Ranking Scientific Papers Using Preference Learning

Nils Dycke,1 Edwin Simpson,2 Ilia Kuznetsov,1 Iryna Gurevych1
1 Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt.
ukp.informatik.tu-darmstadt.de
2 Intelligent Systems Lab
University of Bristol
edwin.simpson@bristol.ac.uk

Abstract

Peer review is the main quality control mechanism in academia. Quality of scientific work has many dimensions; coupled with the subjective nature of the reviewing task, this makes final decision making based on the reviews and scores therein very difficult and time-consuming. To assist with this important task, we cast it as a paper ranking problem based on peer review texts and reviewer scores. We introduce a novel, multi-faceted generic evaluation framework for making final decisions based on peer reviews that takes into account effectiveness, efficiency and fairness of the evaluated system. We propose a novel approach to paper ranking based on Gaussian Process Preference Learning (GPPL) and evaluate it on peer review data from the ACL-2018 conference. Our experiments demonstrate the superiority of our GPPL-based approach over prior work, while highlighting the importance of using both texts and review scores for paper ranking during peer review aggregation.

Introduction

Peer review (PR) has a longstanding tradition as the prevalent quality assurance mechanism in science (Birukou et al. 2011). In the context of scholarly publication, the goal of PR is to select a fraction of best papers from the pool of submissions that exceed a certain quality threshold (Jefferson, Wager, and Davidoff 2002). During a classical PR process, a group of referees each assign one or multiple scores and a textual review to an allocated paper. Next, the program chairs weigh and aggregate the peer reviews for each paper to make a final acceptance decision. This process requires careful consideration of all aspects of paper quality including soundness, presentation, impact potential and rightfulness (Jefferson, Wager, and Davidoff 2002; Aksnes, Langfeldt, and Wouters 2019). The difficulty of this process is amplified by noise, bias and high disagreement in the peer reviews (Bakanic, McPhail, and Simon 1987; Lee et al. 2013; Walker and Rocha da Silva 2015). Moreover, the exponentially growing amount of research across a range of disciplines (Bornmann and Mutz 2015) call for new approaches that make PR more time-efficient. To alleviate acceptance decision making, the ranking by average review scores per paper is used in practice as one of its main inputs. However, due to the inconsistent usage of rating scales by the referees (Wang and Shah 2019; Lee 2015), average review scores provide an unreliable ranking to program chairs.

To combat the aforementioned issues, computer-aided assistance in PR has become more popular (Price and Flach 2017). Most works focus on the early editorial stages in the PR process, like the allocation of referees to papers (Wang, Chen, and Miao 2008). To support acceptance decision making, consensus ranking approaches have been applied (Cook et al. 2007; Baskin and Krishnamurthi 2009). They approximately solve the NP-hard task of finding a ranking of submissions that is maximally consistent with referees’ relative preferences. These approaches, however, neglect textual review information and are prone to noise in review scores. Related methods from the field of Natural Language Processing (NLP) consider textual information, like a paper’s text, to predict surrogate labels of paper quality, such as acceptance decisions (Kang et al. 2018) or citation counts (Li et al. 2019). As these approaches are trained on past conferences and do not consider shifted research focus over time, their transferability to novel works submitted to new PR processes may be limited.

In this paper, we propose to represent submissions based on both review scores and texts and rank them according to referees’ relative preferences on this space using preference learning. We investigate three hypotheses: First, a ranking model on submissions benefits from including peer review texts in addition to scores. Second, preferences expressed by human referees over the set of papers under review serve as a valuable supervision signal. According to this perspective, training requires no external paper quality estimates, like citation counts, which are typically not available during the time of review. Third, preference learning techniques can effectively mitigate the impact of noise, disagreement and bias in peer review data.

To formalize our preference-learning-based approach, we introduce a more general formulation of the review aggregation task, which we call the Paper Ranking Problem (PRP). To validate our hypotheses on real data, we derive a new generic evaluation framework from the science-of-science literature on peer review, which is based on past acceptance labels and citation counts. We select Gaussian process preference learning (GPPL) (Simpson and Gurevych 2018) to tackle the PRP. GPPL is a preference learning method
that has been applied to NLP tasks, such as ranking arguments by convincingness (Simpson and Gurevych 2018), and proved to be robust against noisy preference labels. We apply the aforementioned framework on the highly structured peer review data from the 2018 conference of the Association for Computational Linguistics (ACL) (Gao et al. 2019) to measure the performance of our method.

We show that our GPPL-based approach achieves the best balanced performance on both paper quality proxies compared to previous methods and baselines. During our ablation study, we find that review texts increase ranking performance by citation counts drastically, while review scores alone have small predictive validity for the future impact of accepted submissions. Finally, we find that our approach is less susceptible to noise in the review scores caused by unreliable referees and bias induced by heterogeneous weighing of quality criteria.

