Universität Leipzig
Faculty of Mathematics and Computer Science
Institute for Computer Science

Evaluation Framework for Argument Retrieval in a Crowdsourced Setting

A Multi-Aspect Approach to Argument Quality Assessment Based on the Bradley-Terry Model

Bachelor's Thesis

Leipzig, April 25, 2019

Submitted by:
Lukas Gienapp
Digital Humanities, B.Sc.

Supervised by:
Jun.-Prof. Dr. Martin Potthast

Abstract

This work develops an evaluation framework to meaningfully assess the performance of different retrieval models for the task of argument retrieval. A new annotation procedure for argument quality is developed, reducing the annotation effort by over 90% while achieving an annotation accuracy on par with previous relative rating methods and better than previous absolute rating methods. This novel approach is then applied to evaluate the retrieval results of three different retrieval models in the domain of argument search, comparing the new evaluation procedure based on argument quality with a classic TREC-like approach based on item relevance.

Contents

Chapter 1

Introduction

Argument search is a novel field in information retrieval which is concerned with identifying arguments in sources of text and ranking them relative to a proposed search issue. Contrary to classic information retrieval, which is mostly aimed at providing the best objective answer to a search query, argument retrieval specializes in controversial topics, where no unique best answer is desired, but rather a spectrum of results capturing the controversial and opinionated nature of the information need. It aims to deliver preformulated argumentative text and give insight about a matter from different perspectives. Also, arguments are not necessarily constructed around factual knowledge, which leads to a multitude of possible results with diverse characteristics for a given search query. Although vast resources that argumentative text could be sourced from are available in different forms, like specific knowledge corpora, ontologies, argument corpora or just the web as a whole, retrieving the best arguments on an issue remains a complex task.

Several approaches that try to resolve this problem have emerged recently. Early work in the field of argument retrieval was restricted to domain-specific search, such as mining arguments from legal texts [58]. Subsequently, systems that provide domain-independent debate support were developed, most notably IBM's *Project Debater* [41].[1] However, it is designed to take on humans in real debates and is not (yet) available to provide support in everyday use. Two alternative solutions have emerged recently to make the task of argument retrieval publicly accessible through search engines specializing in argument search, namely `args.me`[2] [53] and `ArgumentText`[3] [49]. They aim at enabling users to search for arguments unrestrained by search domain in a broad variety of topics.

[1] `https://www.research.ibm.com/artificial-intelligence/project-debater/`, unless otherwise noted, all URLs in this work have been last accessed on April 25, 2019

[2] `https://www.args.me`

[3] `https://www.argumentext.de`

However, none of these search engines have yet undergone a comprehensive performance evaluation, and the minimal evaluation work that has been done is constructed in diverse ways and does not follow a systematic evaluation methodology. A general methodological framework for argument retrieval evaluation is therefore warranted.

Since the two publicly available search engines treat argument search as an ad-hoc retrieval task, the classic TREC-style evaluation procedure could be employed. Here, the performance of engines is usually estimated through relevance judgments relative to a search topic. For argument search however, the ranking task for a search engine should not only take topical relevance into account, but also incorporate argument quality, as an argument of higher inherent argumentative quality is more desirable to the user than one of lower quality, even though they may have the same level of relevance. Modifying the TREC evaluation method to take both relevance and quality into account in a mixed approach creates a novel problem for the evaluation task: a reliable way of measuring the argumentative quality of text has to be found.

Usually, the assessment of argument quality is done through expert annotations, as crowd-based approaches to the issue show a lack of annotation quality. However, employing an assessment based on expert annotations on the scale needed for a comprehensive evaluation is problematic: reliable, objective expert annotations are difficult to produce in the desired quantities and are only available at a high cost. Using laymen annotations, which are available in abundance for a comparatively cheap price through crowdsourcing is problematic, too: they often do not match the quality criteria needed for an empirical evaluation of argument search engines.

This work therefore focuses on defining an annotation procedure that produces high-quality annotations of argument quality from crowdsourced data, and how to incorporate it into the classic TREC information retrieval evaluation process. The derived evaluation framework was then applied to gain information about the performance of different retrieval models for argument search to showcase the capabilities of the framework, and secondly, to provide a deeper insight into which existing retrieval models may be suited best for the task of argument retrieval, as no specialized retrieval models for argument search have been developed as of now. This is the first systematic evaluation of retrieval models regarding their performance for argument search.

Existing work in the field of information retrieval evaluation and argument quality assessment is reviewed in Chapter 2, developing prototypical guidelines for the framework. A pilot study is analyzed in Chapter 3, supplying additional information to refine the evaluation process. Based on the identified challenges, a finalized method for argument retrieval evaluation is described in Chapter 4. This framework is then applied to evaluate different retrieval models, with the

data acquisition process detailed in Chapter 5, and the evaluation based on this data given in Chapter 6. Chapter 7 provides a conclusion on the work done.

Chapter 2

Theoretical Background

This section provides a first outline of the proposed evaluation framework. Existing work in the field of argument retrieval is surveyed for suited evaluation procedures (Section 2.1) and extended upon by describing in detail the general IR-evaluation procedure as employed in many TREC-experiments (Section 2.2). Based on that, the special challenges associated with argument retrieval evaluation are addressed (Section 2.3). Concerning the assessment of argumentative relevance (Section 2.4) and quality (Section 2.5), several existing studies are taken into account. Additionally, multiple evaluation metrics are considered with regard to their suitability to the task (Section 2.6). Additional attention is payed to the crowdsourced setting the annotations will be compiled in (Section 2.7).

2.1 Related Work

In general, an argument search task is comprised of two steps: argument mining and argument retrieval. Argument mining deals with identifying such parts of a text which can be regarded as argumentative. Argument retrieval then tries to accurately identify which arguments are most desirable to a user regarding their search query. Based on this, two main strategies to argument search can be identified. They mainly differ in the order the two steps take place.

One approach first mines arguments offline and builds an index of argumentative text. Retrieval is then performed on this index. This strategy is employed by Wachsmuth et al. [53] and has the advantage of faster retrieval times and generally higher quality content, although it can be limited in topic coverage. The other approach first performs the retrieval step, identifying those documents in a large collection of text that most likely contain the desired arguments. Mining is then applied online on the identified documents, extracting the argumentative parts. This strategy is used by Stab et al. [49],

which allows for a much bigger index and in turn a high probability of topic coverage, but comes at the price of higher retrieval runtimes.

This work focuses on the evaluation of the retrieval step, but not the argument mining step. Therefore, a constraint is put on the developed method by assuming that the retrieval operates on an index comprised of already segmented arguments, such as the `args.me` index. While both described approaches could be evaluated in a comparative manner, in order to do so, they would have to operate on the same index. As `args.me` and `ArgumenText` operate on different datasets, for practical reasons, the `args.me` index was chosen to test the capabilities of different retrieval models. However, this limitation is due to practical, not conceptual reasons.

So far, only minimal systematic evaluation has been conducted in work published on the topic of argument retrieval. Wachsmuth et al. [53] provide a first performance insight through examining the topic coverage of their engine by checking against a list of 1082 controversial issues on Wikipedia[1] to assess the completeness of the underlying argument corpus. A ranking evaluation was not given. The evaluation work done by Stab et al. [49] is mainly concerned with assessing the capabilities of their mining approach and no systematic retrieval quality evaluation has been done. Similarly, Wyner et al. [58] give precision and recall estimations for their mining procedure, but do not include a retrieval evaluation. The evaluation given by Rinott et al. [41] is not really applicable for an IR setting, as their approach focuses on testing the conversational abilities of their system, not the quality of the search result.

It is therefore necessary to develop a TREC-style IR evaluation methodology for argument retrieval in order to establish a baseline for future comparisons between retrieval approaches.

2.2 General Experiment Design

In order to meaningfully evaluate the performance of a search engine, regardless of domain, evaluation criteria have to be defined. Insight about them is gained through an evaluation procedure. Gordon and Pathak [17] differentiate between two types of search engine evaluation: testimonials and shootouts. Testimonials compare engines based on features apparent to the users. Such features can include speed, ease of use, interface design, and search capabilities like query operators. In essence, testimonials cover the qualitative assessment of a search engine. Shootouts on the other hand represent the quantitative assessment of a search engines capabilities.

[1] `https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues`

The classic approach to quantifying the performance of information retrieval systems, which is also largely employed in TREC evaluations, follows the *Cranfield paradigm* [13]. It derives three axiomatic assumptions to simplify the evaluation task: (1) retrieval performance can be approximated by topical relevance (2) a single set of judgments for a topic is representative of the user population (3) the list of relevant documents for each topic is complete. The term *topic* here refers to a statement of information need, as is usual TREC terminology.

These assumptions can rarely be fully met in real experiment conditions, for different reasons. The collection size may be simply too big to provide complete relevance judgments, not all relevant documents may be equally desirable in a search output or judgments may significantly differ between users. In order to resolve this conflict, a standard experiment design has emerged, adhering to the assumptions as close as possible while still being practical to use. It is described by Voorhees [51] in the following way: each to-be-compared retrieval model produces a ranked list (decreasing by relevance as assessed by the model) of documents for each topic in a test collection. The length of these lists, i.e. the number of retrieved items is called *depth*. These lists are combined into a pooling per topic. The models' performance for a single topic is computed as a function operating on the pooled lists, as it is assumed that the pooling reflects a near-complete and representative subset of the total index. The effectiveness of the model is then computed as the average score across the set of topics in the test collection. In conclusion, three different parameters can be adjusted to adapt the general evaluation experiment to the specialized task of argument retrieval: (1) the pooling, through adaption of number and depth of topics as well as the search queries used (2) the annotation process, quantifying the items in the pooling (3) the evaluation metric operating on the annotated scores

Additionally, Gordon and Pathak [17] formulate several requirements an evaluation should adhere to in order to make it accurate and informative: (1) topics should be motivated by real information needs (2) a sufficiently large number of topics should be used (3) the evaluation should encompass most major search engines in the domain (4) relevance judgments should be user-based (5) performance measurements should use accepted IR metrics, be statistically tested, and be taken in well-designed experiments

2.3 Differences to general retrieval evaluation

The described general experiment design is widely used to evaluate ad-hoc retrieval tasks. Argument retrieval, as a special domain of ad-hoc retrieval, faces

unique retrieval challenges that have to be addressed in the evaluation. Ad-hoc retrieval assumes topical relevance as the usual parameter for evaluation. However, for argument retrieval, relevance is not the only parameter that the search result should maximize for: having search results of high argumentative quality is desirable as well. This needs to be reflected in the ranking: for two items of similar topical relevance, the one with higher argumentative quality is to be ranked higher. Therefore, the second axiom of the Cranfield paradigm has to be extended to not only include topical relevance but also inherent item quality. This creates the need for a definition and viable annotation strategy for argument quality. The concept of topical relevance for arguments has to be revisited as well.

But: before assumptions about relevance and quality of arguments can be derived, the term itself has to be characterized. While numerous different definitions of "argument" exist, in the context of argument search, it refers to a text span which represents a unit of reasoning. This notion of arguments as *justification* describes an argument as a set of premises which together justify a conclusion [22]. An argument is the product of a rational thought process, where a set of claims is connected in a logical way to support another claim [18, p. 1].

2.4 Argument Relevance

Topical relevance of items is always related to the information needs motivating a search, as topics should aim to represent those. Therein lies a problem: no studies about information needs in the domain of argument search engines have been conducted yet. No established data source for real world search queries is available, nor can the logs of argument search engines be used to extract them, as they are not yet widely used in public and are likely prone to contain a significant amount of test queries by the search engine creators themselves.

However, approaching the issue from the view of argumentative theory, two different kinds of information need can be theorized: the *supportive* need, which assumes that the user utilizes the search engine to find supportive arguments for their predetermined opinion about the topic and the *deliberative* need, which supposes that the user utilizes the search engine to gain knowledge about a topic they were uninformed about prior to the search.

These two hypothesized information needs can be closely mapped to two different interpretations of the concept of argumentation: it can refer to a communicative process between parties, where either one is trying to win over the other (argument as *controversy*, [25]) or to a shared collaborative communication process to gain insight about an issue (argument as *debate*, [42]).

The first one reflects the *supportive* information need, the second more more closely relates to the *deliberative* information need. This distinction is further supported by the work of Mohammed [32], which differentiates between *intrinsic* and *extrinsic* argumentation goals. An *intrinsic* argumentation goal is convincing an opponent of the acceptability of an opinion [50]. *Extrinsic* argumentation goals are expressed as inquiry to base decision-making on [27, p. 12].

Relating these information needs to a single definition of topical relevance proves to be difficult, as both warrant for different kinds of arguments and therefore a different notion of item relevance would have to be utilized.

As of now, the available search engines treat argument search as an ad-hoc task. This is slightly problematic for catering to the *supportive/intrinsic* information need: while users could specify their own opinion through opinionated and/or stanced queries, the debate context is not available to the search engine, as supplying this context would be more akin to task-based retrieval. Satisfying the *deliberative* information need is not problematic, since it is more closely resembled in the standard ad-hoc search environment. This makes accepting the *deliberative* need as default option regarding information needs seem more appropriate, which also simplifies the evaluation process, as the traditional evaluation framework for relevance can be reused here. Therefore, the standard ad-hoc-relevance annotation process is used, employing graded relevance scales as measuring device. This procedure is reviewed for the use on argument data in the pilot study that is part of this work (Chapter 3).

This default choice is additionally motivated by the fact that using relevance as evaluation criterion for task-based retrieval has recently been challenged. Belkin, Cole, and Bierig [6] propose "usefulness" as more fitting criterion, which is closely related to aspects of argumentative quality. Thus, the *supportive* information need being more akin to task-based retrieval would also be reflected in a quality-based evaluation. Therefore, treating the ad-hoc aspect as default for relevance annotations and splitting the evaluation task into relevance and quality of items is additionally affirmed, since both information needs can be captured more accurately.

2.5 Argument Quality

The notion of argument quality stems from the assumption that different arguments can be unequally attractive to a person engaging in a debate and the audience of such a debate. By extension, the basic assumption about argument quality is that different arguments can be placed on a scale ranging from low to high quality. However, it is difficult to define what this scale should mea-

sure, as an argument may exhibit various quality traits that would be judged differently on such a scale. To accurately assess an arguments' quality, these traits have to be identified and measured independently of each other.

Extensive groundwork in the field of argument quality assessment has been done by Wachsmuth et al. [52], suggesting a detailed taxonomy on which argument quality estimations can be based. They differentiate between three main aspects of argument quality, each of which are associated with several quality dimensions: the *rhetorical*, *logical*, and the *dialectical* aspect. This trichotomy is widely accepted when describing argument quality: Wenzel [57] describes arguments as being a *rhetorical* process, a *dialectical* procedure, and a *logical* product; this distinction is also made by Habermas [19]. Following Wenzel [57] and Blair [7], the three aspects can be described in the following way:

Logic - the *logical* aspect of an argument refers to its structure, its parts and how they are combined. An argument of high *logical* quality is based on true premises and combines them in a valid way to support the arguments' conclusion. It has a clearly stated claim that is supported by acceptable, relevant, and sufficient evidence.

Rhetoric - the *rhetorical* aspect of argument quality groups notions of persuasive effectiveness, correct language, vagueness, and style of speech. An argument of high *rhetorical* quality is well-written and appealing to the audience.

Dialectic - the *dialectic* aspect captures an arguments' contribution to the discourse. Dialectics can be described as cooperative method to base decision-making on. An arguments' ability to contribute to that procedure, its ability to resolve the argumentative conflict at hand, or in short, its usefulness is the core principle of the dialectic perspective on quality. This usefulness also includes the arguments' robustness against possible refutation by the opposing debate party.

While Johnson [28] agrees with the aspect trio in general, he argues that the three are often defined differently by different people: there is not *one* understanding of each aspect, but rather different approaches to define them. This poses a significant challenge when generating data from annotations made by human judges to gain information about argument quality, because, as a consequence of the different views on how to define the aspects, the annotation of argument qualities is inherently different from classical annotation tasks.

In contrast to common annotations tasks in other NLP domains, such as POS tags, dependencies, etc., which are essentially driven by an underlying, well-researched common ground (grammar) that every judge adheres to to

some extent, annotation of argument quality is a much more subjective task. This is reflected in the three central challenges for argument quality assessment as identified by Wachsmuth et al. [52]: (1) argumentation quality is assessed on different levels of granularity; (2) some parts of argumentation quality are subjective in nature; (3) overall (e.g. non-specific) quality is hard to measure.

When tasking test persons with assessing argument quality, these three central problems need to be reflected upon in the experiment design. The measurement tool used needs to be able to deal with subjective measurements, e.g., be able to extract a general trend from single judgments. Also, a clear operationalization of the aspects has to be given to the annotators.

Different existing studies have concerned themselves with measuring argument quality, albeit the specific measured aspects differ. In the *Dagstuhl-15512 ArgQuality Corpus*[2] published as part of Wachsmuth et al. [52], 320 arguments were annotated for all of the 15 theorized quality dimensions by three experts. A 1 to 3 Likert scale was used to express each arguments quality per dimension. Albeit the scale encompasses only few steps and experts can be assumed to have a somewhat similar common ground for judgements, the Krippendorff's α agreement between the three is fairly low ranging from $\alpha = 0.27$ to $\alpha = 0.51$ depending on dimension. The overall mean agreement is $\alpha = 0.41$. Usually, the lower bound for α-values in reliable data is put at 0.667 [30, p. 354].

An alternative approach to the task includes ranking from pairwise comparison data. The `UKPConvArg1` corpus compiled by Habernal and Gurevych [20] includes pairwise annotations of arguments gathered in a crowdsourced setting. For a total of over 16 000 pairs of arguments over 32 topics, each comparison was annotated by five different crowd workers on Amazon Mechanical Turk. The annotation was made with respect to an arguments' convincingness. While they do not provide α statistics, they conclude that relative assessment in a crowdsourced setting is sufficiently accurate since the best ranked rater for each pair achieves 0.935 accuracy in comparison to the gold label.

When dealing with pairwise comparisons, two problems are apparent. First, the amount of needed annotations is much higher than in absolute quality assessments. At worst $\binom{n}{2} \cdot x$ annotations have to be made, where n is the number of to-be-annotated items and x is the amount of workers tasked for each comparison. Secondly, a mathematical model is needed to convert the pairwise comparisons into a ranking. Different models are available for such a task. If merely a total order of elements is needed, simple sorting algorithms based on item-wise comparison suffice. However, for the task at hand, an interval scale level is desired. Therefore, the model must produce an accurate interval scaling from pairwise data and preferably accurately operate on an incomplete

[2]Data is available via `http://argumentation.bplaced.net/arguana/data`

set of comparison data, since the cost for producing comparison data is very high with larger item counts.

Habernal and Gurevych [20] propose the use of *PageRank* [37] to embed the items into a scale. However, this approach is problematic for a number of reasons: first, no data was collected on how the *PageRank* method performs on incomplete comparison data; therefore, no strategies to effectively reduce the annotation workload can be implemented. Secondly, cycles in the argument graphs may form rank sinks, distorting the latent rankings. Habernal and Gurevych deal with this problem by constructing a directed acyclic graph from the collected data prior to applying *PageRank*, under the assumptions that argument quality exhibits the property of total order. However, no prior evidence for this property is apparent. Also, the conversion to an acyclic graph may introduce data bias.

In conclusion, pairwise assessments are a promising tool to collect argument quality judgements in a crowdsourced setting, as it overcomes many of the problems found in absolute quality ratings. However, using *PageRank* for rank embedding of arguments is problematic at best, which creates a need for a better embedding model. To additionally get a better understanding of why a Likert-based annotation procedure is not optimal and to subsequently develop a method that is not prone to its shortcomings, a pilot study was carried out, tasking crowd workers with assessing the three argument quality aspects on 1 (low) to 4 (high) Likert scale. The process and results of this pilot study are described in detail in Chapter 3.

2.6 Evaluation metrics

Most of the frequently used evaluation metrics in the field of information retrieval are derived from *recall* and *precision*. However, these metrics have been shown to not be robust with incomplete relevance judgments [9]. While extensions of precision-based metrics (AP) exist for incomplete data through random sampling, AP metrics usually operate on binary relevance data – an item is either relevant or not. However, in argument retrieval, the relevance of items could be considered non-binary, as test persons can accurately decide between different levels of relevance (see Chapter 3). The quality of items should be regarded as well. It is inherently a non-binary value, as it can be expressed on a quality scale. Therefore, the evaluation metric should be able to integrate both non-binary relevance and non-binary quality of items into the performance assessment.

A metric that has found widespread application and can meaningfully incorporate non-binary judgements is *normalized discounted cumulative gain* (NDCG, [26]). It computes a performance score by comparing the retrieval

models score against an ideal score, with both scores being calculated by adding up a gain value associated with each item up to a position k and discounting items that appear later in the ranking. Different versions of the NDCG metric exist, varying in the discount factor or associated gains. In this work, the NDCG formula provided by `trec_eval`,[3] a standard implementation of different metrics used throughout the IR community is used. It is given by

$$\text{NDCG} = \frac{\text{DCG}}{\text{IDCG}}. \tag{2.1}$$

DCG is the discounted cumulative gain of a given model on a given topic, formulated as

$$\text{DCG} = \sum_{i=1}^{k} \frac{g_i}{\log(i+1)}. \tag{2.2}$$

Here, g_i is the relevance score for item i and k is the number of items in the topic (k-depth). IDCG represents the ideal ranking computed on the pooled items of all models on this topic and represents the highest possible (ideal) DCG score a model could possibly obtain on this topic under the assumption that the pooling is representative.

The NDCG formula can thus be rewritten, considering a given permutation (ranking) of elements $\sigma \in \mathcal{P}$ where $\mathcal{P}$ is the set of all permutations of items in the topic and g is a function that assigns a gain to each item of the permutation:

$$\text{NDCG}(\sigma) = \frac{1}{\max\limits_{\sigma \in \mathcal{P}} \sum_{i=1}^{|\sigma|} \frac{g(\sigma_i)}{\log(i+1)}} \cdot \sum_{i=1}^{|\sigma|} \frac{g(\sigma_i)}{\log(i+1)}. \tag{2.3}$$

Apart from being the standard choice, on a variety of datasets, using a logarithmic discount function shows near-optimal performance in comparison to the hypothetical optimal function for a given dataset, while also being more stable than by example Zipfian or linear discount functions [29].

NDCG proves to be sufficiently robust with incomplete judgements [60]. It has to be mentioned that AP metrics can be extended to incorporate graded relevance, but NDCG offers several advantages by taking both the degree of relevance and the rank position into account [26].

The performance of a retrieval model is judged by the average score over all topics. In order to meaningfully interpret NDCG results, confidence intervals can be calculated using a bootstrap method [45]. Reporting results from simple metrics alongside the results from more complex ones is deemed redundant,

[3] `https://github.com/usnistgov/trec_eval`

since they do not provide additional or better information [56]. Therefore, the NDCG metric will be used as sole basis for the evaluation.

2.7 Crowdsourced data annotation

Recently, crowdsourcing has become a viable platform for conducting data annotation tasks. It allows for experiments to be conducted extremely fast, with good results and at low cost [3]. Given a well-constructed experiment design, annotation tasks conducted on crowdsourcing platforms achieve results at a similar level as expert-curated data [4]. Habernal and Gurevych [20] deemed crowdsourcing to be a viable solution to conduct data annotation in the domain of argument quality.

Given a well-designed experimental setup, using crowdsourcing platforms to conduct the data annotation procedure of this work is an acceptable source of data. Following the considerations of Alonso and Baeza-Yates [3], when designing a crowdsourced annotation task, three central challenges have to be addressed: (1) data preparation, (2) interface design, and (3) worker filtering. Snow et al. [47] provide further insight into constructing annotation studies on crowdsourcing platforms. They suggest addressing the challenges by formulating tasks as minimal and simple as possible and give demonstrative examples. Tasks should also be restricted to multiple-choice or fixed range numeric inputs.

Most platforms also provide filtering procedures to only allow qualified workers to complete a task. Such procedures should be employed to increase data reliability. To further strengthen reliability, annotations can be made independently multiple times and then be aggregated to achieve high reliability.

Thus, an aggregation method has to be formulated to accurately combine multiple annotations into a gold label. For binary response data, a simple majority vote may be used. For scaled responses, the mean response may be accepted as gold label [47].

Chapter 3

Pilot Study

3.1 Objective

The objective of this pilot study is twofold: on the one hand, to aquire first insight into the retrieval performance of the four different retrieval models in the domain of argument retrieval. On the other hand, to review the process of annotating argument quality and relevance using Likert scales in a crowdsourced setting, since no such study has been carried out yet. This allows to subsequently develop a better methodology for annotating argument quality. Of special interest is the annotation quality in regards to potential annotation bias and overall agreement of annotators. Also, the chosen quality aspects will be critically evaluated regarding whether they can indeed be accurately reflected by crowd annotations. A related expert study [54] is used as baseline to compare the attained results to.

Concerning the evaluation of the Likert scale, a first problem are the statistical tests in use: recent papers on the matter of argument quality use parametric statistics for insight into argument quality [54, 52]. In expert annotations, it can be assumed that the items of the Likert scale used for quality assessment are perceived as equidistant by test persons, which allows the scale to be interpreted as interval data, permitting the use of parametric statistics [34]. However, in crowdsourced annotations, this assumption is not valid, which restricts the analysis to non-parametric statistics. This complicates a direct comparison between the two. While this analysis mainly relies on non-parametric statistics, for the sake of comparison, a parametric version is additionally calculated if appropriate, as for most statistics, the difference between parametric and non-parametric versions is minimal on Likert data [33].

	Search issue: *plastic bottles*	
	Unbiased Version	Biased Version
Query	plastic bottles	ban plastic bottles
Description	You read about the risk of plastic bottles in a newspaper article. Trying to form your personal opinion, you search for pro and con arguments.	You recently watched a film describing the dangers plastic poses to human health. You now want to persuade your friends to support a ban of plastic bottles. You use the search engine to find suitable arguments.

Table 3.1: Example for differences in biased and unbiased topics

3.2 Design

20 search issues were formulated. For each, a biased and an unbiased version were derived, slightly altering the query to reflect a preexisting stance or not, and supplementing a description of the search scenario and stance the annotator should take. For an example, see Table 3.1.

Each of the resulting 40 topics was assigned to a test person. Depending on whether the topic was biased or unbiased, the test person was further asked to adopt a given stance as context for their annotations. Test persons were then tasked to judge each argument in the topic on a 4-point Likert scale ranging from 1 (low) to 4 (high) for every quality dimension and for relevance. Non-arguments automatically received a score of -2 in all categories. The arguments were presented to the test persons in random order to minimize order effects.

The test persons were recruited from a group of 170 students / 20 instructors having received a national scholarship for gifted students. A high educational background, personal integrity, and interest in societal issues can be assumed, although test persons by design only provided very little personal data in the study. About 77% of the test persons were male, and 23% were female. The mean age was 26, with the oldest annotator being 53 and the youngest being 18 years old. Participation was on a voluntary basis. The test persons did not receive compensation for their work.

3.3 Data

The argument data stems from the runs of four different retrieval models on the `args.me` index: a modified BM25F model [53] and three retrieval models from the Terrier collection [36], namely DirichletLM [61], TF/IDF [48], and DPH [5]. The pooling is done at a depth of $k = 5$. All results in the pooling were

flagged regarding whether they are indeed an argument or not. Only those which represent an argument are part of the annotation study.

Ranking data was aggregated for each engine, with some items appearing in the ranking being duplicates. The duplicate entries were unified, and only one of the duplicates was annotated. Finally, the annotation data was compiled into four different datasets, one for arguments, one for annotators, one for topics, and one for rankings. Full dataset schemes describing the contained data in detail are included in Appendix A.1.

In the ranking dataset, Topic 9 is missing for engine BM25F, as seemingly no items were retrieved by BM25F for that topic. In total, the argument dataset includes 494 items, consisting of 437 arguments and 57 non-arguments. 242 arguments were annotated twice, being part of different topics. No arguments appear more than twice. The 5 argument difference between theoretical upper bound for multiply-annotated arguments (247) is due to the missing 5 annotations for Topic 9 in the results of BM25F. The difference between the actual number of arguments and the theoretical upper bound of annotations (494 vs. 800, as 5 items were retrieved for 40 topics by 4 models) is due to some models sharing arguments in their results. The annotator dataset includes 40 persons, with one one person deciding to not disclose their age.

The pandas framework[1] was used for data import and conversion, as well as simple analysis tasks. Additional functions used for statistical analyses were imported from the scikit-learn [38][2] and scipy[3] frameworks. The dataset is publicly available.[4]

3.4 Analysis

3.4.1 Argument Quality

Distribution

A fairly even number of pro (208) and con (195) arguments is included. For a small subset of arguments, no stance is given (34). Figure 3.1 shows the distribution of annotated scores per quality dimension. While *Logical* and *Rhetorical Quality* have a similar distribution spiking at 3, *Dialectical Quality* has a relatively uniform distribution. However, they all aggregate towards the middle. Apart from being a data-inherent, this aggregation could partly be attributed to the central tendency and reluctance to give extreme answers

[1] https://pandas.pydata.org
[2] https://scikit-learn.org/
[3] https://www.scipy.org
[4] https://git.informatik.uni-leipzig.de/lg80beba/argument-quality-evaluation/

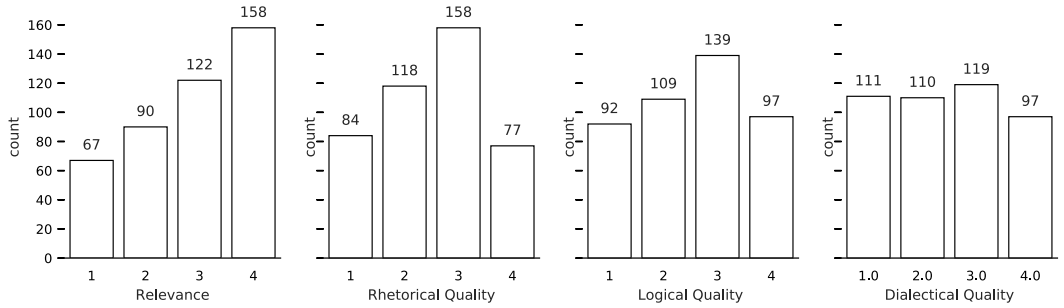

Figure 3.1: Distribution of argument scores

of test persons in Likert-based questions, as the Likert-scale used to question these qualities only spans 4 steps [2]. Also, no neutral or 'cannot judge' option was given, thus forcing a decision, which will then likely fall in the middle ground. The absence of these tendencies in the distribution for *Dialectical Quality* may be due to the unclear operationalization of this attribute, i.e. test persons being unsure about how the attribute is to be judged at all – an issue that was remarked in test persons comment's. Operationalization of the qualities was problematic in itself, as only minimal to no descriptions for argument qualities were given in the study. The distribution of *Relevance* being skewed towards the upper end (a lot of highly relevant arguments retrieved) is a promising sign for the analysis of retrieval performance later on, as the models seemingly retrieve a lot of highly relevant arguments in their first few results, thus accomplishing the retrieval task.

Correlation

Spearman's ρ correlation coefficients for combinations of quality aspects are given in Table 3.2. Overall, a high correlation between all of the measured quality dimensions can be identified. However, when comparing the annotated quality with the annotated relevance of the argument, differences between the quality dimensions are apparent: relevance correlates the most with dialectical quality of an argument, less so with logical and rhetorical quality. This is nevertheless expected for two reasons: dialectical quality and topical relevance are also closely related in argumentative theory [54], and, judging from comments on the study, annotators found it generally hard to differentiate between the two. Comparing correlation in pro and con arguments only, no significant differences can be found, substantiating a consistent annotation.

To compare correlation coefficients with expert annotations, Pearson's ρ was additionally calculated (Table 3.3), showing only minimal differences to Spearman's ρ. The crowd annotations largely reproduce the results found in

	Relevance		Rhetorical		Logical		Dialectical	
Rhetorical	0.40				0.65		0.59	
Logical	0.48		**0.65**				**0.69**	
Dialectical	**0.71**		0.59		**0.69**			
	Pro	Con	Pro	Con	Pro	Con	Pro	Con
Rhetorical	0.45	0.35			0.65	0.63	0.61	0.56
Logical	0.49	0.46	**0.65**	**0.63**			**0.70**	**0.67**
Dialectical	**0.69**	**0.73**	0.61	0.56	**0.70**	**0.67**		

Table 3.2: Spearman's ρ correlation coefficient cross-table, per stance, maximum per combined column marked

	Relevance	Rhetorical		Logical		Dialectical	
		Crowd	Expert	Crowd	Expert	Crowd	Expert
Rhetorical	0.38	-	-	0.65	**0.81**	0.60	0.75
Logical	0.48	**0.65**	**0.81**	-	-	**0.69**	**0.78**
Dialectical	**0.70**	0.60	0.75	**0.69**	0.78	-	-

Table 3.3: Pearsons' ρ correlation coefficient in expert[5]and non-expert annotations per argument quality, maximum per column marked

	Rhetorical	Logical	Dialectical
Non-Expert	0.27	0.26	0.32
Expert	0.45	0.44	0.50

Table 3.4: Krippendorff's α coefficient for expert[5] and non-expert annotations

expert annotations as far as relative comparison of scores goes, i.e., *Rhetorical Quality* correlating more with *Logical Quality* than *Dialectical Quality*. The absolute scores on the other hand are lower by 0.09 to 0.16.

For further insight into how the correlation occurs, Figure 3.2 shows contingency plots for every combination of attributes. While the quality correlations are more or less evenly spread throughout the whole spectrum, for *Relevance*, most of the correlating pairs occur on high scores, which can be interpreted as an argument of high relevance generally also being of high quality, while high quality arguments not necessarily being also highly relevant. Analogous, low relevance does not indicate bad argument quality.

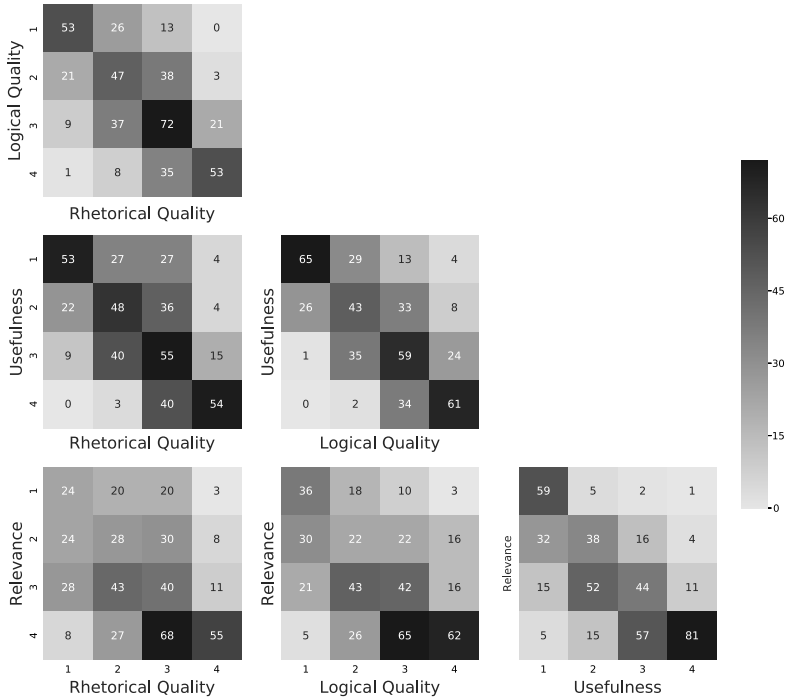

Figure 3.2: Contingency plot for every combination of attributes

Agreement

As some arguments were annotated once as part of a biased topic and once as part of an unbiased one, two ratings from different test persons are available per argument. As the argument-inherent attributes *Rhetorical, Logical*, and *Dialectical Quality* do not depend on the topic context, but only on the argument itself, a direct comparison between the two can be drawn for these attributes, even though topic query and topic description are different for both of the annotations. The inter-rater agreement between the two was measured using Krippendorff's α. Results for every attribute except *Relevance* are shown in Table 3.4. As expected, the data shows a much lower agreement than the expert baseline. This once again could be partly attributed to the missing operationalization to form a common ground between annotators. Such a common ground is implicit in expert annotations, but not necessarily given for non-experts, thus creating the need for an explicit formulation and description of the variables.

[5]Taken from Wachsmuth et al. [54], Table 3

			Argument Stance									
			Relevance		Rhetorical		Logical		Dialectical		$n = 199$	
			Pro	Con	Pro	Con	Pro	Con	Pro	Con	Pro	Con
Annotator Stance	Mean	Pro	**2.88**	2.54	**2.56**	2.35	**2.67**	2.38	**2.49**	2.21	57	48
		Con	**2.83**	2.81	2.51	**2.86**	**2.76**	2.54	**2.49**	2.43	37	35
		Neutral	2.78	**2.92**	2.54	**3.00**	2.31	**2.56**	2.23	**2.44**	13	9
	p-value	Pro	0.1112		0.1575		0.0880		0.0888			
		Con	0.4766		0.0908		0.1774		0.4054			
		Neutral	0.3349		0.0951		0.2612		0.2815			

Table 3.5: Mean argument scores and Mann-Whitney-test p-values cross-tabulation per argument stance / annotator stance. Maximum per Pro/Con pair row-wise marked. Significant p-values for $\alpha = 0.05$ marked.

Annotation Bias

A sufficient number of multiply annotated arguments is not available to conduct a full examination of annotation bias. However, since each annotator provided his own stance towards the topic, further insight into a potential systematic bias can be obtained, assuming that systematic bias would manifest itself through higher average annotated scores when annotator and argument share the same stance and lower average scores if they oppose. For calculations of mean, only arguments in unbiased topics were used.

Table 3.5 shows mean argument scores per annotator stance and argument stance. The first half of the table includes mean scores for every cross-category. While mean calculation is inherently a parametric measurement, and therefore not suited to derive conclusions from non-parametric data, in this case it is merely used to illustrate an effect. The hypothesis of a divergence between Pro and Con annotations is tested using a non-parametric test (Mann-Whitney). The p-values of that test for the pro/con-argument samples are shown in the second half of the table. Category sizes are shown on the upper right of the table. No significant divergence pattern in the average scores by stances is apparent. However, it cannot be determined whether a potential bias is masked by data-inherent differences, i.e., arguments in one category scoring higher on average because the arguments of this category in the dataset are objectively better. Such an effect could overshadow an underlying systematic bias. While inconclusive about the existence of a bias itself, it can be concluded that even if there is an undetected bias, its influence is low.

	Relevance			Dialectical			Rhetorical			Logical		
	Mean	95% CI		Mean	95% CI		Mean	95% CI		Mean	95% CI	
DPH	**0.65**	**0.61**	**0.70**	0.65	0.59	0.70	0.67	0.62	0.72	0.64	0.58	0.69
DirichletLM	0.64	0.60	0.70	**0.65**	**0.60**	**0.70**	**0.70**	**0.65**	**0.74**	**0.69**	**0.64**	**0.74**
TF_IDF	0.56	0.49	0.54	0.47	0.41	0.54	0.46	0.39	0.54	0.49	0.42	0.55
BM25F	0.46	0.41	0.48	0.42	0.37	0.48	0.45	0.40	0.49	0.42	0.37	0.47

Table 3.6: $\alpha = 0.95$ confidence intervals for mean NDCG@5 Scores per model and attribute, maximum per column marked

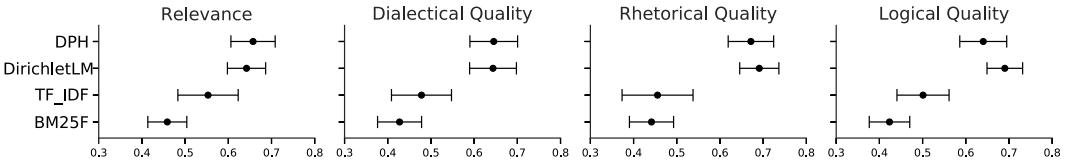

Figure 3.3: Confidence intervals for mean NDCG@5 scores

3.4.2 Ranking

To gain insight into the ranking performance of the different retrieval models, the NDCG was computed for every model/attribute combination. An ideal ranking was established by ordering all annotated arguments for the current topic descending by their attribute score. The mean NDCG scores across all topics calculated for $k = 5$. Confidence intervals were obtained using bootstrapping ($n = 10\,000$). Results are shown in Table 3.6 and the corresponding Figure 3.3. The significance of difference of the mean NDCG-Scores was additionally tested using a 1-way ANOVA test.

As a result, the following performance evaluation can be given: for *Relevance*, TF_IDF performs on par with BM25F and DirichletLM ($p > 0.05$ for all pairings); DirichletLM and DPH perform similar and better than BM25F ($p < 0.05$ for both pairings); DPH is better than TF_IDF ($p = 0.0313$). For *Dialectical, Rhetorical*, and *Logical Quality*, DirichletLM and DPH perform similar ($p < 0.05$ in all cases), and better than BM25F and TF_IDF ($p > 0.05$ in all cases), which are on par with each other ($p < 0.05$ in all cases). Since DPH and DirichletLM perform very similar judging from their NDCG-scores, the assumption could be derived that they often perform good on the same topics, essentially retrieving the same arguments. This hypothesis can be rejected, since the two models have only an overlap of 38% in all retrieved arguments and for only 5% of the topics the same arguments are retrieved, which then are mostly ranked differently. For no topic, the two models score the same.

Another criterion that allows better assessment of the models performances is the variance of scores [55], since even if the mean score of an engine is

	Relevance	Dialectical	Rhetorical	Logical
DPH	0.037	0.045	0.045	0.048
DirichletLM	**0.033**	**0.041**	**0.028**	**0.032**
TF_IDF	0.061	0.065	0.079	0.060
BM25F	0.040	0.041	0.037	0.038

Table 3.7: Variance of NDCG scores per engine and dimension, minimum per column marked

	Relevance	Dialectical	Rhetorical	Logical
DPH	**18**	16	**15**	11
DirichletLM	13	**18**	**15**	**19**
TF_IDF	8	3	5	7
BM25F	5	6	5	4

Table 3.8: Amount of times a model scored highest in a topic. Total number per column can exceed 40 due to ties. Maximum per column marked.

very high, a consistent performance across topics would be desirable. For score variance per model and attribute, refer to Table 3.7. While DPH and DirichletLM are not distinguishable by means of their mean NDCG scores only, taking the variance of these into account, DirichletLM consistently achieves the lowest NDCG variance for every attribute. Comparing BM25F and TF_IDF, BM25F achieves a much lower variance across the board. A similar trend is apparent when taking the absolute number of topics a model scored highest out of the 4 into account (Table 3.8). Here, DirichletLM outperforms in 2 categories, being tied in 1. DPH scores best for *Relevance*.

Taking into account mean performance, variance, and absolute number of highest scored topics, a clear ranking of the retrieval models for the task of argument search on the given corpus can be established: DirichletLM performs best, closely followed by DPH. Seperating TF_IDF and BM25F still proves difficult, as TF_IDF achieves higher scores on average, but BM25F being more consistent. If only judged by *Relevance* of search results, not their argument quality, DPH seems to be the best performing model.

3.5 Conclusion

It has been shown that non-expert annotations of argument quality largely reproduce the data found in expert annotations. The distinction between the three quality dimensions is warranted and adequate. Test persons were able to produce consistent data and a potential annotation bias could not be verified in practice. Asking annotators to adopt a stance or provide information about

their own opinion therefore seems not necessary, simplifying future studies. However, agreement and operationalization of quality dimensions remain a problem. In future studies, a clear explanation of what each of the quality aspects encompasses should be given.

The use of a Likert-scale as measuring tool for the annotation task of argument quality should be discarded, as the agreement between annotators is too low. It is not suited to capture the subjective nature of the latent quality scale. Thus, while relying on crowdsourced annotations for evaluation data is fine, it is recommended to employ a different annotation method in future studies. A pairwise comparison approach seems promising.

Test persons successfully distinguished different levels of relevance thus the graded relevance approach to performance assessment should be used in argument retrieval evaluation. The usage of the Likert scale to obtain graded relevance judgements was adequate for the task. Using 4 steps seems to be a good choice as well, since the resulting score distribution is rising monotonously, thus not hinting at a too small category size. Using NDCG with crowdsourced relevance assessments provided a clear separation of the retrieval performance of the different models. It is suitable to incorporate both Relevance and Quality assessments. The metric can therefore be used in future evaluations as well. Separation per variance and won topics is optional and can be employed if two models give no indication based on their NDCG scores alone.

Chapter 4

Methodological Approach

4.1 Experimental Requirements

As outlined in Chapter 2 and confirmed in Chapter 3, the Likert scale is adequate to collect relevance ratings from test persons, but a ranking from pairwise comparison data is the more promising approach to assess argument quality. This creates the need for an embedding model to project the items onto a scale based on the available comparison data.

Multiple methods are available to generate rankings from relative comparisons of items. While other ranking methods which use comparisons between more than two items at once, compare in groups, use sorting groups etc., are available, a pairwise comparison is favorable for the task at hand. A direct comparison between two items requires less information to be processed by the test person than in any other ranking method such as triad test or pile sorts. Only comparing two items at a time also excludes the possibility of framing effects, i.e., two very similar arguments winning against a stronger one if they appear together, as they complement each others argumentation. While such framing effects for argument annotations have not been verified in practice yet, they are also not ruled out. Thus, the cautious approach of only comparing two items at a time is chosen. Also, an existing data set with pairwise comparisons for argument data exists, simplifying the evaluation of the model before applying it in practice.

Another benefit of employing pairwise comparisons and an appropriate model for embedding them is that the resulting quality scale is of an interval level, allowing the use of parametric statistical tests and in turn a high statistical power. Also, score distributions can be meaningfully interpreted and the collected data can be reused for multiple other applications such as machine learning tasks.

Symbol	Explanation
S	set of annotators
s	annotator $s \in S$
N	set of items
n	items $n \in N$
C	set of pairwise comparisons between items in N
C_s	subset of comparisons made by annotator s
$i \succ j$	item $i \in N$ being of greater perceived quality than item $j \in N$
$i \approx j$	items $i, j \in N$ have no perceivable difference in quality
γ	merit vector, with γ_n being the merit of item n
λ	regularization parameter
τ	minimum difference threshold

Table 4.1: Terminology

4.2 Annotation of Argument Quality

In the following section, a mathematical model is introduced to derive a ranking of items from pairwise comparison data. A description of the basic model used is given in Section 4.2.1. It is then extended in Section 4.2.2 and Section 4.2.3. Section 4.2.5 derives conditions on the experiment design optimal for the model. Its performance is evaluated in Section 4.2.6, showcasing its capabilities on real data. An overview on the mathematical symbols and terminology used throughout this section is given in Table 4.1.

4.2.1 Model Description

The Bradley-Terry Model [8] is a method of ranking items based on paired comparison data. It is often used to infer a latent ranking when no natural ranking of items is readily available. The wide range of recent applications includes sports [11], marketing [21], multi-class classification problems [23] or genetics [46].

The Bradley-Terry model usually assumes a binary comparison, i.e., an item is either better or not, with no information about the magnitude of difference. Such information could be obtained by a scaled preference rating and be integrated into the model, e.g., like described by Okamura, Kiyota, and Hiramatsu [35]. However, gathering information about the magnitude of difference is undesirable in our case as described before, since no common reference frame for this magnitude can be expected from annotators. For other experiments this may prove beneficial though.

The latent ranking can be expressed as the items in a set N having a true preference rating γ_i, called merit. The parameter space of γ is a continuous scale constrained by $\sum_i \gamma_i = 1$, as γ is scale-invariant. The items are compared pairwise with mutually different outcomes, which are assumed to be independent. The probability of item i beating j is defined as

$$P(N_i \succ N_j) = \frac{\gamma_i}{\gamma_i + \gamma_j} \tag{4.1}$$

with $i, j \in N$. Using exponential score functions $p_i = e^{\gamma_i}$ reduces the model to a logistic regression on pairs of individuals [1]. A maximum-likelihood approach can then be used to infer the merit vector γ [24]. The log-likelihood equation for a pool of comparisons C is:

$$\mathcal{L}(\gamma) = \sum_{(i,j)\in C} \log\left(\frac{p_i}{p_i + p_j}\right) \tag{4.2}$$

The maximization is guaranteed to converge to the unique maximum likelihood estimator in finite steps under the assumption that in every possible partition of the items into two nonempty subsets, some subject in the second set beats some subject in the first set at least once [16, 24]. In order to conform to this assumption, the number and shape of pairwise comparisons is restricted in essentially two ways:

(i) The matrix formed by the comparisons must construct a strongly connected graph

(ii) The comparisons between the partitions cannot all be won by subjects from the same group, i.e., no item has losses or wins exclusively.

The model could be extended to compare more than two items at once [24]. However, due to the framing effects discussed prior, this is unfavorable.

4.2.2 Incorporating ties

The Bradley-Terry model can be extended to accommodate for ties [40, 14]. It is assumed that there is a difference threshold between two items under which test persons cannot meaningfully decide which item is better. This threshold parameter τ is incorporated into the model by Rao and Kupper [40] in the following way with $\theta = e^{\tau}$:

$$P(N_i \succ N_j) = \frac{p_i}{p_i + p_j\theta} \tag{4.3}$$

for the probability of preference of N_i over N_j and for the probability of no preference between the two

$$P(N_i \approx N_j) = \frac{p_i p_j(\theta^2 - 1)}{(p_i + p_j\theta)(p_i\theta + p_j)} \tag{4.4}$$

For $\tau = 0$, i.e., test persons being able to differentiate every item, these equations reduce to the standard Bradley-Terry model.

4.2.3 Regularization

The unique convergence of the maximum likelihood equations is only guaranteed if the convergence constraints (i) and (ii) (Section 4.2.1) are met. However, this depends on the collected data rather than the experiment design. Even though the adherence becomes asymptotically likely given an appropriate experiment design [59], additional strategies can be employed to ensure a unique solution for extreme data cases, such as outlier items that always loose in comparisons.

Following the regularization approach by Chen et al. [12], a dummy item t_0 is added with score e^{s_0}. This dummy item is assumed to compare against every item with exactly one win and one loss, thus transforming the model into a regularized maximum likelihood problem. Convergence is now ensured, as the graph is guaranteed to be strongly connected. A second benefit is that if e^{s_0} is fixed, the scale invariance problem does not occur, thus allowing the condition $\sum_i \gamma_i = 1$ to be dropped. By fixing s_0 at 1, we can define a regularization term $\mathcal{R}$ for the log likelihood equation:

$$\mathcal{R}(\gamma) = \sum_{i=0}^{|N|} \left[\log\left(\frac{e^1}{p_i + e^1}\right) + \log\left(\frac{p_i}{p_i + e^1}\right) \right], \tag{4.5}$$

which is multiplied by a regularization parameter λ and added to the unregularized log likelihood equation. Note that the fixation score of 1 is chosen rather arbitrarily, choosing another value would just shift the score distribution.

4.2.4 Log-Likelihood Maximization

The final log-likelihood equation, where the regularization parameter λ, and the difference threshold τ are fixed takes the form

$$\mathcal{L}(\gamma, \tau, \lambda) = \sum_{(i,j)\in C} \log \left[\begin{cases} P(N_i \succ N_j) & \text{if } N_i \succ N_j \\ P(N_i \approx N_j) & \text{if } N_i \approx N_j \end{cases} \right] + \lambda\mathcal{R}(\gamma) \tag{4.6}$$

4.2.5 Comparison Sparsity

Paired comparison experiments suffer from one drawback, that restricts their widespread usage in data annotation studies: the amount of comparisons needed to produce an accurate ranking is quite high, if a full comparison set is to be obtained, since the asymptotic bound for comparisons is $\binom{n}{2}$ for n items. For example, a ranking involving 20 items would require 190 comparisons, and a ranking involving 50 items requires 1225 comparisons. Therefore, sampling strategies are needed to reduce the amount of comparisons, while still maintaining a high annotation quality. Different strategies exist to accomplish this task.

The active learning approach formulated by Chen et al. [12] could be reused. After each comparison added to the total set of annotated pairs, they identify the next pair by calculating which comparison would reduce the overall model uncertainty the most, until a fixed number of comparisons is reached. While this approach would probably use a minimum amount of comparisons, it has two major drawbacks for the application proposed in this work: 1. The uncertainty evaluation is based on the Bradley-Terry model. While this ensures that the set of pairs is perfectly suited for this specific model, the reusability of the collected data for other purposes is diminished, as such a specific method of choosing pairs introduces data bias when other models are applied to the collected data. 2. An active learning approach is more complicated to implement on a crowdsourcing platform. Since this extra work comes with little benefit, active learning is deemed not suited for the task at hand.

Concerning other strategies, Burton [10] describes a design where items are arranged in a cyclical way. A main design feature is that each item is required to appear in the same number of pairs, in order to gain the same amount of information about each item. For a randomly ordered set of items, N, each item n_i is compared with with item n_{i+1}. Item $n_{|N|}$ is compared with item n_1, thus forming a cycle. This can be generalized to higher step sizes s, for example if $s = 2$, all items that are separated by two positions around the ring are compared. However, to ensure that the comparisons form a fully connected graph, different step sizes have to be combined for higher comparisons per item. For example, to have 2 comparisons per item, one cycle with step size $s = 1$ and two cycles with step size 2, offset by 1, can be used. However, just combining the two cycles of step size 2 would leave the comparison graph with 2 unconnected components. Finding the right combinations of step sizes for different desired amounts of comparisons can be a challenging task. While the assumption that every item should have the same amount of information available is a sensible approach, the high complexity of such designs is a drawback.

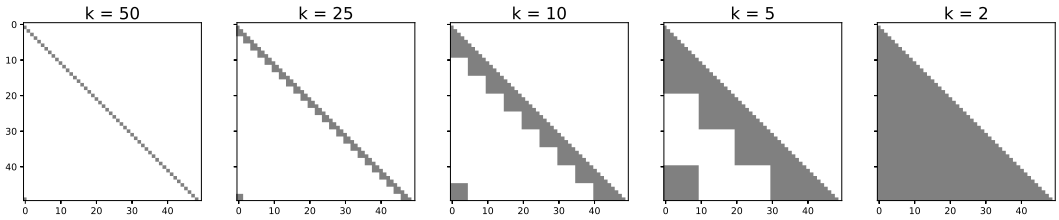

Figure 4.1: Comparison matrices for $n = 50$ and different values of k

Yan, Xu, and Yang [59] propose a method of sparse grouped comparisons, where N is the set of all items. N can be partitioned into K disjoint and nonempty subsets $N_k, k = 1, ..., K$ of equal size such that:

(i) for each N_k, $|C_{ij}| > 0$ when $i, j \in N_k, i \neq j$

(ii) $n_{ij} > 0$ when $i \in N_k, j \in N_{k+1}$ for $k = 1, ..., K - 1$

A favorable aspect of this design is the high probability of a strongly connected comparison graph and the low complexity of the comparison design. A drawback is that not all items have the same amount of comparisons.

Combining the two approaches, a cyclical grouped comparison design can be derived by also including comparisons between group N_1 and group N_k. In that way, every item has the same number of comparisons but the overall construction of the experiment design remains simple. All combinations of items in the same group and the cartesian product of adjacent groups are included. Therefore $k \cdot \binom{n/k}{2}$ intra-group comparisons and $k \cdot \left(\frac{n}{k}\right)^2$ inter-group comparisons are needed. Thus, the total amount of comparisons is

$$c = k \left(\left(\frac{n}{k}\right)^2 + \binom{n/k}{2} \right) \tag{4.7}$$

Example comparison matrices for $n = 50$ and different values of k are shown in Figure 4.1. Note that the comparison matrix is inherently symmetric – however, to reflect the true count of comparisons, only one half is depicted.

The number of groups should be a natural factor of the number of items, to ensure even group sizes, and thereby, that each item has the same amount of comparisons. This design combines the advantages of both described approaches while still being easy to implement in a real experiment setting. One may also opt to have every comparison annotated by multiple people to increase the data quality; in this case, c has to be additionally multiplied by the number of annotations per comparisons x.

Annotators	k	Annotations	Annotations %	Comparisons %	$\bar{\rho}$	95% CI	
	4	1840	74	74	1.00	0.99	1.00
5	8	880	35	35	0.97	0.92	0.99
	16	400	16	16	0.87	0.79	0.94
	4	1472	60	74	0.99	0.98	1.00
4	8	704	28	35	0.96	0.93	0.98
	16	320	12	16	0.84	0.78	0.92
	4	1104	45	74	0.98	0.97	1.00
3	8	528	21	35	0.94	0.86	0.98
	16	240	10	16	0.77	0.58	0.91
	4	736	30	74	0.98	0.95	0.99
2	8	352	14	35	0.90	0.85	0.97
	16	160	6	16	0.75	0.56	0.88
	4	368	15	74	0.94	0.89	0.97
1	8	176	7	35	0.83	0.74	0.91
	16	80	3	16	0.62	0.32	0.85

Table 4.2: Comparison sample rate and mean correlation coefficient for different cyclic group designs. The complete comparison set for $n = 32$ items includes 496 comparisons annotated by 5 annotators each (2480 total annotations).

4.2.6 Model Evaluation

In order to gain a first insight into the performance of the proposed model, it was tested on the `UKPConvArg1` corpus [20] which includes pairwise comparisons of arguments obtained with Amazon Mechanical Turk. The corpus includes full comparisons for a number of topics.

To test the accuracy trade-off between full comparison and sparse comparison designs, ten topics were randomly selected from the corpus. For each, 32 items were randomly chosen, since 32 has a lot of natural factors and therefore allows for multiple group sizes. For these 32 items, a complete comparison set was sampled. In the first step, the model was fitted on the gold labels provided in the corpus to establish a baseline. In the second step, different group sizes were used to sample a subset of the comparisons. The proposed model was fitted with each of the sampled comparison sets and different numbers of annotations per comparison. The merit ranking obtained was compared against the gold label ranking using Pearson's ρ. All merit vectors in the experiment were normalized to a $[0, 1]$ interval before the calculation of correlation coefficients. Confidence intervals were calculated using bootstrapping ($n = 10\,000$). Results are shown in Table 4.2. The derived model and comparison design are able to produce near-perfect rankings ($\rho = 0.94 \pm 0.04$) using only 15% and acceptable rankings ($\rho = 0.83 \pm 0.07$) using only 7% of the full comparison

set. This significant reduction is a promising sign for employing the model in crowdsourced studies on large item counts.

4.3 Annotation of Argument Relevance

As the pilot study has shown, the method of annotating graded relevance with Likert scales can be used. A scale spanning four steps seems to be a reasonable choice. As the models (currently) do not offer an option to specify the desired stance or contextual information for the search, the defined exploratory information need is accepted as default, therefore topical relevance can be judged just like in classic TREC-style IR evaluation.

4.4 Topic and Query Selection

In order to retrieve a test collection for evaluation from the search engines at hand, for each chosen topic, a query has to be formulated. Simulated test collections are problematic in the sense that they have to accurately portray the information need of potential users to serve as basis for meaningful evaluation. As simulated queries are used to build such a test collection, it is of high importance that the queries used are representative.

Using real-world queries, i.e., queries taken from a search engines' log as those are made by real users is not possible in our case, since the number of real users that use the search engines to satisfy an information need is low and the search logs consists mainly of test queries.

An alternative solution is to take debate issues from online debate platforms. However, crowdsourced debate platforms often feature poorly worded questions and the amount of possible queries is simply too high: filtering a diverse and representative range of topics from the whole set of available debates is a strenuous task. Another possibility to source search queries from are debate platforms that let experts aggregate arguments.

ProCon.org[1] is a non-profit online platform presenting expert-curated arguments on a variety of topics in a pro-con format. *ProCon* states to choose topics that are "1. Important to many Americans 2. Controversial 3. Useful to promote critical thinking, education, and informed citizenship 4. Complementary to ProCon.org's diverse subject offering" [39]. For these topics, *ProCon* formulates questions which are "usually worded deliberately so that a Pro response is generally considered to be Pro the topic and a Con response is generally considered to be Con the topic" [39]. These characteristics make *ProCon* a useful resource for topic queries, as it can be expected that they

[1] `https://www.procon.org`

cover a diverse range of real information needs and are adequately worded to retrieve meaningful and diverse search results. Choosing topics in such a structured and non-random way provides results with higher reliability [43]. Additionally, *ProCon* provides categories that the topics belong to.

4.5 Z-Transformed NDCG

Given the developed model for rank embedding, a problem presents itself: the fixation point for the regularization and therefore the fixation point for the derived score distribution is chosen arbitrarily. This is problematic if the scores are used to calculate the NDCG performance of models, since the scoring does not behave consistent and is dependent on this arbitrary parameter.

This is due to the sensitivity of NDCG to negative scores. When shifting the distribution around 0, the share of positive and negative scores changes, in turn increasing or decreasing the resulting NDCG value, since more negative scores increase the penalization. While the relative performance of models measured on the same score distribution stays the same to some degree, as they are equally prone to the increased and decreased number of negative items, the scores cannot be meaningfully interpreted, anymore, especially in comparison to prior experiments where a different score distribution/vote collection method was used.

While the first part of the issue can be addressed by choosing a sensible fixation point based on past experiments, the latter part remains problematic. To address this, an adapted formulation of NDCG is proposed: before calculating the NDCG performance index, the underlying score distribution is z-transformed, thus having a mean of 0 and a variance of 1.

This allows for a new interpretation of the resulting NDCG value: it now resides in the interval $[-1, 1]$. A score of 1 represents the perfect ranking, a score of -1 represents the inverse perfect (worst) ranking. The first benefit is that it allows for NDCG scores derived from different score distributions to be compared directly, since the distributions were normalized beforehand. The second benefit is that this interpretation gives rise to a new kind of information to be metered explicitly: performance in comparison to a random ranking.

This information is implicitly contained in the classic NDCG formula: the performance of a random ranking would equal the NDCG score if all entries in the ranking had the mean score. However, this value changes when the mean of the underlying score distribution changes and has to be calculated additionally to allow for the comparison to randomness to be derived. When applying a z-transform to the score data, the mean is standardized to be 0, thus the NDCG performance of a random ranking necessarily is close to 0 as well.

Therefore, the comparison is explicitly included in the new interpretation: if a score is in the intverval $(0, 1]$, the ranking is better than random and as the score approaches 1, the more accurate the ranking gets. If a score is in the interval $[-1, 0)$, the ranking is worse than a random one.

Applying a z-transformation to the item scores can also be justified from a qualitative point of view: the user of a search engine would like to receive the items from the index corresponding best the search task. Thus, each item that is better than the mean signifies an improvement to the search result, warranting a positive score. Each item that is worse than the mean signifies a decline in search quality, thus penalizing it with negative scores is justified as well. z-transforming the score data eliminates the need to explicitly define a penalizing threshold, thus also eliminating an interference factor if this threshold is chosen poorly.

One drawback to this interpretation of NDCG is a ranking which only includes items of the same score: since each ranking would essentially be random, the NDCG score under the new formulation necessarily equals 0. At the same time, each ranking would be perfect, implying an NDCG score of 1. However, it could be argued that if every ranking is equally good, no ranking provides more benefit to the user than others, thus justifying the 0-score.

In the context of this work, this new interpretation of NDCG is referred to as z-NDCG.

Chapter 5

Data Acquisition

5.1 Topic Data

To build a test collection, *ProCon.org* was crawled for search topics.[1] From the original set, all questions were omitted that were deemed as too US-specific (for example *"Was Bill Clinton a good president?"*), represent a multi-topic, covering an event (e.g., *"2016 Presidential Election"*) or were not an adequate question (e.g., *"School Vouchers – Top 4 Pros and Cons"*). As a result, a set of 50 topics was compiled, each featuring a query in long (*"Should Marijuana Be a Medical Option?"*) and short form (*"Medical Marijuana"*) as well as the category.

Based on the pilot study, sample size calculations were carried out using G*Power [15]. Given the expected NDCG score differences, an α-error probability of 0.05 and a β-error probability of 0.05, the resulting sample size is 9 topics to separate the search engines with the desired confidence. Similar results were obtained using the methodology for estimating topic counts in retrieval evaluation experiments described by Sakai [44], with 7 topics being the recommendation here.

These numbers seem unusually low, but are likely due to the fact that the BM25F model performs significantly worse, while DPH and DirichletLM perform nearly equal, thus easily allowing to distinguish the former, while the latter two remain not separable even with higher topic counts.

However, the sample size was increased to 20 topics, to account for potential methodological flaws influencing the pilot study. Thus, a reasonable tradeoff between statistical power and study cost can be made. From the 50 available queries, 20 were randomly sampled, which include at least one from every query category to ensure topic diversity.

[1] Website was last accessed on February 11, 2019

5.2 Retrieval Models

As shown in the pilot study, the four retrieval models BM25F, TF_IDF, DPH, and DirichletLM show different performance ratings on the corpus. However, the pilot study has methodological flaws and was only extensive enough to draw tentative conclusions. In order to gain a more accurate insight into the retrieval models performance, they will be subject to be tested again using the described evaluation framework. Due to limitations in experiment size, TF_IDF will not be included, as it has not shown promising performance in the pilot study. DPH and DirichletLM need to be more closely reevaluated to be able to gain conclusive insight into their performance, as those are nearly indistinguishable in the pilot study. BM25F is still included, too, since it is the model `args.me` currently runs on and therefore serves as a baseline.

5.3 Pooling

For selected 20 topics, a pooling was compiled by letting each of the three chosen models (BM25F, DPH and DirichletLM) retrieve items for the respective query at a depth of $k = 50$ from the `args.me` index [53]. It comprises more than 300 000 arguments sourced from debate platforms. The items in the index are pre-parsed and are already tagged with an argument stance (Pro/Con). Duplicates in the pooling were unified, keeping them as separate entities in the ranking, but pointing to only one unique argument in the annotated argument dataset.

In total, 3000 results were pooled, amounting to 1606 unique arguments.

5.4 Relevance Annotation

Amazon Mechanical Turk (MTurk)[2] was used for relevance annotations. A test person was presented with 5 spans of text and was tasked for each of them to: (1) decide if the text contains an argument or not; and (2) judge its relevance to a given topic on a Likert-scale ranging from 0 (not at all relevant) to 3 (highly relevant). The option to comment on the task was available. The questionnaire layout in use is included in Figure B.2. The classification into whether the text is argumentative or not was done to reduce the experiment size for quality annotations later, since the quality of non-arguments is not of interest. Also, since the search engine should ideally only retrieve texts that are indeed arguments, any non-arguments can be penalized in the evaluation later on.

[2] `https://www.mturk.com`

Each annotation was done by five test persons, to reduce the data error. To additionally ensure annotation quality, only workers were accepted for the task that have an acceptance rating of at least 95%. Each test person could opt to complete multiple sets of 5, which a high number of test persons did – this hints at adequate rate of pay, understandable operationalization, and a pleasant task design in general.

Majority vote (for argument classification) and mean (for relevance annotation) were used to derive a gold label. In total, 329 HITs (Human Intelligence Tasks) were carried out, resulting in 1645 unique assignments to test persons. Each assignment was paid \$0.08, putting the total annotation cost for relevance annotations, including Amazons fees, at\$148.28. The resulting data set is made available.[3]

Concerning the reliablitity of the collected data, Krippendorff's α was calculated for both the relevance and the is-argument classification. For relevance, $\alpha = 0.27$, for the is-argument classification $\alpha = 0.21$. While these values seem very low, the α-value for relevance judgements is inside the expected range for such a classification task [31]; the α value for the is-argument classification as well, and since it represents more of a preprocessing step for quality annotation the required reliability is low. However, since the scores are aggregated using mean/majority, further insight into the reliability is provided by measuring percentual class agreement. For relevance, the mean percentual agreement (as in: ratio of of test persons choosing the most selected option per item) is 79.4%. In 67% of the items, at least 3 of the 5 votes are for the same score, establishing a clear majority option. In nearly all of the remaining items, the deviation of voted scores is at most ± 1, which still allows for a reliable mean to be derived. For the is-argument-classification, two thirds of the items receive a classification with at least four of the five votes specifying one option, thus establishing a clear majority. One third of the items was classified unanimously. Therefore, the largest part of the items has a clear majority indication and the data can be seen as reliable.

Spammer detection was carried out by comparing votes against the majority label. When the disparity was bigger than one category size (± 1), the ratio of potentially mislabeled items to all items of the worker in question was calculated. Only about 6% of the workers "missclassified" over 30% of their items – however, the vast majority of those only classified 1 or 2 tasks (5 to 10 items), thus not giving a spamming indication. No case of recurring missclassifications of a volume that would indicate systematic spamming is apparent. Similar results are given when investigating spamming using inter-item variance as detection metric. No test persons shows a vote distribution that hints at spamming.

[3] `https://git.informatik.uni-leipzig.de/lg80beba/argument-quality-evaluation/`

5.5 Quality Annotation

The quality annotation task was carried out on MTurk as well. A test person was presented with pairs of two texts. For each pair, the test person was tasked to select the text that exhibits a higher quality than the other, in regard to a given description of the respective quality aspect. The annotation was carried out separately for each of the three quality aspects. To make the task accessible to test persons without prior knowledge of argumentative theory, the three quality dimensions were operationalized in a simplified way: "Which text has the better logical structure?" (logical aspect), "Which text has the better style of speech?" (rhetorical aspect), and "Which text would be more useful in a debate?" (dialectical aspect). An example questionnaire layout for the rhetorical aspect is included in Figure B.1. The questionnaires for logical and dialectical aspects only differ in the task description (see above) and examples. Five comparisons were presented together as one task. The comparison sets of five were compiled randomly to minimize order effects.

For the paired comparisons, a cyclic group comparison design as described in Section 4.2.5 with $k = 8$ was employed, with each pair annotated by one test person. On average, a topic pooling consists of $n = 64$ unique items, with $c_{\text{sampled}} = 698$ and $c_{\text{full}} = 2043$. The mean comparison sample rate regarding the full comparison set therefore is 0.342. Similar sample rates have shown a good performance prior, as such designs achieved a mean correlation of $\bar{\rho} = 0.83 \pm 0.07$ with a "perfect" ranking (see Table 4.2).

While the accuracy in comparison to a perfect ranking could be increased by tasking more than one person to judge each comparison (for example, having each annotated by two test persons would increase the correlation to $\bar{\rho} = 0.90 \pm 0.06$), the extra expenditure was not considered as warranted for this proof-of-concept work. As the individual scores are used in NDCG calculations later on, the impact of small score differences are minimal.

The annotation effort for this study was reduced by 93.17% using such a design.[4] Also, if higher data accuracy is needed in future experiments, the comparison set can easily be extended by adding additional votes per comparison or increasing the group size.

One drawback is that the number of items per topic differs and is not always divisible by 8. Thus, one of the groups has fewer members and therefore those receives less comparisons than the other items in the topic. While this could be circumvented by allowing duplicates in the item set to boost the item count to the same divisible number (i.e. 80) for every topic, it would greatly increase the annotation effort. As this comes with little reward in terms of annotation quality, the extra expenditure is not deemed necessary. This decision accounts

[4] In comparison to a full comparison set annotated by 5 test persons each

for little deviations of actual comparison count and theoretical comparison count derived using the formula given in Equation (4.7).

Spam detection is complicated, since no comparisons between different persons on the same task can be drawn. Therefore, the test person requirements were raised to a higher level: in addtion to the 95%-HIT-acceptance-rate already used for the relevance annotation, each test person had to have at least 20 HITs accepted as correct. Also, to lower the probability of one person introducing personal bias to the judgements, each test person could complete at most 25 items. Furthermore, a simple spam detection was carried out measuring the intra-rater variance (i.e., "Did a test person always vote for A?") and deviation from the mean time to task completion. Likely due to the high test person standards, no indication of systematic spam was found.

In total, 2797 HITs were carried out. A reward of 0.08 \$ per HIT was given, amounting to 805.54 \$ total cost for the quality annotation (including MTurk fees). From the gathered data, argument quality scores were calculated using the extended Bradley-Terry model described in Chapter 4. Raw comparison data as well as processed score data are available.[5]

In order to derive meaningful ranking data from pairwise comparisons using the described model, it is important to pay special attention to the parameters used to derive such a ranking. As described in Chapter 4, the model depends on two fixed parameters: λ, the regularization parameter; τ, the difference threshold. The score value of the fixed dummy element could theoretically be seen as a parameter as well, but since the scores are z-normalized, this parameter would have no effect. For λ, a value of 0.1 was chosen, which has shown a good performance on larger item counts [12]. For τ, a relatively low value of 0.05 was chosen, since the actual number of ties in the collected comparisons is fairly low, hinting at test persons being able to differentiate small quality differences.

[5] `https://git.informatik.uni-leipzig.de/lg80beba/argument-quality-evaluation/`

Chapter 6

Results

This chapter provides a quantitative analysis of the collected data and gives a performance assessment of the selected retrieval models. Throughout, comparisons to the pilot study will be drawn. Section 6.1 presents general metrics on the distribution of key attributes. These attributes and their interactions are more closely explored in Section 6.2. Addressing the main objective of this work, the ranking evaluation is conducted in Section 6.3, providing a conclusive evaluation of the different performance of models. This is extended upon in Section 6.4, where the notion of general argumentative quality is explored. Dataset schemes for the collected data are included in Appendix A.2.

6.1 Distribution

In total, the collected corpus encompasses 1610 items. 1271 of those were flagged as arguments, the remaining 339 were not. The arguments are divided fairly evenly into Pro (675) and Con (596) on the respective search issue.

Figure 6.1 shows the distribution of scores for the three quality aspects. Non-arguments are not included in the distribution plots for quality, as they received no quality ratings to reduce the annotation effort. The distributions of quality appear similar to each other, with only some slight deviations near the mean. Some outliers are apparent. Concerning the distribution of relevance scores, the trend implied in the pilot study is continued: the distribution is skewed towards the higher end, thus hinting at a high number of relevant items included in the pooling.

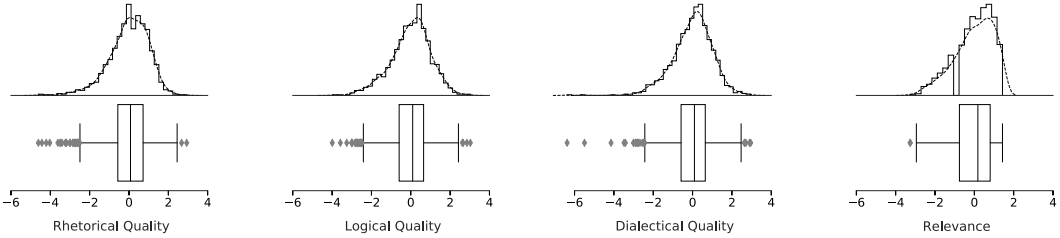

Figure 6.1: Quality aspect score distributions

	Relevance		Rhetorical		Logical		Dialectical	
Rhetorical	0.13				**0.63**		**0.61**	
Logical	0.10		**0.63**				0.55	
Dialectical	**0.15**		0.61		0.55			
	Pro	Con	Pro	Con	Pro	Con	Pro	Con
Rhetorical	0.14	0.12			**0.62**	**0.63**	**0.60**	**0.61**
Logical	0.11	0.09	**0.62**	**0.63**			0.60	0.51
Dialectical	**0.14**	**0.17**	0.60	0.61	0.60	0.51		
	Overall	$l > 100$	Overall	$l > 100$	Overall	$l > 100$	Overall	$l > 100$
Word Count	0.10	0.00	0.65	0.37	0.64	0.37	0.63	0.28

Table 6.1: Pearson ρ correlation cross-tabulation between quality aspects and relevance and per stance. $n_{\text{Pro}} = 675, n_{\text{Con}} = 596$

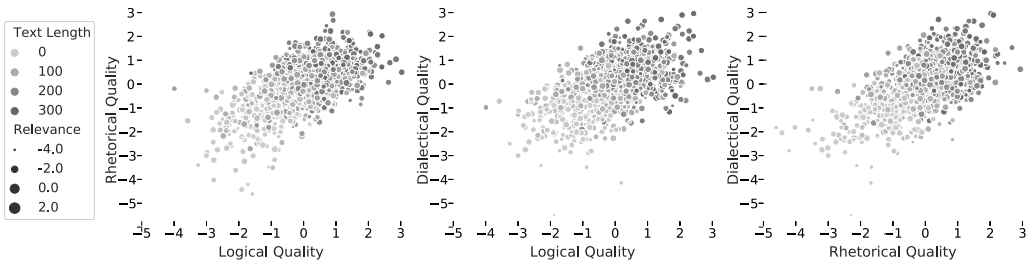

Figure 6.2: Scatterplots for the three quality aspects

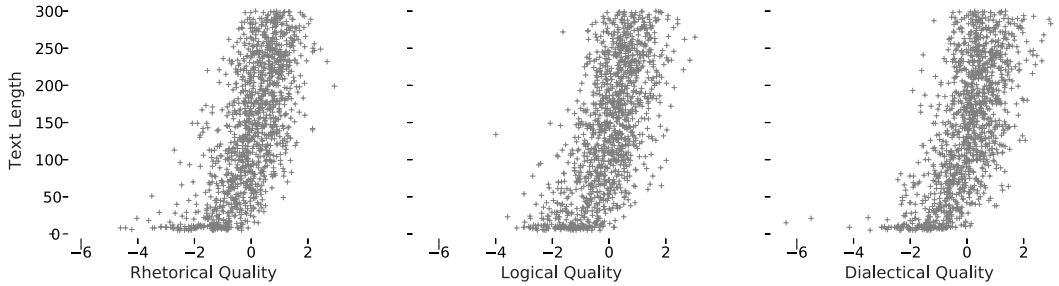

Figure 6.3: Scatterplot of Text Length and Text Quality

6.2 Correlation

Table 6.1 shows correlation coefficients between all three quality aspects and relevance. The inter-quality correlation appears similar to the pilot study. Although the maximum correlation per quality being different than before, the value differences are too small to draw any conclusions regarding whether two of the three are more intertwined than the third.

In contrast to the pilot study, the correlation between relevance and quality is not given anymore, possibly due to the increased topic depth. This effect substantiates the assumption that both relevance and quality are to some extent independent of each other and should both be taken into account when evaluating argument search engines, as no assumptions about the other can be derived from one of them. The correlation between the quality aspects being fairly high, hints at them being dependent on a latent variable, which could be the overall argumentation quality.

A correlation of quality and text length (measured as word count) is also apparent. While this could hint at a data bias, with test persons just voting for longer texts in the comparison but not actually reading all of it, the effect is much less pronounced when only measuring the correlation in texts longer than 100 words. Thus, much of the pronounced effect can be explained by short texts receiving justified low scores rather than longer texts being voted higher regardless of content. This effect is also apparent in Figure 6.3. Towards the lower end of the length spectrum, a cluster of short texts is visible, which mostly receive a low scoring. From a qualitative point of view, a correlation effect between length and quality would also be expected, since a solid argumentative reasoning (claim and justification) usually requires longer text lengths. Thus it can be concluded that the correlation is data-inherent and not due to an annotation bias.

6.3 Ranking Evaluation

To assess the ranking performance of the three chosen retrieval models, NDCG scores were calculated at $k = 50$ over 20 topics. The NDCG scores given are obtained on z-transformed score distributions as described in Section 4.5 and are to be interpreted as such. Confidence intervals were calculated using the bootstrap method (n =10 000). Mean z-NDCG scores and 95% confidence intervals are given in Table 6.2 and Figure 6.4.

Additional insight in comparison to the pilot study is apparent: even though the two models DirichletLM and DPH are still not separable with the desired statistical significance ($p_{\text{rhet}} = 0.0663$, $p_{\log} = 0.1781$, $p_{\text{dial}} = 0.1101$), a strong indication is given towards DirichletLM being the best performing model

	Rhetorical Quality			Logical Quality			Dialectical Quality			Relevance		
	Mean	95% CI		Mean	95% CI		Mean	95% CI		Mean	95% CI	
DPH	0.13	0.06	0.21	0.17	0.09	0.26	0.11	0.01	0.21	0.11	−0.07	0.27
DirichletLM	**0.18**	0.19	0.33	**0.27**	0.19	0.35	**0.24**	0.15	0.32	**0.18**	0.05	0.30
BM25F	−0.18	−0.24	−0.12	−0.17	−0.22	−0.11	−0.17	−0.24	−0.10	−0.23	−0.38	−0.10

Table 6.2: Mean z-NDCG scores and 95% confidence intervals for three retrieval models. Maximum per column marked bold.

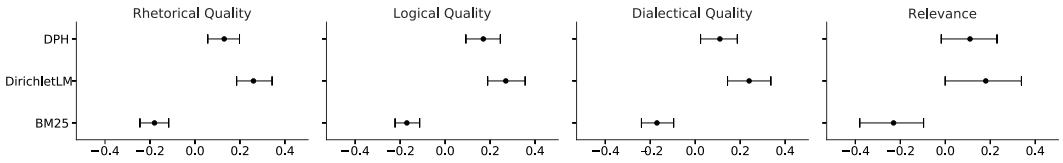

Figure 6.4: Mean z-NDCG scores and 95% confidence intervals for three retrieval models

across all three quality aspects as well as relevance. Both perform much better than a random ranking. However, given that the mean scores are around 0.22 (DirichletLM) and 0.13 (DPH), room for improvement is definitely given.

Regarding the performance of BM25F, the results of the pilot study are reproduced as well: it still ranks lowest among the compared models ($p < 0.01$ for all attributes and both models). Additionally, it can be derived that it performs worse than a random ranking, since its z-NDCG score is negative. While odd at first, this is likely due to BM25F favoring shorter texts ($\bar{l}_{\text{BM25F}} = 75.0, \bar{l}_{\text{DPH}} = 149.4, \bar{l}_{\text{DirichletLM}} = 165.6$). As argumentative quality correlates with text length, a random ranking would have a better performance, since the mean text length in the corpus is 123.5 words.

6.4 Combined Argument Quality

Even though it is argued that a general argument quality is hard to measure, the 3 different explored aspects could be combined to derive such a rating. The high correlation of the different quality aspects implies such a latent variable. As a working hypothesis, the overall argument quality could be interpreted as a 3-dimensional vector, with each of the quality aspects corresponding to a dimension. Based of this, two essential questions have to be explored: (1) Are the different aspects equally influential on the overall argument quality? (2) How can the overall quality be derived from such a vector?

Step	Variance	Rhetorical	Logical	Dialectical
1	0.73	−0.5866	−0.5715	−0.5738
2	0.15	0.1050	0.6489	−0.7536
3	0.12	−0.8031	0.5023	0.3206

(a) Component vectors and explained variance for PCA steps on argument quality

	z-NDCG	95% CI	
DPH	0.16	0.07	0.25
DirichletLM	0.31	0.23	0.38
BM25F	−0.19	−0.26	−0.13

(b) z-NDCG scores for combined quality

Table 6.3

To address the first question, Principal Component Analysis (PCA) was carried out to measure the influence of the aspects on the hypothesized latent variable. Results are given in Table 6.3a. The first step of the PCA accounts for 73% of the data variance, and is equally influenced by all three quality aspects. Therefore, evidence is given towards the hypothesis.

As for how to derive a numerical value for this overall argument quality, since the influence of all aspects is equal, the euclidean vector length is proposed. However, since the quality scores derived in this work are positive as well as negative, the length of a vector is the same as of its negative counterpart. To account for this, the distance is calculated to a negative point outside of the distribution range of all aspects instead of the origin[1] and the resulting vector lengths are then z-transformed again. The performance assessment was repeated on the combined quality scores to estimate which model would perform best when judged with equal attention to all three quality dimensions and to provide a scalar value that represents the quality performance of that model. Results are given in Table 6.3b.

On the combined scores, DirichletLM performs almost significantly better than the other two models ($p_{\mathsf{DPH}} = 0.053$, $p_{\mathsf{BM25F}} = 0$). The indication derived prior is therefore confirmed on combined quality scoring.

[1] This is equivalent to shifting the score distributions into the positive domain

Chapter 7

Conclusion

A general evaluation framework for argument retrieval has been proposed, adapting the classic TREC design to the domain of argument search. The new procedure incorporates different aspects of item relevance and item quality, thus allowing for an in-depth assessment of the search performance.

The existing approach utilizing topical relevance as evaluation parameter was found to be adequate for argument search, as it can be related to information needs stemming from argumentative theory. To expand the evaluation beyond relevance as performance criterion, a novel procedure for annotating argument quality has been developed, outperforming existing approaches in terms of annotation quality, annotation effort and annotation detail. The collected corpus is the largest accessible collection of arguments with quality annotations for different aspects available at the time. The annotation quality itself sets a new standard for future work. The collected data set can additionally be used for a multitude of purposes. Besides serving as basis for retrieval evaluation, it could be suitable to train new ranking models thus improving the retrieval performance of future engines. A second field of application is debate systems, where a dataset of quality-tagged arguments is of use for training system to formulate new arguments.

Also, the developed annotation procedure is not only limited to rate item quality: it can easily be transferred to any other question or criteria that can be rated by comparison. In information retrieval evaluation in particular, this approach could lead to a new notion of item relevance. Even though the annotation cost is higher compared to the classic absolute rating approach, even when employing the described optimization procedures, the derived data is much more detailed and allows for conclusions with higher statistical power.

Insight into argument quality was derived on a larger scale than in previous studies. The three major aspects have been shown to be adequate to capture the argumentative quality of a text and can be successfully annotated by lay-

men when using the described annotation procedure. The correlation patterns found in previous studies were reproduced, showing the aspects to be highly correlating with each other. This is likely due to them being dependent on a latent overall quality. An approach to derive a numerical value for this latent variable was explored as well.

Based on the proposed framework, a first insight into the retrieval performance of three retrieval models in the domain of argument search was gained. The performance was assessed using a new interpretation of the NDCG metric, adapted for the special data at hand. A clear performance ranking of the tested engines can be given: DirichletLM performs best, closely followed by DPH. BM25 consistently performs worse than a random ranking and its usage for the domain of argument retrieval is therefore discouraged. The engines perform similar in all three tested quality aspects, albeit consistently scoring highest when judged for *Logical Quality*. Overall, the evaluation can be regarded as successful.

Future research could include parameter optimization for the introduced scoring model, since the parameters chosen in this work are rather an educated guess than empirically derived. Improvements to the scoring model itself are also possible. Further work in other areas includes improving existing retrieval models using the collected data. The proposed z-NDCG metric should be formalized and further investigated in regards to its statistical properties. While DirichletLM shows promising performance for the taks of argument retrieval, room for improvement is given, warranting the development of specialized retrieval and/or ranking models for the domain.

Bibliography

[1] Alan Agresti. *Categorical data analysis*. 2nd ed. Hoboken, NY: John Wiley & Sons, 2003, pp. 436–439. DOI: 10.1002/0471249688.

[2] Gerald Albaum. "The Likert scale revisited". In: *Market Research Society. Journal.* 39.2 (1997), pp. 1–21. DOI: 10.1177/147078539703900202.

[3] Omar Alonso and Ricardo Baeza-Yates. "Design and Implementation of Relevance Assessments Using Crowdsourcing". In: *Proceedings of the 33rd European Conference on Advances in Information Retrieval - Volume 6611*. Ed. by Paul Clough et al. ECIR 2011. Dublin, Ireland: Springer-Verlag, 2011. DOI: 10.1007/978-3-642-20161-5_16.

[4] Omar Alonso and Stefano Mizzaro. "Using crowdsourcing for TREC relevance assessment". In: *Information Processing & Management* 48.6 (2012), pp. 1053–1066. DOI: 10.1016/j.ipm.2012.01.004.

[5] Giambattista Amati. "Frequentist and bayesian approach to Information Retrieval". In: *Advances in Information Retrieval*. Ed. by Mounia Lalmas et al. Berlin, Heidelberg: Springer, 2006, pp. 13–24. DOI: 10.1007/11735106_3.

[6] Nicholas J. Belkin, Michael Cole, and Ralf Bierig. "Is relevance the right criterion for evaluating interactive information retrieval?" In: *Proceedings of the ACM SIGIR 2008 Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments.* 2008.

[7] J Anthony Blair. "Rhetoric, dialectic, and logic as related to argument". In: *Philosophy & Rhetoric* 45.2 (2012), pp. 148–164.

[8] Ralph Allan Bradley and Milton E. Terry. "Rank analysis of incomplete block designs: The method of paired comparison". In: *Biometrika* 39.3-4 (1952), pp. 324–345. DOI: 10.1093/biomet/39.3-4.324.

[9] Chris Buckley and Ellen M Voorhees. "Retrieval evaluation with incomplete information". In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2004, pp. 25–32.

[10] Michael L Burton. "Too many questions? The uses of incomplete cyclic designs for paired comparisons". In: *Field Methods* 15.2 (2003), pp. 115–130. DOI: 10.1177/1525822X03015002001.

[11] Manuela Cattelan, Cristiano Varin, and David Firth. "Dynamic Bradley–Terry modelling of sports tournaments". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62.1 (2013), pp. 135–150. DOI: 10.1111/j.1467-9876.2012.01046.x.

[12] Xi Chen et al. "Pairwise ranking aggregation in a crowdsourced setting". In: *Proceedings of the sixth ACM international conference on web search and data mining*. WSDM '13. Rome, Italy: ACM, 2013, pp. 193–202. DOI: 10.1145/2433396.2433420.

[13] Cyril Cleverdon. "The Cranfield tests on index language devices". In: *Aslib proceedings*. Vol. 19. 6. MCB UP Ltd, 1967, pp. 173–194. DOI: 10.1108/eb050097.

[14] Roger R Davidson. "On extending the Bradley-Terry model to accommodate ties in paired comparison experiments". In: *Journal of the American Statistical Association* 65.329 (1970), pp. 317–328. DOI: 10.1080/01621459.1970.10481082.

[15] Franz Faul et al. "G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences". In: *Behavior research methods* 39.2 (2007), pp. 175–191. DOI: 10.3758/BF03193146.

[16] Lester R. Jr. Ford. "Solution of a ranking problem from binary comparisons". In: *The American Mathematical Monthly* 64.8P2 (1957), pp. 28–33. DOI: 10.2307/2308513.

[17] Michael Gordon and Praveen Pathak. "Finding information on the World Wide Web: the retrieval effectiveness of search engines". In: *Information Processing & Management* 35.2 (1999), pp. 141–180. DOI: 10.1016/S0306-4573(98)00041-7.

[18] Trudy Govier. *A practical study of argument*. 7th ed. Belmont, CF: Wadsworth, 2010.

[19] Jürgen Habermas. *The theory of communicative action*. Vol. 2. Boston: Beacon Press, 1984.

[20] Ivan Habernal and Iryna Gurevych. "Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1589–1599. URL: http://www.aclweb.org/anthology/P16-1150.

[21] Reinhold Hatzinger and Josef A Mazanec. "Measuring the part worth of the mode of transport in a trip package: An extended Bradley–Terry model for paired-comparison conjoint data". In: *Journal of business research* 60.12 (2007), pp. 1290–1302. DOI: 10.1016/j.jbusres.2007.04.010.

[22] Michael H. G. Hoffmann. "The Elusive Notion of "Argument Quality"". In: *Argumentation* 32.2 (2018), pp. 213–240. DOI: 10.1007/s10503-017-9442-x.

[23] Tzu-Kuo Huang, Chih-Jen Lin, and Ruby C. Weng. "Ranking individuals by group comparisons". In: *Journal of Machine Learning Research* 9.Oct (2008), pp. 2187–2216. DOI: 10.1145/1143844.1143898.

[24] David R. Hunter. "MM algorithms for generalized Bradley-Terry models". In: *The annals of statistics* 32.1 (2004), pp. 384–406. DOI: 10.1214/aos/1079120141.

[25] Sally Jackson. "Design Thinking in Argumentation Theory and Practice". In: *Argumentation* 29.3 (2015), pp. 243–263. DOI: 10.1007/s10503-015-9353-7.

[26] Kalervo Järvelin and Jaana Kekäläinen. "Cumulated gain-based evaluation of IR techniques". In: *ACM Transactions on Information Systems (TOIS)* 20.4 (2002), pp. 422–446. DOI: 10.1145/582415.582418.

[27] Ralph H. Johnson. *Manifest rationality: A pragmatic theory of argument*. London: Routledge, 2012.

[28] Ralph H. Johnson. "Revisiting the logical/dialectical/rhetorical triumvirate". In: *OSSA Conference Archive* 84 (2009).

[29] Evangelos Kanoulas and Javed A Aslam. "Empirical justification of the gain and discount function for nDCG". In: *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM. 2009, pp. 611–620. DOI: 10.1145/1645953.1646032.

[30] Klaus Krippendorff. "Testing the reliability of content analysis data". In: *The content analysis reader* (2009), pp. 350–357. DOI: 10.1111/j.1468-2958.2004.tb00738.x.

[31] Eddy Maddalena et al. "Considering assessor agreement in IR evaluation". In: *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM. Amsterdam, The Netherlands, 2017, pp. 75–82. DOI: 10.1145/3121050.3121060.

[32] Dima Mohammed. "Goals in argumentation: A proposal for the analysis and evaluation of public political arguments". In: *Argumentation* 30.3 (2016), pp. 221–245. DOI: 10.1007/s10503-015-9370-6.

[33] Jacqueline Murray. "Likert data: what to use, parametric or non-parametric?" In: *International Journal of Business and Social Science* 4.11 (2013).

[34] Geoff Norman. "Likert scales, levels of measurement and the "laws" of statistics". In: *Advances in health sciences education* 15.5 (2010), pp. 625–632. DOI: 10.1007/Fs10459-010-9222-y.

[35] Hiroshi Okamura, Masashi Kiyota, and Kazuhiko Hiramatsu. "Quantitative analysis of paired comparison data using the Bradley-Terry model with a normal distribution". In: *Japanese Journal of Biometrics* 21.2 (2001), 2_1–14. DOI: 10.5691/jjb.21.2_1.

[36] Iadh Ounis et al. "Terrier: A high performance and scalable information retrieval platform". In: *Proceedings of the OSIR Workshop*. 2006, pp. 18–25.

[37] Lawrence Page et al. *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab, 1999.

[38] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[39] ProCon.org. *FAQs (Frequently Asked Questions)*. https://www.procon.org/faqs.php. [Online; accessed 11-03-2019]. 2018.

[40] P.V. Rao and Lawrence L. Kupper. "Ties in paired-comparison experiments: A generalization of the Bradley-Terry model". In: *Journal of the American Statistical Association* 62.317 (1967), pp. 194–204. DOI: 10.1080/01621459.1967.10482901.

[41] Ruty Rinott et al. "Show me your evidence - An automatic method for context dependent evidence detection". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 440–450. URL: https://aclweb.org/anthology/D/D15/D15-1050.

[42] Horst W. J. Rittel and Melvin M. Webber. "Dilemmas in a general theory of planning". In: *Policy Sciences* 4.2 (1973), pp. 155–169. DOI: 10.1007/BF01405730.

[43] Stephen Robertson. "On the Contributions of Topics to System Evaluation". In: *Proceedings of the 33rd European Conference on Advances in Information Retrieval - Volume 6611*. Ed. by Paul Clough et al. ECIR 2011. Dublin, Ireland: Springer-Verlag, 2011, pp. 129–140. DOI: 10.1007/978-3-642-20161-5_14.

[44] Tetsuya Sakai. "Designing test collections for comparing many systems". In: *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*. ACM. 2014, pp. 61–70. DOI: 10.1145/2661829.2661893.

[45] Tetsuya Sakai. "Evaluating evaluation metrics based on the bootstrap". In: *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*. ACM. 2006, pp. 525–532. DOI: 10.1145/1148170.1148261.

[46] PC Sham and D Curtis. "An extended transmission/disequilibrium test (TDT) for multi-allele marker loci". In: *Annals of human genetics* 59.3 (1995), pp. 323–336. DOI: 10.1111/j.1469-1809.1995.tb00751.x.

[47] Rion Snow et al. "Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, 2008, pp. 254–263. URL: http://www.aclweb.org/anthology/D08-1027.

[48] Karen Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of documentation* 28.1 (1972), pp. 11–21. DOI: 10.1108/eb026526.

[49] Christian Stab et al. "Cross-topic Argument Mining from Heterogeneous Sources". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, 2018. URL: https://aclweb.org/anthology/D18-1402.

[50] Frans H Van Eemeren and Rob Grootendorst. *Speech acts in argumentative discussions: A theoretical model for the analysis of discussions directed towards solving conflicts of opinion*. Vol. 1. Berlin, New York: Walter de Gruyter, 2010.

[51] Ellen M. Voorhees. "The Philosophy of Information Retrieval Evaluation". In: *Evaluation of Cross-Language Information Retrieval Systems*. Ed. by Carol Peters et al. Berlin, Heidelberg: Springer, 2002, pp. 355–370. DOI: 10.1007/3-540-45691-0_34.

[52] Henning Wachsmuth et al. "Argumentation Quality Assessment: Theory vs. Practice". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 250–255. URL: https://aclweb.org/anthology/P17-2039.

[53] Henning Wachsmuth et al. "Building an Argument Search Engine for the Web". In: *Proceedings of the Fourth Workshop on Argument Mining (ArgMining 2017)*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 49–59. URL: https://aclweb.org/anthology/W17-5106.

[54] Henning Wachsmuth et al. "Computational Argumentation Quality Assessment in Natural Language". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*. 2017, pp. 176–187. URL: http://aclweb.org/anthology/E17-1017.

[55] Jun Wang. "Mean-Variance Analysis: A New Document Ranking Theory in Information Retrieval". In: *Advances in Information Retrieval*. Ed. by Mohand Boughanem et al. Berlin, Heidelberg: Springer, 2009, pp. 4–16. DOI: 10.1007/978-3-642-00958-7_4.

[56] William Webber et al. "Precision-at-ten Considered Redundant". In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. Singapore: ACM, 2008, pp. 695–696. DOI: 10.1145/1390334.1390456.

[57] Joseph W Wenzel. "Three perspectives on argument: Rhetoric, dialectic, logic". In: *Perspectives on argumentation: Essays in honor of Wayne Brockriede* (1990), pp. 9–26.

[58] Adam Wyner et al. "Approaches to Text Mining Arguments from Legal Cases". In: *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*. Ed. by Enrico Francesconi et al. Berlin, Heidelberg: Springer, 2010, pp. 60–79. DOI: 10.1007/978-3-642-12837-0_4.

[59] Ting Yan, Jinfeng Xu, and Yaning Yang. "Grouped sparse paired comparisons in the Bradley-Terry model". In: *arXiv preprint* (2011). URL: https://arxiv.org/abs/1111.5110v3.

[60] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. "A simple and efficient sampling method for estimating AP and NDCG". In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2008, pp. 603–610. DOI: 10.1145/1390334.1390437.

[61] Chengxiang Zhai and John Lafferty. "A study of smoothing methods for language models applied to ad hoc information retrieval". In: *ACM SIGIR Forum*. Vol. 51. 2. ACM. 2017, pp. 268–276. DOI: 10.1145/383952.384019.

Appendix A

Dataset Schemes

Each key on the left side of the table represents a column name with details about the contained data in the explanation field. Primary keys are marked in **bold**. If a combined key is used, all entries that the combined key is composed of are marked. Primary keys are separated from the rest of the table with a line. Foreign keys that can be used to reference other tables are marked in *italics*.

A.1 Pilot Study Dataset

Topic ID	Unique identifier for the topic context the item was judged for
Argument ID	Unique identifier for the item in regard to the discussion it was part of
Discussion ID	Unique identifier of the discussion the item was part of
Is Argument?	Boolean value, indicating wether the item is an argument, or not
Stance	Denoting the stance of the item; can be *Pro, Con* or *Not specified*
Relevance	Relevance score as judged by an annotator on a scale of 1 (not relevant) to 4 (very relevant)
Logical Quality	Logical quality score as judged by an annotator on a scale of 1 (very bad) to 4 (very good)
Rhetorical Quality	Rhetorical quality score as judged by an annotator on a scale of 1 (very bad) to 4 (very good)
Dialectical Quality	Dialectical quality score as judged by an annotator on a scale of 1 (very bad) to 4 (very good)
Premise	Text of the items premise
Conclusion	Text of the items conclusion
Comment	Optional comment made by the annotator

Table A.1: Data scheme for the argument dataset

Topic ID	ID of the topic judged by the annotator
Age	Age of the annotator
Gender	Gender of the annotator
Comment	Annotators' comments about the study

Table A.2: Data scheme for the annotator dataset

Topic ID	Unique identifier for the topic
Biased	Boolean value indicating whether the topic is biased or not, i.e., if the annotator was tasked to adopt a predetermined stance (topic thesis)
Annotator Stance	Annotator stance, can be 'I agree', 'I disagree' or 'Neutral'; if the topic is *biased*, the stance is determined in regard to the topic description; if the topic is *unbiased*, the stance is determined in regard to the topic thesis
Thesis	Predetermined stance for unbiased topics, empty otherwise
Description	Text description of the topic
Query	Query used for this topic as input for the retrieval models

u

Table A.3: Data scheme for the topic dataset

Topic ID	Unique identifier for the topic context
Engine	Name of the engine the ranking this entry stems from was created with
Rank	The rank of the argument in the respective engines ranking
Argument ID	Unique identifier for the argument in regards to the discussion it was part of
Discussion ID	Unique identifier of the discussion the argument was part of

Table A.4: Data scheme for the ranking dataset

A.2 Final Dataset

Topic ID	Unique identifier for the topic context the item was judged in
Argument ID	Unique identifier for the item in regards to the discussion it is part of
Discussion ID	Unique identifier of the discussion the item is part of
Is Argument?	Boolean value, indicating wether the item is an argument, or not
Stance	Denoting the stance of the item, can be *Pro, Con.* Mapped to boolean values, *True* for *Pro*, *False* for *Con*
Relevance	Relevance score, z-normalised
Logical Quality	Logical quality score, z-normalised
Rhetorical Quality	Rhetorical quality score, z-normalised
Dialectical Quality	Dialectical quality score, z-normalised
Combined Quality	Combined quality score, z-normalised
Premise	Text of the items' premise
Text Length	Word count of the premise

Table A.5: Data scheme for the argument dataset

Topic ID	Unique identifier for the topic
Category	Thematical category the topic belongs to
Long Query	Long query, used as input for the retrieval models
Short Query	Shortened form of the query

Table A.6: Data scheme for the topic dataset

Topic ID	Unique identifier for the topic context
Model	Name of the engine the ranking this entry stems from was obtained with
Rank	The rank of the argument in the respective engines ranking
Discussion ID	Unique identifier of the discussion the argument is part of
Argument ID	Unique identifier for the argument in regards to the discussion it is part of

Table A.7: Data scheme for the ranking dataset

Appendix B

Questionnaires

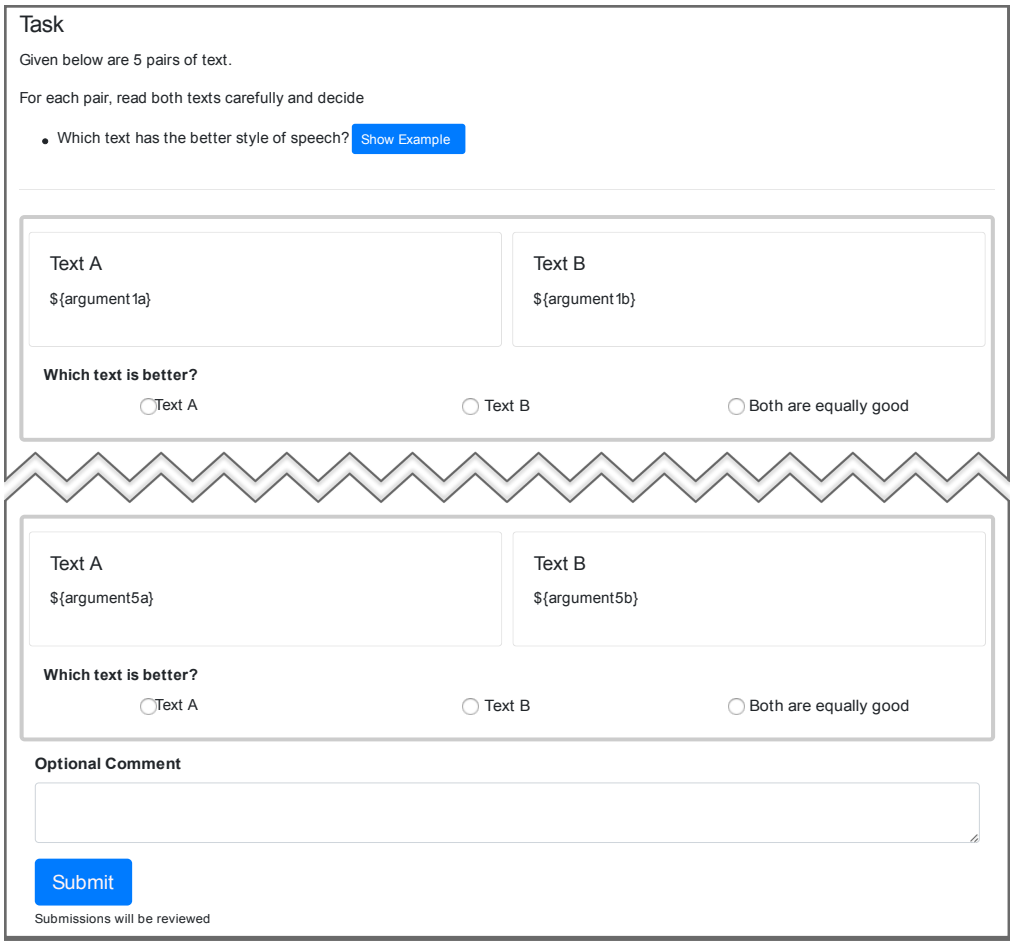

Figure B.1: Questionnaire layout used for quality annotation tasks on Amazon Mechanical Turk. Pairs 2 - 4 not shown.

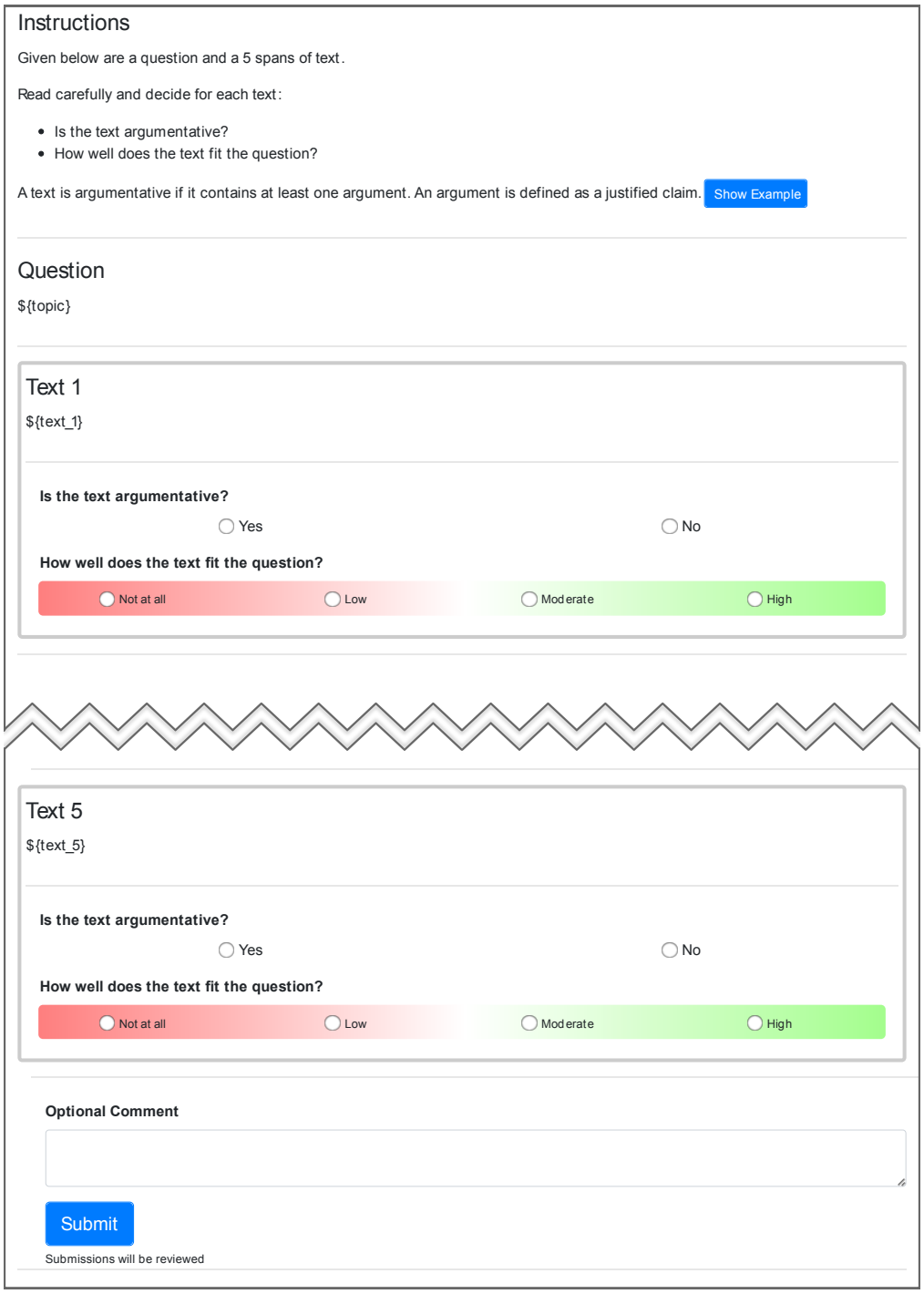

Figure B.2: Questionnaire layout used for relevance annotation tasks on Amazon Mechanical Turk. Text 2 - 4 not shown.

Appendix C

Query List

ID	Category	Query
1	Economy & Taxes	Should the Federal Minimum Wage Be Increased?
2	Economy & Taxes	Is a Universal Basic Income beneficial?
3	Economy & Taxes	Should Daylight Saving Time be kept?
4	Education	Should Students Have to Wear School Uniforms?
5	Education	Is the Use of Standardized Tests Improving Education?
6	Education	Should Corporal Punishment Be Used in Schools?
7	Elections	Do Electronic Voting Machines Improve the Voting Process?
8	Elections	Should Felons Who Have Completed Their Sentence Be Allowed to Vote?
9	Entertainment & Sports	Are Social Networking Sites Good for Our Society?
10	Health & Medicine	Should People Become Vegetarian?
11	Health & Medicine	Should Marijuana Be a Medical Option?
12	Health & Medicine	Should people have the Right to Health Care?
13	Politics	Should Adults Have the Right to Carry a Concealed Handgun?
14	Politics	Should the Death Penalty Be Allowed?
15	Science & Technology	Should Bottled Water Be Banned?
16	Science & Technology	Should police officers wear body cameras?
17	Science & Technology	Should Animals Be Used for Scientific or Commercial Testing?
18	Sex & Gender	Is Sexual Orientation Determined at Birth?
19	Sex & Gender	Should Gay Marriage Be Legal?
20	World & International	What Are the Solutions to the Israeli-Palestinian Conflict?

Table C.1: Topic Queries