Assisting Decision Making in Scholarly Peer Review: A Preference Learning Perspective

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 primary means of quality control in academia; as an outcome of a peer review process, program and area chairs make acceptance decisions for each paper based on the review reports and scores they received. Quality of scientific work is multi-faceted; coupled with the subjectivity of reviewing, this makes final decision making difficult and time-consuming. To support this final step of peer review, we formalize it as a paper ranking problem. We introduce a novel, multi-faceted generic evaluation framework for ranking submissions based on peer reviews that takes into account effectiveness, efficiency and fairness. We propose a preference learning perspective on the task that considers both review texts and scores to alleviate the inevitable bias and noise in reviews. Our experiments on peer review data from the ACL 2018 conference demonstrate the superiority of our preference-learning-based approach over baselines and prior work, while highlighting the importance of using both review texts and scores to rank submissions.

1 Introduction

Peer review (PR) is the prevalent quality assurance mechanism in science [Birukou et al., 2011]. During peer review, multiple referees evaluate a paper providing scores and a textual report. Next, the program or area chairs weigh the reviews of all papers to make final acceptance decisions selecting the top papers exceeding a quality threshold [Jefferson et al., 2002]. This requires careful comparison of papers based on the reviewers’ assessment of soundness, presentation, impact potential and rightfulness [Jefferson et al., 2002; Aksnes et al., 2019]. The difficulty of this process is amplified by noise, bias and high disagreement in the peer reviews [Lee et al., 2013; Walker and Rocha da Silva, 2015] and the ever-growing submission numbers in academia.

To cope with the complexity of this task, decision makers settle for heuristics to guide and prioritize the assessment process. Typically, statistics on review scores, like the mean overall score per paper, are used to rank papers as an input. Hereby, decision makers can focus on borderline regions close to the quality threshold. This heuristic is insufficient due to inconsistent usages of rating scales by the referees [Wang and Shah, 2019; Lee, 2015] and the aforementioned issues of noise. Consequently, agreement on absolute scores is considerably below substantial (in computer science) [Ragone et al., 2013] making the mean on typically three to four peer reviews per paper highly unreliable.

Peer review can gain in accuracy and efficiency when providing a reliable input ranking to the complex and time-consuming decision making task. In the past, so-called consensus ranking has been applied to rank submissions [Cook et al., 2007; Baskin and Krishnamurthi, 2009]; yet these approaches neglect textual review information and are prone to noise in review scores. Related methods from the field of Natural Language Processing (NLP) use textual information, like a paper’s abstract, to predict binary acceptance decisions [Kang et al., 2018] or citation counts [Li et al., 2019]. The output of these methods is coarse-grained or relies on information that is not available during the time of review restricting their applicability to assist acceptance decision making.

In this paper, we propose preference learning to rank submissions represented by their review scores and texts using referees’ preferences as supervision on this feature space. According to this perspective, training requires no external paper quality estimates, like citation counts, which are typically not available during the time of review. We investigate three hypotheses: First, preferences expressed by human referees over the set of papers under review serve as a valuable supervision signal. Second, a ranking model on submissions benefits from including peer review texts in addition to scores. Third, preference learning techniques can effectively mitigate the impact of noise, disagreement and bias in peer review data. To formalize our new preference learning perspective, we introduce a more general formulation of the assistance task, which we call the Paper Ranking Problem (PRP). To validate our hypotheses on real data, we define a novel generic evaluation framework motivated by science-of-science literature on peer review, using past acceptance labels and citation counts as a reference. 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 noise.

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]. We show that our preference-learning-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 substantially, 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 additional noise in the review scores caused by unreliable referees and bias induced by heterogeneous weighing of quality criteria.

2 Related Work

Cook et al. [2007] propose the conversion from exact review scores to partial rankings on a per-referee basis to mitigate the issue of score miscalibration bias in PR [Wang and Shah, 2019]. For example, if referee A reviewed the papers x, y and z assigning scores of 2, 1 and 3 (higher is better), respectively, this is expressed by the ordering A: y ≺ A x ≺ A z. To aggregate reviewed submissions into a ranking, the authors cast the task as an NP-hard consensus ranking problem: given the partial rankings on overlapping paper subsets, the goal is an output ranking on all papers that violates the least precedence pairs. The authors propose a branch-and-bound algorithm that can find optima on small artificial datasets in reasonable time. Baskin and Krishnamurthi [2009] propose a more efficient neighborhood-based optimization algorithm for the same task. During our experiments on a real-world dataset, we show that both consensus ranking approaches perform inferior in the presence of bias and noise compared to our preference-learning-based method.

Scholarly document quality assessment (SDQA) covers various problem settings from NLP judging paper quality aspects based on paper texts or associated review texts. Kang et al. [2018] introduce the task of paper acceptance prediction (PAP): given papers from past conferences their binary acceptance decisions are the target of prediction. This task has been approached in many flavors [Ghosal et al., 2019; Stappen et al., 2020]. 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 augment paper texts with structural tags and apply it on both tasks in isolation. Li et al. [2019] approach CCP using cross-attention between review texts and the paper abstract extended by author metadata like their h-index as a feature. Several factors hinder the applicability of SDQA methods for assistance in PR: PAP approaches make coarse-grained recommendations and often neglect review texts. Methods from CCP (and PAP) train models on historic papers from broad domains having citation counts (and acceptance decisions) available. As scientific merit is domain- and time-dependent [Lee, 2015], there is a substantial distribution shift between training and inference time when applying these models for a particular, new PR process. Consequently, their recommendations are likely biased to a historic state of the art.

3 Paper Ranking Problem

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

3.1 Problem Definition

The PRP is the task of ranking the submissions to a PR system 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 dynamically.

Let P denote the set of papers submitted to the venue, where each paper p ∈ P is represented by its text. For the set of reviews R, each review r ∈ R is characterized by its text and a vector of scores. Each review is associated with exactly one referee ref(r) ∈ E from the set of referees E and exactly one paper pap(r) ∈ P. Hence, each referee e ∈ E generates a set of reviews R_e = {r | r ∈ R, ref(r) = e} corresponding to a set of reviewed papers P_e = {pap(r) | r ∈ R_e}. The Paper Ranking Problem is then defined as the task of predicting an overall ranking O_P (implying a total order on P', given R, P and E with all associated information). O_P should have minimal ranking distance to the ranking of papers by their true total quality ordering O_t. In case of making acceptance decisions on all submissions together, we have P' = P. To make individual acceptance decisions on subsets of P grouping submissions, for instance, by track or paper type, we model this task as multiple PRP on disjoint subsets P_1', ..., P_n' partitioning P. For conciseness we refer to this set of subsets by P', as well.

In this definition, we assume that there is one true ranking by paper quality for each considered subset of papers. This is an inherent assumption in peer review that is merely reflected in this formalization and which is most plausible for the given context limited to papers within a single PR system.

3.2 Performance Criteria

A useful assistance system increases the efficiency of a PR process, while at least maintaining the quality of output decisions. Hence, we adapt quality criteria from the science-of-science literature on peer review, to measure the performance on the PRP. Specifically, we consider effectiveness, completeness, fairness and efficiency described below. We draw on the concept of random ranking models [Critchlow et al., 1991] to describe a paper ranking algorithm A. The quality of a paper p ∈ P', u(p), is drawn from the true quality distribution. For any x, y ∈ P' the pairwise precedence x ≺_M y is distributed according to the random ranking model M implied by A: this means that x precedes y in output rankings of A with the probability defined by M.
Effectiveness and Completeness Effectiveness and completeness are classical criteria from information retrieval. 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 \( y \).

\[
P(u(x) > u(y)|x \preceq_M y) = 1 \quad (1)
\]

For a complete ranking model, \( x \) should always precede \( y \) in the ranking, given that \( x \) has a higher quality than \( y \).

\[
P(x \preceq_M y|u(x) > u(y)) = 1 \quad (2)
\]

Fairness Fairness in PR is often regarded as the absence of bias in the reviews or as the replicability of the PR process [Walker and Rocha da Silva, 2015]. This perspective neglects the subjectivity of reviewing and desirable types of bias in reviews, like a credit of trust towards potentially groundbreaking works [Bornmann, 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.

\[
P(x \succeq_M y|u(x) = u(y)) = P(y \succeq_M x|u(x) = u(y)) \quad (3)
\]

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.

3.3 Evaluation Framework

To evaluate approaches to the PRP in real scenarios, we approximate the latent true quality per paper \( u(p) \). Being multifaceted in nature, we combine multiple weak indicators for different aspects of quality to approximate \( u(p) \).

Despite uncertainty as to what citation counts measure exactly [Aksnes et al., 2019], they are common proxies for the impact of a paper. Within a fixed scientific field and a fixed time since publication, we use them as indicators for paper impact in our framework. Furthermore, historic acceptance decisions relate to paper quality. While being venue-specific, noisy and highly selective, due to restrictive acceptance quotas, they can serve as valuable quality indicators within the context of a fixed PR process. Although these measures should not be considered in isolation, a balanced, high performance on both of them reveals 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 measure effectiveness and completeness against a ranking by citation counts and binary acceptance decisions practically, we use Spearman’s rank correlation (\( \rho \)) and the area under the receiver operating characteristic (AUROC), respectively. As the probability distribution of precedence pairs \( P(x \succeq_M y) \) is typically not available, we approximate fairness as 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. We hereby ground the commonly used sensitivity analysis on artificial PR data (e.g. [Cook et al., 2007; Wang and Shah, 2019]) on real datasets, to make it more adequate for application scenarios. The evaluation of efficiency follows directly from its definition. After sub-sampling the set of reviews randomly, we measure the performance decrease in terms of effectiveness and completeness compared to the full dataset. During experiments we apply this general evaluation schema on real data.

4 Preference Learning for Paper Ranking

The Paper Ranking Problem can be naturally approached by preference learning to incorporate review texts and account for bias and noise in review scores.

4.1 Preference-based Model of the PRP

In preference learning, relative preferences on the item space serve as the supervision signal. For that, we elicit preferences as for consensus ranking: we convert the review scores of a referee \( e \in E \) into a partial ranking on the set of reviewed papers \( R_e \). This makes a latent assumption: referees consciously or subconsciously compare papers during review. While effects like order bias [Birukou et al., 2011] suggest that this is in fact true, there might be violations for papers on very different topics reviewed by the same referee. However, this assumption is not only induced by the preference perspective, but it is also inherent to the direct use of review scores: 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 partial rankings \( P_{OE} = \{ O_{P_e} | e \in E \} \) are given as training supervision. Here \( O_{P_e} \) is the total order on papers \( P_e \) reviewed by \( e \), which is induced by the order on associated scores. We train the model on all papers \( P_e \), where each paper is represented by a feature vector derived from the review texts and score vectors. The goal is to predict \( O_{P_e} \), assuming that the observed partial rankings are sampled from the true order \( \hat{O}_{P_e} \) with random permutations. Each partial ranking implies a set of preference pairs including tie preferences for non-strict partial orders. In this learning setup, the model orders only items seen during training.

4.2 GPPL for the Paper Ranking Problem

Gaussian processes (GPs) are fully Bayesian regression learners with Gaussian priors achieving high robustness against noise in small-data domains [Rasmussen and Williams, 2006]. Chu and Ghahramani [2005] adapt GPs for preference learning. The authors assume that observed preference pairs \( x \succ y \) (“\( x \) is preferred over \( y \)”), follow a likelihood function where the probability of \( x \succ y \) depends on the delta between
the hidden quality of $x$ and $y$ plus 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 a more scalable approximation for the posterior using stochastic variational inference. The authors show that scalable GPPL effectively ranks arguments by convincingness based on embeddings and linguistic features. Scalable GPPL is suitable for our purpose given that review scores can be noisy and the number of submissions range from below hundred to thousands. Hence, we utilize the method by Simpson and Gurevych [2018] for our approach. To apply GPPL for the PRP, the partial orders $P_{OE}$ are converted into a set of preferences. We enumerate all implied precedence pairs for each $O_{P_S} \subseteq P_{OE}$ 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 to rank all submissions.

4.3 Feature Set Design

We focus on review scores and texts to represent submissions. Hereby, we investigate the importance of review texts for the PRP, and enforce that primarily the human judgements of referees are reflected in the output rankings. We consider three feature sets to define a vector representation, such that two papers are close in the vector space if they are of similar quality.

- **Score Features**: These features are derived from the reviews’ 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 should reflect controversy 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 summaries. The distribution of these argumentative units relates to paper quality, as they reflect its strengths and weaknesses. In the AMPERE dataset [Hua et al., 2019], 400 computer science reviews are annotated with sentence-level discourse labels, including e.g. request or non-argumentative. We fine-tune the last layer of BERT [Devlin et al., 2019] for 4 epochs on 90% of samples in AMPERE resulting in a 0.7969 micro-F1-score on the remaining 10%. We apply this model on each review sentence. The distribution of discourse labels and the proportion of non-argumentative sentences across the reviews of a paper are added as a feature.

- **Embedding 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 be-

| ACL-2018 Dataset Statistics |
|----------------------------|
| # Papers | 1538 |
| # Reviews | 3875 |
| # Reviews per paper | 2.52 ± 0.67 |
| # Reviews per referee | 3.04 ± 1.35 |
| Krippendorff’s $\alpha$ | 0.3596 |
| Intra-class correlation coefficient (ICC) | 0.365 |
| # Preference pairs | 5109 |
| # Comparisons per paper | 6.65 ± 2.62 |

Table 1: On overall scores, we use Krippendorff’s $\alpha$ [Krippendorff, 1980] with an ordinal metric and ICC [McGraw and Wong, 1996].

5 Experiments

We rely on the anonymized PR data from consenting referees at ACL-2018 kindly provided by the collecting parties [Gao et al., 2019]. According to Gao et al., 85% of referees opted-in the collection making the dataset close to complete. ACL-2018 contains anonymous referee identifiers, as in real PR systems, which we use to infer partial rankings. To our knowledge, ACL-2018 is the only available dataset with this information. Nevertheless, access to 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 reviews and acceptance labels for 1538 submissions from 21 tracks. We only consider before-rebuttal reviews to avoid social biases [Gao et al., 2019]. The acceptance rate lies at roughly 25%. Each review has an overall score on a six-point scale and six aspect scores originality, soundness, substance, replicability, meaningful comparison and readability on five-point scales. Additionally, there are five text sections including summary and contributions, strengths, weaknesses, questions and additional comments in each review. Table 1 summarizes relevant statistics of ACL-2018. We observe that the preference signal is ample: each paper is compared on average 6-times. Treating PR as an annotation study, the agreement on overall scores reveals the level of consistency in referees’ judgements. For ACL-2018, the agreement is low for a computer science venue [Ragone et al., 2013], but high compared to social science journals [Bornmann, 2011].

To realize the proposed evaluation strategy on both citation counts and acceptance labels as gold standards, we match the accepted papers in ACL-2018 with the NLPScholar dataset.
converting the hereby acquired citation counts into a ranking. The papers have identical age making their citation counts comparable. As an additional reference, we normalize citation counts per track, which eliminates a preference to topics with broad audiences. We normalize by the sum of citation counts $cc(p)$ of papers in track $t$ to form a ranking: $ncc(p) = \frac{cc(p)}{\sum_{q \in t} cc(q)}$. 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\(^1\). Instead of length-scale optimization or the proposed median-heuristic we use standard normalization on the non-embedding features, as this increased performance on all gold standards substantially. We only report on variations of the feature sets, as no other hyper-parameter configuration improved performance during experiments. 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 all papers and their reviews, but make predictions for paper subsets. We split ACL-2018 into a 20% development and an 80% test set. The sets are randomly sampled while ensuring that the distribution of acceptance labels and positions in the citation count ranking are consistent with the overall population. We report the mean and standard deviation of the performance metrics over five runs with randomly shuffled inputs.

**Baselines** We consider score aggregation strategies including the mean overall score weighted by referee confidence (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] (DCON) and the neighborhood-based [Baskin and Krishnamurthi, 2009] (NCON) consensus rankers. Both algorithms are re-implemented in Python based on the author-provided code. They receive the same input as the GPPL model excluding tie preferences, as they cannot account for them.

**Experimental Scenarios** In the first experiments, we investigate the performance of different feature sets. We optimize two feature configurations, where the first is selected based on acceptance labels (by AUROC) and the second one is selected based on the citation rank counting (using Spearman’s $\rho$) on the development set. Hereby, we identify which features boost performance on each gold standard. In the end, we select the configuration with the best balanced performance on both gold standards for all further experiments.

The evaluation of effectiveness and completeness follows directly from the generic 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 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 integers) 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 $\alpha \in \{30\%, 60\%\}$ 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 guaranteeing at least one review per paper. Again we measure the performance decrease on both gold standards.

**Results and Analysis**

Our implementation of DCON is stopped early after 20h of computation (8 CPUs, 16Gb of RAM). NCON on average converges in 11h. Our GPPL model using the final pre-computed feature set terminates on average in 4.5min.

**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 decision making. To rank consistently with citation counts, embedding-based features are crucial: the best features include only 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 also report the performance using only embedding-based and only score fea-
Table 2: Effectiveness and completeness on the test set of ACL-2018. The mean and standard deviation over five runs are given. ”norm.” refers to the citation counts normalized per track. ”PRAUC” is area under the precision-recall curve.

Table 3: Performance for the fairness scenarios. The noise scenarios refer to the case with $\sigma = 1.0$ with 30% and 60% affected referees.

...
[Walker and Rocha da Silva, 2015] Richard Walker and Pascal Rocha da Silva. Emerging trends in peer review—a survey. *Frontiers in Neuroscience*, 9, May 2015.

[Wang and Shah, 2019] Jingyan Wang and Nihar B. Shah. Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings. In *Proceedings of the 18th AAMAS*, volume 18, 2019.