**Review Score Aggregation**

To solve the review aggregation task, Cook et al. (2007) propose the conversion from exact scores to partial rankings on a per-referee basis to mitigate the issue of miscalibration bias (Wang and Shah 2019) in review scores. As an example, if referee $A$ reviewed the papers $x$, $y$, and $z$ assigning them scores of 2, 1, and 3 (higher is better), respectively, this can be expressed by the ordering $\prec_A$ with: $y \prec_A x \prec_A z$. The authors cast the review aggregation task as a consensus ranking problem. Given the partial rankings on overlapping paper subsets, the goal is to find an output ranking on all papers that violates the smallest number of precedence pairs in the observed partial rankings. To solve this NP-hard problem the authors propose a branch-and-bound algorithm that can find optima on small artificial datasets in reasonable time. Similarly, Baskin and Krishnamurthi (2009) propose a more efficient neighborhood-based optimization algorithm for the same task and problem sizes.

Both works rely on the analytical optimization against partial rankings, which implies two inherent weaknesses on real-world data: First, they are susceptible to erroneous precedence pairs induced by review score noise, because there is no rational strategy for resolving contradictory partial rankings. Second, when the number of partial rankings and the overlap between their item sets is low, many supposedly optimal rankings exist from which an arbitrary one is chosen without a sensible strategy. During our experiments, we show that our approach outperforms both algorithms on a real-world dataset.

**Scholarly Document Quality Assessment**

We use scholarly document quality assessment (SDQA) as an umbrella term for various problem settings from NLP. All of them assess aspects of the quality of a scientific paper based on parts of the text or associated review texts. Kang et al. (2018) first introduced the task of paper acceptance prediction (PAP). Here papers from past conferences constitute the input. Their associated binary acceptance decisions are the target of prediction. This task has been approached in various flavors and learning setups: Shen et al. (2019) propose a model based on multi-modal embeddings of the paper and evaluate it on PAP and Wikipedia article quality assessment. Ghosal et al. (2019) use a multi-task learning setup based on review texts to jointly predict the review scores and acceptance decisions. Similarly, Stappen et al. (2020) predict acceptance decisions and review scores based on merged review texts, but they do not use a multi-task learning setup. Pytas, Rizos, and Specia (2021) apply SciBERT (Beltagy, Lo, and Cohan 2019) on review texts to predict paper acceptance and analyze how the reviewers’ vocabulary relates to the acceptance of a paper. Maillette de Buy Wenniger et al. (2020) combine PAP with the task of citation count prediction (CCP), where the target of prediction is the paper’s future citation count. The authors define an approach based on paper texts augmented with structural tags and apply it on both tasks in isolation. Similarly, Li et al. (2019) approach CCP using cross-attention between review texts and the paper abstract.

All of these models are trained on historic papers to have acceptance decisions or citation counts available. Most approaches rely on large corpora of scientific works from general domains regardless of their year of publication or topics. Several issues with the basic learning setup exist that hinder their applicability for assistance in a PR system. First, as models are trained on historic papers and make isolated predictions for each paper, they would need to learn very general, topic-independent concepts of paper quality. If not, they are inherently biased towards a historic state of the art leading to systematic conservatism. Second, scientific merit is context- and time-dependent (Lee 2015). A breakthrough work in one community might be insignificant for another – even within the same scientific domain. This makes binary acceptance decisions from different venues or different years a highly inconsistent proxy for paper quality.

All in all, existing strategies to review aggregation make substantial simplifying assumptions hindering their real-world applicability.

**Paper Ranking Problem**

In this section we introduce the Paper Ranking Problem, which models the task of aggregating reviews into a ranking of submissions to assist acceptance decision making.

**Problem Definition**

We define the PRP as the task of finding a ranking of the submissions according to their estimated quality. Unlike the problem settings in PAP or CCP, the quality estimate should be produced in comparison to the other submissions, hereby accounting for their scientific context and allowing program chairs to determine the acceptance threshold.

Let $P$ denote the set of papers submitted to the venue, where each paper $p \in P$ is represented by its text. For the set of reviews $R$ each review $r \in R$ is characterized
by its text and vector of scores. Furthermore, each review is associated with exactly one referee \( ref(r) \in E \) from the set of referees \( E \) and exactly one paper \( pap(r) \in P \). Hence, each referee \( e \in E \) generates a set of reviews \( R_e = \{ r \mid r \in R, ref(r) = e \} \) corresponding to a set of reviewed papers \( P_e = \{ pap(r) \mid r \in R_e \} \). The Paper Ranking Problem (PRP) is then defined as the task of predicting an overall ranking \( \mathcal{O}_P \) implying a total order on \( P \), given \( R \), \( P \) and \( E \) with all associated information. \( \mathcal{O}_P \) should have minimal ranking distance to the ranking of papers by their true total quality ordering \( \hat{\mathcal{O}}_P \).

We make two assumptions: First, there is exactly one true ranking by paper quality. We argue that this assumption is most plausible for the given context limited to a single venue. Due to the subjectivity of quality, this assumption might be partially violated. Second, the true quality ordering is total, implying that all papers are comparable to all others. Rogers and Augenstein (2020) raise reasonable doubt about the validity of this assumption, due to the heterogeneity of submissions. As we outline later, the PRP formulation can be projected to multiple target rankings on paper subsets, which weakens the restrictiveness of these assumptions.

**Performance Criteria**

A useful assistance system increases the efficiency of a PR process, while at least maintaining the quality of output decisions. Hence, quality criteria applied on PR in general need to be adopted to measure performance in the PRP. Inspired by science-of-science on PR, we derive the four performance criteria effectiveness, completeness, fairness and efficiency described below. For this purpose, we treat approaches to the PRP as random ranking models describing a probability distribution over all item permutations through decomposition into simpler probabilistic decisions (Critchlow, Fligner, and Verducci 1991). We consider pairwise precedences as these decisions. Here, the quality of \( p \), \( u(p) \), is drawn from the true quality distribution for paper \( p \in P \). For any \( x, y \in P \): \( x \preceq_M y \) is distributed according to the random ranking model \( \mathcal{M} \) and means that \( x \) precedes \( y \) in output rankings with the probability defined by \( \mathcal{M} \).

**Effectiveness and Completeness**

Effectiveness and completeness are classical criteria from information retrieval, typically measured by precision and recall. In the context of PR, effectiveness is often described by the predictive validity (Ragone et al. 2013) or the ability to filter out low quality works (Birukou et al. 2011). Lack of completeness is associated with low acceptance rates (Church 2005). If \( x \) is ranked higher than \( y \), effectiveness requires that the quality of \( x \) is higher than the one of \( y \).\[
\text{P}(u(x) > u(y) | x \preceq_M y) = 1
\]

For a complete ranking model \( x \) should always precede \( y \) in the ranking, given that \( x \) has a higher quality than \( y \).\[
\text{P}(x \preceq_M y | u(x) > u(y)) = 1
\]

**Fairness**

Fairness in PR is commonly regarded as the absence of bias in the reviews (Walker and Rocha da Silva 2015) or as the replicability of the PR process (Langford and Guzdial 2015). This perspective neglects the subjective nature of reviewing, which cannot be bias-free, and certain desirable types of bias in reviews, like a credit of trust towards potentially ground-breaking works (Bormann 2011). We propose an output-oriented criterion for fairness in the PRP: a ranking model is fair if papers of the same quality precede each other with the same probability.\[
P(x \gtrless_M y | u(x) = u(y)) = P(y \gtrless_M x | u(x) = u(y))
\]

This shifts the focus of bias analysis from the reviews to the ranking model and allows for swapped pairs in the generated output rankings, as long as they do not occur systematically.

**Efficiency**

Efficiency is a meta-criterion of the process of obtaining a ranking model. Ragone et al. (2013) measure the efficiency of PR by the time spent by reviewers to achieve a certain quality standard of acceptance decisions. We transfer this criterion directly to the PRP: the number of reviews required as an input should be minimal to produce an effective, complete, and fair ranking model.

**Evaluation Schema**

To evaluate approaches to the PRP in real domains, the latent true quality distribution per paper \( u(p) \) needs to be approximated. To account for the multi-faceted nature of paper quality, we propose to combine multiple weak quality indicators correlating with different aspects of quality.

Citation counts are common proxies for the impact potential of a paper, although they are noisy and uncertainty persists as to what they measure (Aksnes, Langfeldt, and Wouters 2019). Nevertheless, within a fixed scientific field and normalized by the time since publication (Mingers and Leydesdorff 2015) they can serve as weak indicators for paper impact. Furthermore, historic acceptance decisions relate to paper quality. They are, however, venue-specific, noisy and conservative in the sense that many high quality works are rejected due to restrictive acceptance quotas. While these measures should not be considered in isolation, a balanced, high performance on both of them reveals a certain consistency to paper quality within the limitations of the used proxies. In principle, the provided evaluation framework is applicable on any set of metrics correlating with diverse aspects of paper quality.

To practically quantify the effectiveness and completeness of a paper ranking model against a ranking by citation counts and binary acceptance decisions, we use Spearman’s rank correlation (ρ) and the area under the receiver operating characteristic (AUROC), respectively. As the probability distribution of precedence relations \( P(x \gtrless_M y) \) is typically not available, we approximate fairness by measuring the sensitivity of the PRP approach to bias and noise in the input data. When adding artificial bias and noise to the review scores, a fair approach should output rankings consistent to the unaltered and true one. The evaluation of efficiency follows directly from its definition. The amount of available reviews is artificially reduced by sub-sampling.
randomly. Then the performance decrease in terms of effectiveness and completeness compared to the full data set is measured. During experiments we apply this general evaluation schema on real data.

Gaussian Process Preference Learning for the Paper Ranking Problem

The Paper Ranking Problem can be naturally cast as a preference learning task, which not only ranks papers according to referees’ scores, but also incorporates review texts.

Preference-based Model of the PRP

Preference learning requires relative preferences on the space of items as an input, which we elicit as for consensus ranking, by converting the review scores of a referee \( e \in E \) into the partial ranking on the set of reviewed papers \( R_e \). This makes a latent assumption: The referees compare papers during reviewing consciously or subconsciously. While reported effects like order bias (Birukou et al. 2011) suggest that referees do not make isolated decisions for each paper, this assumption might be violated where papers on different topics or of diverse methods are reviewed by the same referee. However, this assumption is also implied when using review scores directly, as the scores of incomparable papers are mapped to the same numeric space. In fact, the preference-based view is less restrictive, as it allows for explicit filtering of potentially invalid comparisons.

Formally, the set of partial orders \( PO_E = \{ O_{P_e} | e \in E \} \) is given as training supervision. Here \( O_{P_e} \) refers to the total order on papers \( P_e \) reviewed by \( e \), which is induced by the order on associated scores. The training samples are the set of all papers, where each paper \( p \in P \) is characterized by a feature vector derived from the review texts and score vectors. The goal is to predict \( O_{P} \) assuming that the observed partial rankings are sampled from the true order \( O_{P} \) with random permutations. Each partial ranking implies a set of preference pairs including tie preferences if the partial orders are not strict. Consequently, the model orders only items seen during training in this learning setup.

To describe the sparsity of the supervision signal the preference graph (Domshlak et al. 2011) is a useful tool. In the context of the PRP, the nodes of the preference graph represent the papers and the multi-set of pairwise preferences induced by the partial orders \( PO_E \) are the directed edges of the graph. If the graph is highly connected, the supervision signal is rich and preference learning is applicable.

GPPL for Ranking Scientific Papers

Gaussian processes (GPs) are fully Bayesian regression learners relying on Gaussian priors (Rasmussen and Williams 2006). They are typically more robust against noise and perform well in small data domains. Chu and Ghahramani (2005) adapt Gaussian processes for preference learning. In classical GP-based regression the observed real-numbered variable in the training data is assumed to be distributed according to a likelihood function depending on the samples’ true, but hidden target values plus Gaussian-distributed noise. For GPPL, on the other hand, the observations are preference pairs \( x \succ y \) meaning that sample \( x \) is preferred over \( y \). The observable preference pairs are assumed to follow a likelihood function where the probability of a preference \( x \succ y \) depends on the difference between the hidden quality values of \( x \) and \( y \) with additive Gaussian noise. GPPL predicts the quality \( u(x) \) of a sample given the training pairs by marginalizing over the latent variables.

Simpson and Gurevych (2018) propose an approximation for the posterior using stochastic variational inference (Hoffman et al. 2013). The resulting method is more scalable with respect to the number of training samples and preference pairs than previous works. The authors show that scalable GPPL can be effectively applied in the domain of NLP to rank arguments based on sentence embeddings and linguistic features. Scalable GPPL is suitable for the PR domain given that review scores can be noisy and the number of submissions range from below hundred to thousands of submissions depending on the venue. Hence, we utilize the method by Simpson and Gurevych (2018) in our system. To apply GPPL for the PRP, the partial orders \( PO_E \) have to be converted into a set of preference pairs. We enumerate all implied precedence pairs for each \( O_{P_e} \in PO_E \) and join these precedence pairs of all referees into a multi-set as an input. The papers are represented by feature vectors described in the following section. The output of a GPPL model are real-valued quality estimates allowing for the total ranking on the set of all submissions.

Feature Set Design

Although paper texts could be used as an input to feature computation, we focus on review scores and texts to derive a representation of the submissions. This does not only allow us to investigate the importance of review texts for the PRP, but also enforces that primarily the human judgements of referees are reflected in the resulting ranking. We consider three feature types to define a vector representation, such that two papers have a small distance in the vector space if they are of similar true quality.

- **Score-based Features**: These features are derived from the reviewers’ score vectors. Apart from an overall assessment, typically ratings on aspects of paper quality, like soundness or presentation, are provided as aspect scores. For each aspect score and the overall score, the mean, standard deviation, minimum and maximum over the reviews per paper are computed. Additionally, the score vectors of each review of a paper are concatenated in arbitrary order. Score-based features might reflect controversies and multiple quality aspects, but they are expected to be noisy, as they rely on the scores directly.

- **Discourse Features**: Peer reviews contain questions, feedback and summarizing statements. The distribution of these argumentative elements might be an indicator for the soundness of a paper, as they relate to its strengths
and weaknesses. The AMPERE dataset (Hua et al. 2019) consist of 400 computer science reviews annotated with sentence-level discourse labels, including request and non-argumentative. We fine-tune the last layer of a BERT (Devlin et al. 2019) model for 4 epochs on a random subset of 90% of samples in AMPERE resulting in a 0.7969 micro-F1-score on the remaining 10% of the data. This model is applied on each review sentence. The distribution of discourse labels and the proportion of argumentative sentences across all reviews are added as a feature.

- **Embedding-based Features**: Embeddings are widely used in NLP to capture the similarity and meaning of texts. Reviews are structured into sections answering different questions of the review form. We encode each section using mean pooling on sentence-embeddings and concatenate them. Additionally, we form the mean of embeddings per section across reviews. To capture review-relatedness we compute the average cosine similarity of the first sentence per review. To embed sentences, we use distilled SBERT (Reimers and Gurevych 2019) fine-tuned on the natural language inference task, as these embeddings are not domain-specific and suitable to represent general statements in reviews. In this work, we focus on a simple representation of review texts as a proof of concept. We leave sophisticated embedding strategies on paragraph-level, like Longformer (Beltagy, Peters, and Cohan 2020), or domain-specific embeddings, like SciBERT (Beltagy, Lo, and Cohan 2019), for future work.

During experiments, we investigate feature subsets to determine how useful they are with respect to different proxies of paper quality. Additionally, we tested simple reading complexity metrics, like Flesch reading ease (Flesch 1948), but they did not improve performance in any scenario.

### Experiments

For our experiments, we use the anonymized PR data from consenting referees at ACL-2018 kindly provided by the collecting parties (Gao et al. 2019). This dataset contains anonymous referee identifiers, as in real PR systems, which we can use to infer partial rankings. To our knowledge, ACL-2018 is the only available dataset with this information. We also investigated data from the open, post-publication PR service F1000 Research for our work, but the used reviewing scheme does not employ review scores. Nevertheless, we presume that the availability of more complete PR data is likely in the future, due to an increased interest in open peer review (Birukou et al. 2011).

#### Dataset

The ACL-2018 dataset includes review data and acceptance decisions on 1538 submissions. We only consider before-rebuttal reviews to avoid social biases (Gao et al. 2019). The acceptance rate lies at roughly 25%. Each paper is assigned to one of 21 tracks. Each review consists of an overall score on a six-point scale and the six aspect scores originality, soundness, substance, replicability, meaningful comparison and readability on a five-point scale. Additionally, each review is structured into five sections including summary and contributions, strengths, weaknesses, questions and additional comments. Relevant statistics for the ACL-2018 dataset are summarized in Table 1. We observe that the preference graph is quite dense implying roughly 6 comparisons per paper. Treating PR as an annotation study, the agreement on overall scores reveals the level of consistency between referees’ judgements. For ACL-2018, agreement is low compared to other computer science venues (Kagone et al. 2013), but high compared to social science journals (Bormann 2011).

To realize the proposed evaluation strategy using both citation counts and acceptance labels as gold standards, the subset of accepted papers in ACL-2018 was matched with the NLPScholar dataset (Mohammad 2020). The hereby acquired citation counts are converted into a ranking. We omit time-wise normalization, as the papers have identical age. As an additional reference, we normalize citation counts for each track, which eliminates a preference towards topics with a broader audience. We normalize by the sum of citation counts \( cc(p) \) of papers in track \( t \) to form a ranking: \( ncc_t(p) = \sum_{q \in t} \frac{cc(q)}{cc_t} \). Due to privacy restrictions, the citation counts of rejected papers cannot be considered.

#### Hyper-parameter Tuning and Training

We employ the GPPL implementation by Simpson and Gurevych (2018) in its default configuration with a Matérn kernel. Instead of length-scale optimization or the proposed median-heuristic we use standard normalization on the non-embedding features, because this increased performance on all gold standards substantially. During experiments, no other hyperparameter configuration showed a performance increase. Hence, we only report on variations of the feature sets. The goal of the PRP is to predict the true quality ranking of papers seen during training. For all experiments we therefore train the GPPL model on the complete set of input papers and reviews, but make predictions for subsets of this data. We split the ACL-2018 data into a 20% development set for our work, but the used reviewing scheme does not employ review scores. Nevertheless, we presume that the availability of more complete PR data is likely in the future, due to an increased interest in open peer review (Birukou et al. 2011).
Baselines  We consider naïve score aggregation strategies including the mean overall score weighted by review confidence score (MEAN-S-w), median overall score (MEDIAN-S) and majority voting on overall scores (MAJOR-S) falling back to the mean score for tied votes. The MEAN-S-w weights based on referee-provided confidence scores, as this improved performance on the development set. Additionally, we compare to the decision-based (Cook et al. 2007) (DECI-CONS) and the neighborhood-based (Baskin and Krishnamurthi 2009) (NEIGH-CONS) consensus rankers. Both algorithms are re-implemented in Python based on the code kindly provided by the authors. The consensus rankers receive the same pairs as the GPPL model excluding tie preferences, because they cannot account for them.

Experimental Scenarios  The first set of experiments serves to investigate the performance of different feature sets. We optimize two feature constellations, where the first is selected only based on acceptance labels (by AUROC) and the second one is selected based on the unnormalized citation-based ranking (using Spearman’s ρ) for the development set. Hereby, we identify which features boost performance on the different gold standards. In the end, we select the one with the best balanced performance on both gold standards for all further experiments.

The evaluation of effectiveness and completeness is directly given by the evaluation strategy described earlier. We measure the performance of the baselines and best model on the test set of the ACL-2018 dataset. To judge the fairness of the ranking models, we consider two scenarios of rating errors and measure their impact on performance: First, we add random noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma \in \{0.75, 1.0\}$ to the aspect and overall scores (rounded to full scores) of $\alpha \in \{30\%, 60\%\}$ of the referees. Hereby, we simulate the effect of unreliable referees. Commensuration bias refers to the heterogeneous weighting of paper aspects by different referees to derive the overall score of a paper (Lee 2015). We simulate this by replacing the overall score by a weighted sum of the aspect scores and adding low normal noise $\sigma = 0.5$. We apply this for the reviews of $\alpha = 30\%$ of the referees. To analyze different scenarios, we use equal weights (COMM-EQ), an over-emphasis on readability (COMM-READ) and discarding of the originality score (COMM-CON). To evaluate the efficiency of the PRP approaches we sub-sample the reviews per paper discarding $\alpha \in \{30\%, 60\%\}$ of the reviews while ensuring that there is at least one review per paper. Again we measure the performance decrease on both gold standards.

Results  Our Python implementation of DECI-CONS is stopped early after 20h of computation on a machine with 8 CPUs and 16Gb of RAM. NEIGH-CONS on average converges in 11h. Our GPPL model using the final pre-computed feature set terminates on average in 4.5min on the same machine.

Feature Selection  The best features on acceptance labels include score- and embedding-based features, but no discourse information. This configuration achieves 0.8463 AUROC and $\rho = 0.1930$ on citation counts. The performance drops to 0.7687 AUROC without score-based features. Not surprisingly, review scores are central to predict acceptance labels, because they strongly influence acceptance decisions. To rank consistently with citation counts, embedding-based features are crucial. Here, the best features include only the embeddings of the reviews’ summary and contributions sections and discourse features. This achieves $\rho = 0.2952$ and 0.6948 AUROC. Discarding embedding-based features leads to a drastic drop of $\rho$ by 0.1049. This is reasonable, as impact is strongly linked to the contributions of a paper. All further experiments use the features optimized on acceptance labels, because they offer the best performance trade-off on both gold standards.

Effectiveness and Completeness  In Table 2 the effectiveness and completeness of our approach is compared to the baselines. As an ablation study, we report the performance using only embedding-based and only score features. While MEAN-S-w performs best on acceptance labels according to AUROC, the difference to our model is close to zero ($-0.0099$). At the same time the performance gain of our model according to $\rho$ on the raw ($+0.0805$) and normalized ($+0.0621$) citation rankings is substantial. The consensus ranking baselines perform consistently worse. This shows the limitations of other methods on real peer review data. The GPPL models using subsets of the best features confirm the importance of scores for acceptance labels and of the embedding-based features for citation counts. This also shows that our approach can be tuned to the desired trade-off on these gold standards. Finally, the equally high performance on track-wise normalized citation counts hints that our approach does not simply learn to favor topics with a broader audience, as this effect is mitigated in this ranking.

Fairness  The results for simulating unreliable referees are consistent for added noise at levels $\sigma = 0.75$ and $\sigma = 1.0$. Hence, we only report on the setting with $\sigma = 1.0$ on varying rates of unreliable referees. For all algorithms except the consensus rankers, the ranking produced on the noisy reviews is mostly consistent with their original rankings ($\rho > 0.9$). In Table 3 the AUROC and Spearman’s $\rho$ on the test set are given for different proportions of affected referees. It becomes visible that GPPL shows the smallest decay in performance for acceptance labels. On the citation-count-based ranking, the performance of GPPL remains the highest, while the other baselines show a drastic drop for 60% of noisy referees. Although adding noise only to scores favors approaches that rely on review texts, this suggests that the GPPL model is less affected by additional score noise.
For the three scenarios of commensuration bias, the performance of the best baselines and our approach is reported in Table 3. The GPPL model outperforms all other methods in all scenarios. Surprisingly, the COMM-EQ scenario leads to an improved performance on the citation-based ranking for nearly all algorithms. Apparently, substituting the actual overall score by the average of aspect scores acts as a de-biasing approach. This supports the idea of pre-processing data samples and preference pairs to further improve the performance of our approach.

**Efficiency** In the first scenario of efficiency evaluation ($\alpha = 30\%$) on average 1.60 reviews per paper and 1.93 reviews per referee are left. For $\alpha = 60\%$ of removed reviews, 1.01 reviews per paper and 1.23 reviews per referee are given. The consensus ranking algorithms are not applicable in both scenarios, because not all papers are included in a partial ranking of more than one paper. Likewise, the GPPL model has to make predictions on submissions not seen during training. As shown in Figure 1, the performance of all algorithms drops drastically. Our model deals slightly better with sparsity of reviews than the MEAN-S-w. The reduction of the number of reviews to increase efficiency appears, however, contradictory to output quality for all approaches.

**Conclusion and Future Work**

In this paper, we defined the Paper Ranking Problem together with a generic evaluation framework formulating the task of review aggregation more realistically than existing work. We derived a preference-oriented view and specified a ranking model based on GPPL. We empirically showed that it offers the best balanced performance on acceptance labels and citation counts. Additionally, we showed that it is more robust against unreliable referees and added commensuration bias. We hereby answered the question on the validity of the preference model of peer review positively and showed the importance of review texts in particular for ranking consistently to citation counts. We think that specialized embeddings of review texts and the combination with paper embeddings proposed in scholarly document quality assessment are promising future directions of research. Additionally, intensive research on the transferability of our method to different PR systems is required.
References

Aksnes, D. W.; Langfeldt, L.; and Wouters, P. 2019. Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. SAGE Open, 9(1): 215824401982957.

Bakanic, V.; McPhail, C.; and Simon, R. J. 1987. The manuscript review and decision-making process. American Sociological Review, 631–642. Publisher: JSTOR.

Baskin, J. P.; and Krishnamurthi, S. 2009. Preference aggregation in group recommender systems for committee decision-making. In Proceedings of the third ACM conference on Recommender systems - RecSys ’09, 337. New York, New York, USA: ACM Press. ISBN 978-1-60558-435-5.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pre-trained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3615–3620. Hong Kong, China: Association for Computational Linguistics.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150.

Birukou, A.; Wakeling, J. R.; Bartolini, C.; Casati, F.; Marchese, M.; Mirylenka, K.; Osman, N.; Ragone, A.; Sierra, C.; and Wassef, A. 2011. Alternatives to Peer Review: Novel Approaches for Research Evaluation. Frontiers in Computational Neuroscience, 5.

Bornmann, L. 2011. Scientific peer review. Annual Review of Information Science and Technology, 45(1): 197–245.

Bornmann, L.; and Mutz, R. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology, 66(11): 2215–2222. Publisher: Wiley Online Library.

Chu, W.; and Ghahramani, Z. 2005. Preference learning with Gaussian processes. In Proceedings of the 22nd international conference on Machine learning: 137–144.

Church, K. 2005. Reviewing the reviewers. Computational Linguistics, 31(4): 575–578. Publisher: MIT Press.

Cook, W. D.; Golany, B.; Penn, M.; and Raviv, T. 2007. Creating a consensus ranking of proposals from reviewers’ partial ordinal rankings. Computers & Operations Research, 34(4): 954–965.

Critchlow, D. E.; Fligner, M. A.; and Verducci, J. S. 1991. Probability models on rankings. Journal of mathematical psychology, 35(3): 294–318. Publisher: Elsevier.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Domshlak, C.; Hullermeier, E.; Kaci, S.; and Prade, H. 2011. Preferences in AI: An overview. Artificial Intelligence, 175(7-8): 1037–1052.

Flesch, R. 1948. A new readability yardstick. Journal of applied psychology, 32(3): 221. Publisher: American Psychological Association.

Fytas, P.; Rizos, G.; and Specia, L. 2021. What Makes a Scientific Paper be Accepted for Publication? arXiv:2104.07112.

Gao, Y.; Eger, S.; Kuznetsov, I.; Gurevych, I.; and Miyao, Y. 2019. Does My Rebuttal Matter? Insights from a Major NLP Conference. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 1274–1290. Minneapolis, Minnesota: Association for Computational Linguistics.

Ghosal, T.; Verma, R.; Ekbal, A.; and Bhattacharyya, P. 2019. DeepSentiPeer: Harnessing Sentiment in Review Texts to Recommend Peer Review Decisions. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1120–1130. Florence, Italy: Association for Computational Linguistics.

Hoffman, M. D.; Blei, D. M.; Wang, C.; and Paisley, J. 2013. Stochastic variational inference. Journal of Machine Learning Research, 14(S).

Hua, X.; Nikolov, M.; Badugu, N.; and Wang, L. 2019. Argument Mining for Understanding Peer Reviews. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2131–2137. Minneapolis, Minnesota: Association for Computational Linguistics.

Jefferson, T.; Wager, E.; and Davidoff, F. 2002. Measuring the quality of editorial peer review. Jama, 287(21): 2786–2790. Publisher: American Medical Association.

Kang, D.; Ammar, W.; Dalvi, B.; van Zuylen, M.; Kohlmeier, S.; Hovy, E.; and Schwartz, R. 2018. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 1647–1661. New Orleans, Louisiana: Association for Computational Linguistics.

Krippendorff, K. 1980. Content Analysis: An Introduction To Its Methodology. Beverly Hills: Sage Publications.

Langford, J.; and Guzdial, M. 2015. The arbitrariness of reviews, and advice for school administrators. ACM New York, NY, USA.

Lee, C. J. 2015. Commensuration Bias in Peer Review. Philosophy of Science, 82(5): 1272–1283.

Lee, C. J.; Sugimoto, C. R.; Zhang, G.; and Cronin, B. 2013. Bias in peer review. Journal of the American Society for Information Science and Technology, 64(1): 2–17.

Li, S.; Zhao, W. X.; Yin, E. J.; and Wen, J.-R. 2019. A Neural Citation Count Prediction Model based on Peer Review Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 4913–4923. Hong Kong, China: Association for Computational Linguistics.
Maillette de Buy Wenniger, G.; van Dongen, T.; Aedmaa, E.; Kruitbosch, H. T.; Valentijn, E. A.; and Schomaker, L. 2020. Structure-Tags Improve Text Classification for Scholarly Document Quality Prediction. In Proceedings of the First Workshop on Scholarly Document Processing, 158–167. Online: Association for Computational Linguistics.

McGraw, K. O.; and Wong, S. P. 1996. Forming inferences about some intraclass correlation coefficients. Psychological methods, 1(1): 30.

Mingers, J.; and Leydesdorff, L. 2015. A review of theory and practice in scientometrics. European Journal of Operational Research, 246(1): 1–19.

Mohammad, S. M. 2020. Examining Citations of Natural Language Processing Literature. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5199–5209. Online: Association for Computational Linguistics.

Price, S.; and Flach, P. A. 2017. Computational support for academic peer review: A perspective from artificial intelligence. Communications of the ACM, 60(3): 70–79.

Ragone, A.; Mirylenka, K.; Casati, F.; and Marchese, M. 2013. On peer review in computer science: analysis of its effectiveness and suggestions for improvement. Scientometrics, 97(2): 317–356.

Rasmussen, C. E.; and Williams, C. K. I. 2006. Gaussian processes for machine learning. Adaptive computation and machine learning. Cambridge, Mass: MIT Press. ISBN 978-0-262-18253-9. OCLC: ocm61285753.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Rogers, A.; and Augenstein, I. 2020. What Can We Do to Improve Peer Review in NLP? In Findings of the Association for Computational Linguistics: EMNLP 2020, 1256–1262. Online: Association for Computational Linguistics.

Shen, A.; Salehi, B.; Baldwin, T.; and Qi, J. 2019. A joint model for multimodal document quality assessment. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 107–110. IEEE.

Simpson, E.; and Gurevych, I. 2018. Finding convincing arguments using scalable bayesian preference learning. Transactions of the Association for Computational Linguistics, 6: 357–371. Publisher: MIT Press.

Stappen, L.; Rizos, G.; Hasan, M.; Hain, T.; and Schuller, B. W. 2020. Uncertainty-Aware Machine Support for Paper Reviewing on the Interspeech 2019 Submission Corpus. In Interspeech 2020, 1808–1812. ISCA.

Walker, R.; and Rocha da Silva, P. 2015. Emerging trends in peer review—a survey. Frontiers in Neuroscience, 9.

Wang, F.; Chen, B.; and Miao, Z. 2008. A Survey on Reviewer Assignment Problem. In Nguyen, N. T.; Borzemski, L.; Grzech, A.; and Ali, M., eds., New Frontiers in Applied Artificial Intelligence, Lecture Notes in Computer Science, 718–727. Berlin, Heidelberg: Springer. ISBN 978-3-540-69052-8.

Wang, J.; and Shah, N. B. 2019. Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings. In Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems, volume 18. Montréal, Canada: International Foundation for Autonomous Agents and Multiagent Systems.