the plague. VM all the way. Or, just use a single OS that does what you want. Windows with Cygwin provides a lot of the Unixy stuff that people need.
First object: to dual-boot, second object: run a VM.
Explanation: The answer provides a strong opinion that a VM is better than a dual-boot. Note, that the text passage may not use the same object names as the question, e.g., it can contain synonyms or abbreviations or just mention only one object. Stance: pro second object.

Question: Who would win in a battle, a squirrel or a bird?
Answer passage: First of all, it depends on the bird's size. The bird has the initial advantage of flying away. [...] But if it is small, it would fly

away. And you know, the winner never runs away from the battlefield.
First object: squirrel, second object: bird.
Explanation: The answer suggests that under some condition a bird would win, but without the condition a squirrel would. This means both could win a fight, and they are equal. Stance: both equal.

Question: Which to choose a pie or a tart?
Answer passage: Generally speaking, a pie refers to a pastry covered with a lid, like a typical apple pie. A tart is open-topped, like a quiche, or a French tartes aux pommes. [. . .] Regional variations also apply.
First object: pie, second object: tart.
Explanation: The answer does not provide any pro or con arguments or opinions. The answer simply describes what a pie and a tart are. According to the definition of stance (see above), there is no stance in the passage. Stance: no stance.

Also, select "no stance" if the text passage does not contain arguments / opinions toward the objects (that is neither the first nor second object nor their synonyms are in the text).

Now, I have a question comparing first object: <first object> and second object: <second object>:
Question: <question>
Identify whether the following text is "pro first object", "pro second object", "both equal", or "no stance". Please, answer only with "pro first object", "pro second object", "both equal", or "no stance":
Answer passage: <passage>
Stance: