Are Key-Phrases All that Reviewers Care About? A Comprehensive Benchmarking of Reviewer Matchmaking Systems

Sourish Dasgupta*, Harsh Sharma*, Devansh Patel*, Prarthee Desai*, Anil K. Roy

Dhirubhai Ambani Institute of Information & Communication Technology, India
{sourish_dasgupta, 202111002, 202001262, 202001257, anil_roy}@daiict.ac.in

Abstract

Reviewer Matchmaking (RM) is a pivotal process in academic publishing that aligns manuscripts with appropriate reviewers based on their expertise and prior publications. The demand for an automated RM system has escalated with the significant surge in submissions over the past decade. State-of-the-art (SOTA) RM models are document-representation-based (DR-RM) and match the manuscript and reviewer's past publication using a similarity method defined on a high-dimensional vector space. However, they are far from accurate despite their large-scale usage. In this paper, we establish that conventional RM evaluation measures are unreliable and instead emphasize that standard correlation measures are adequate. For the first time, we compare the performance of six SOTA DR-RM models with those of fourteen SOTA Key-phrase Extraction-based RM (KPE-RM) models - an alternate unexplored approach. We observe that KPE-RM models show comparable results in many cases, with the new best model being PatternRank-RM - a KPE-RM model beating the best DR-RM model SPECTER2-RM (Pearson: 0.004+, Spearman: 0.006+, Kendall: 0.043+). We conclude that KPE-RM models must be contextualized to the RM task and cannot be used as plug-n-play.

1 Introduction

Academic journals and conferences are crucial platforms for researchers to share their work and receive expert feedback. The process of optimally assigning appropriate manuscripts to reviewers based on the reviewers' area of interest, expertise, and availability such that every manuscript is assigned the minimum required reviewer and, at the same time, no reviewer is overburdened by more than the maximum manuscripts is called **Reviewer Allocation** (**RA**) (Stelmakh, Shah, and Singh 2019; Cousins, Payan, and Zick 2023; Aziz, Micha, and Shah 2023). A necessary step for RA is **Reviewer Matchmaking** (**RM**), also known as paper-reviewer matching, where a set of "*most suitable*" reviewers are matched, thereby providing a score to every manuscript-reviewer pair.

Need for accurate RM systems. There has been a significant surge in the number of manuscript submissions received in the past decade, rendering the manually curated, careful RM process unrealistic. Even with a bidding system, the sheer volume hinders reviewers from thoroughly evaluating a long list of manuscripts before deciding on a bid (Zhang et al. 2023). This has emphasized the need for an automated RM system. Ideally, this ***score should correlate with the reviewers' confidence scores*** during reviewing, failing which the RM model stands unreliable. However, there is much scope for improvement in RM systems. A study on selected papers of NeurIPS 2014 shows that there is *no correlation between the quality scores given by the reviewers to accepted papers and the actual citation-impact* of those papers (Cortes and Lawrence 2021). Besides the inherent subjectivity, such situations are also due to sub-optimal RM.

Document Representation (DR) based RM. Most venues use DR-based RM models such as the Toronto Paper Matching System (Charlin and Zemel 2013), ACL Reviewer (ACL-org 2022), and OpenReview (OpenReview-org 2022). These models are based on document embedding-based matchmaking of manuscripts and the reviewer's past publications. The central objective is to represent various arguments (i.e., the research question and associated claims and justifications) within a manuscript and map that to related arguments made in the reviewers' past publications.

Key Phrase Extraction (KPE) based RM. As an alternative approach to existing DR-based RM models, we investigate for the first time the efficacy of Key-Phrase Extraction (KPE) models for the RM task. The motivation behind analyzing KPE-based RM was two-fold. First, ***DR-based models suffer from representational underfitting*** because it may include content that is either peripheral or generic to a specific track. Secondly, ***KPE models have shown significant success in downstream NLP tasks*** such as summarisation (Glazkova and Morozov 2023) and document clustering (Li and Daoutis 2021). We can design KPE-based RM models using KPE models to extract the top-k keyphrases from the manuscripts and the reviewers' past publications and then apply a suitable similarity algorithm to the two sets of keyphrases. We study fourteen SOTA KPE models of architectural designs that are based on deep neural networks (seven models), graph structures (four models), and statistical approaches (three models). For the first time, we compare KPE-RM models with six SOTA DR-based RM models (Table 1; Appendix).

Observations. In this paper, we first show the ***inadequacy of existing RM evaluation frameworks and associated measures***. More specifically, we argue that the conventional

*These authors contributed equally.

expert-annotation-based measures such as Precision@K, as introduced in Mimno and McCallum (2007) and have been widely followed since (Karimzadehgan, Zhai, and Belford 2008; Singh et al. 2023a), are incomplete. We also demonstrate that a more recent reviewer-confidence-rating-based leaderboard that uses Kendall Loss (Stelmakh et al. 2023) is unreliable. We instead recommend standard correlation measures of Pearson r, Spearman ρ, and Kendall τ w.r.t ground-truth confidence scores. We use the gold standard dataset released by Stelmakh et al. (2023) for our study.[1] Several *KPE-based RM models* ***perform comparably*** *with the best performing DR-based RM model*, with KPE-based PatternRank-RM topping the chart. However, we also conclude empirically that the best-performing KPE-based RM models have barely moderate rank correlation. This is because the KPE task is performed as a prequel to the RM task, thereby not extracting keyphrases that are more useful for the RM task, leaving a ***wide scope for improvement***.

2 Preliminaries

2.1 The Reviewer Matchmaking Problem

Reviewer Matchmaking (RM), also called paper-reviewer match, is the first step in the Reviewer Allocation (RA) process. For each submitted manuscript, the RM model matches reviewers from a fixed candidate reviewer pool based on their expertise to achieve a thorough and fair evaluation of the manuscript, leading to constructive feedback and improved quality of submissions. An RM system represents manuscripts using the content in the title, abstract, and body. On the other hand, it represents the reviewer's profile using past peer-reviewed publications of the reviewer, often the recent ones. We define the problem as follows:

Definition 1. ***(Review Matchmaking)*** *Given a set of manuscripts $\mathcal{M} = \{m : (\textit{title, abstract, body})\}$ and a set of candidate reviewer pool $\mathcal{R}$, where each reviewer $r \in \mathcal{R}$ has publication profile $Q_r = \{q_r : (\textit{title, abstract, body, publication-year})\}$, an RM model Θ is to learn a relevance-score function $f_\Theta^{RM} : \mathcal{M} \times \mathcal{R} \rightarrow \mathbb{R}$ and generate a reviewer rank ($\bar{\mathcal{R}}_m$) for each m w.r.t f_Θ^{Rm}.*

2.2 RM Evaluation Measures: Limitations

In this section, we discuss the limitations of two RM evaluation frameworks - (i) external expert annotation-based measures and (ii) reviewers' confidence rating-based measures.

Expert Annotation-based Measure In this evaluation framework, pairs of ***accepted*** manuscripts and assigned reviewers are rated in terms of their quality of assignment by external annotators who are experts from the same area where the manuscripts belong (Mimno and McCallum 2007). An RM model is evaluated in terms of the number of times the predicted relevance score (w.r.t f_Θ^{RM}) aligns with the human-judgment assignment ratings (i.e., 2+ out of a score between 0-3). More specifically, the central idea is to find the subset of the top-K predicted reviewers ($r_{m,k}$) for any manuscript m that has received good human judgments (i.e., ground-truth). To this end, two measure variants have been proposed - soft Precision @K and hard Precision @K as follows:

$$\begin{aligned}\text{Soft P@K} &= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\frac{\sum_{k=1}^{K}\mathbb{I}(\text{score}(m,r_{m,k})\geq 2)}{K};\\ \text{Hard P@K} &= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\frac{\sum_{k=1}^{K}\mathbb{I}(\text{score}(m,r_{m,k})= 3)}{K}\end{aligned} \tag{1}$$

Theoretical Limitations: This evaluation framework has multiple drawbacks. First, it is a rather indirect evaluation method that ***does not guarantee whether an agreement (or disagreement)*** between the RM model and the annotator ***also correlated with a high confidence rating*** of the actual reviewer who assessed the manuscript. Secondly, since the framework only considers accepted manuscripts, it ***does not consider negative pairs*** (i.e., rejected manuscripts) that can also ***receive a high confidence rating*** from the actual reviewer. Finally, it is overly dependent on the subjective annotation of external annotators, and hence, the ***inter-annotator agreement needs to be considered*** before the results can be taken seriously.

Confidence-Rating-based Measure The actual reviewers' confidence rating-based evaluation framework will always be superior, being more direct. However, due to privacy reasons, in most cases, reviewer profiles are anonymous, thereby making such evaluation infeasible. Earlier results that have been reported cannot be reproduced or reused for newer RM models because of this reason (Rodriguez and Bollen 2008; Qian, Tang, and Wu 2018; Anjum et al. 2019). Stelmakh et al. (2023) proposed a dataset (CMU-RM23) that can simulate the desired situation more directly to some extent (see section 4.1 for details). In this dataset, reviewers are replaced by expert "*readers*" who have already read the manuscripts (i.e., accepted ones) and have given a confidence score had they been assigned the manuscripts for review. The authors used Kendall Loss (L_{KL}) as the measure. The key idea is to calculate the number of alignments of the predicted relevance scores (s_r) of the RM model with that of the confidence ratings provided by the reviewer r (denoted ϵ_r). L_{KL} is defined as:

$$\begin{aligned}L_{KL} &= \frac{\sum_r^{\mathcal{R}} L_r}{\sum_r^{\mathcal{R}}\sum_{\substack{i,j=1\\ i<j}}^{m_r}|\epsilon_r^{(i)}-\epsilon_r^{(j)}|} \text{ where:}\\ L_r &= \sum_{\substack{i,j=1\\ i<j}}^{m_r}\Big(\mathbb{I}_{\mathbf{err.}}\{(s_r^{(i)}-s_r^{(j)})\cdot(\epsilon_r^{(i)}-\epsilon_r^{(j)})<0\}\cdot|\epsilon_r^{(i)}-\epsilon_r^{(j)}|\\ &\quad + \mathbb{I}_{\mathbf{tie}}\{(s_r^{(i)}-s_r^{(j)})\cdot(\epsilon_r^{(i)}-\epsilon_r^{(j)})=0\}\cdot\frac{1}{2}|\epsilon_r^{(i)}-\epsilon_r^{(j)}|\Big)\end{aligned} \tag{2}$$

Theoretical Limitations: The Kendall Loss (L_{KL}), too, has four serious drawbacks in its formulation, explained as follows with examples:

Disproportionate agreement: To illustrate this, let's say a reviewer, Alice, gives confidence rating (i.e., ϵ_r) of 5 to a manuscript m_1 and 4 to another manuscript m_2. Now,

[1] We use the acronym CMU-RM23 dataset from here on.

say the RM model predicts Alice's confidence (i.e., s_r) to be 5 and 1, respectively. Although disproportionate, this will be treated as an agreement with Alice, and hence, no Kendall Loss will be imposed, thereby leading to the ***incorrect*** *non-assignment of m_2 to Alice*. The same would happen for another reviewer, Bob, whose confidence, say, is 2 for m_1 and 5 for m_2, while the RM model predicts Bob's confidence to be 4 and 5, respectively.

Unidentified disagreement: This can happen when, in the previous example, the RM model predicts Alice's rating to be 2 for m_1 and 1 for m_2. Kendall Loss will not be imposed, although clearly, the RM model fails to predict the correct confidence. This would also lead to ***incorrect*** *non-assignment of m_1 and m_2 to Alice*. The same goes for Bob, who rates m_1 as 2 and m_2 as 2.5, while the model predicts Bob's rating as 4.5 for m_1 and 5 for m_2.

Minor disagreement: Continuing with the example, if an RM model predicts Alice's ratings to be 4.3 to m_1 and 4.5 to m_2, then the model will still be imposed a penalty of 1. The same goes when Bob rates m_1 as 2.3 and m_2 as 2.5 while the model predicts Bob's rating to be 2.4 for m_1 and 2.3 to m_2.

Mishandling of tie: Let us take the case where there is a tie in Alice's ratings to both m_1 and m_2 (say, 5), while the RM model predicts Alice's ratings to be 1 for both m_1 and m_2 (i.e., a prediction tie). *Although having a huge disagreement, no Kendall Loss will be imposed*. The same goes for Bob, who rates both m_1 and m_2 as 2, while the model predicts that Bob's ratings will be 5 for both m_1 and m_2. In the case when Alice's ratings do not have a tie (say she rates 5 for m_1 and 1 for m_2), while the RM model, as before, has a tie, then a large Kendall Loss penalty of 2 (= 4/2) is imposed *without any specific reason why the first case should not be*.

2.3 Document Representation (DR) based RM

DR-based RM models are used in most of the peer-review systems. These techniques represent manuscripts and reviewer profiles (i.e., past publications) as high-dimensional embeddings. A document includes the title, abstract, and body[2]. The embeddings (often contextual) are supposed to capture the semantics of the scholarly argumentation. Such representation allows for the computation of similarity relevance scores (pipeline outlined in Figure 3; Appendix).

Limitations: Although DR-based RM methods have shown progress, significant obstacles remain due to representational challenges. First, the representation of the reviewer's profile might not accurately portray his/her recent area of interest. It might also happen that the reviewer's profiles may contain irrelevant information from papers that are marginal or peripheral to the manuscript (or even irrelevant), thereby adding noise to the reviewer's profile representation.

As an alternative to DR-based RM, we propose that KPE-based RM should also be seriously considered. Before we detail the generic outline of KPE-based RM models, we briefly describe the KPE task in the following section.

[2]Additional metadata such as author-keywords, reviewer's area of expertise, and review history may also be found.

2.4 Key-Phrase Extraction (KPE)

KPE is a textual information processing task responsible for automatically extracting *characteristic* and *representative* key phrases covering a document's aspects (i.e., ***key ideas***). KPE models can be broadly categorized into three types based on their underlying architectures and extraction techniques: (i) term-statistics based (El-Beltagy and Rafea 2010; Campos et al. 2020; Sparck Jones 1972), (ii) graphical-based (Boudin 2018; Bougouin, Boudin, and Daille 2013; Florescu and Caragea 2017; Mihalcea and Tarau 2004; Wan and Xiao 2008), and (iii) deep neural network based that are mostly fine-tuned versions of some pre-trained language model (PLM) (Grootendorst 2020; Sun et al. 2021; Schopf, Klimek, and Matthes 2022; Kulkarni et al. 2022; Kong et al. 2023). *We argue that since KPE models are supposed to represent a document's key ideas concisely, they could be promising alternative for the RM task.*

3 KPE-based RM: Alternative Approach

KPE-based RM (KPE-RM) models extract key phrases from the manuscripts and the potential reviewers' past publications. The extracted keyphrases are then used to create profiles that represent the central theme of the manuscript m (i.e., the set $m = \{t_m : t_m \in \mathcal{T}_m\}$ where $g_\Theta^{KPE} : m \mapsto \mathcal{T}_m$; $\mathcal{T}_m$ is the extracted set of key phrases t for m), and the reviewer r's publication profile Q_r (= $\{q_r\}$; where $q_r = \{t_{q_r} : t_{q_r} \in \mathcal{T}_{q_r}\}$ where $g_\Theta^{KPE} : q_r \mapsto \mathcal{T}_{q_r}$; $\mathcal{T}_{q_r}$ is the extracted set of key phrases t for publication q_r), respectively. The KPE-based RM problem can be reformulated as learning $f_\Theta^{RM} : (m, Q_r) \mapsto \bar{\mathcal{R}}_m; \forall m \in \mathcal{M}; \bar{\mathcal{R}}_m$: ranked list of reviewers for m. Here, the key phrase t is represented as an embedding $\mathbf{t}$, and f_Θ^{RM} is a similarity measure on the vector space in which t is defined (we have experimented with cosine and Jaccard). We investigate two types of embeddings for this purpose: GloVe embeddings (Pennington, Socher, and Manning 2014) and Sentence-BERT (SBERT) embeddings (Reimers and Gurevych 2019). For deep neural-based RM models, we also experimented with internal model-generated $\mathbf{t}$ embeddings. To compute the aggregate relevance score over all $q_r \in Q_r$, based on which the rank $\bar{\mathcal{R}}_m$ is generated, we use mean and max (termed *Mode*). Figure 1 outlines the top-level KPE-RM pipeline.

4 Evaluation Setup

4.1 Model Benchmarking Dataset

As discussed in section 2.2, to reliably evaluate the comparative accuracy performance between DR-based and KPE-based RM models in a direct way, we need an evaluation dataset that contains gold-reference reviewer confidence ratings and reviewer profiles (i.e., past publications). We, therefore, selected the CMU-RM23 gold-standard dataset (Stelmakh et al. 2023). The dataset consists of 463 manuscripts (i.e., $\mathcal{M}$) and 477 self-reported ground-truth confidence scores (i.e., set of ϵ_r) provided by 58 researchers (i.e., $\mathcal{R}$) who are domain experts and whose profiles (i.e., Q_r) are publicly available (see Table 1). In this framework, domain experts select accepted manuscripts they have read and substitute the actual (anonymous) reviewers.

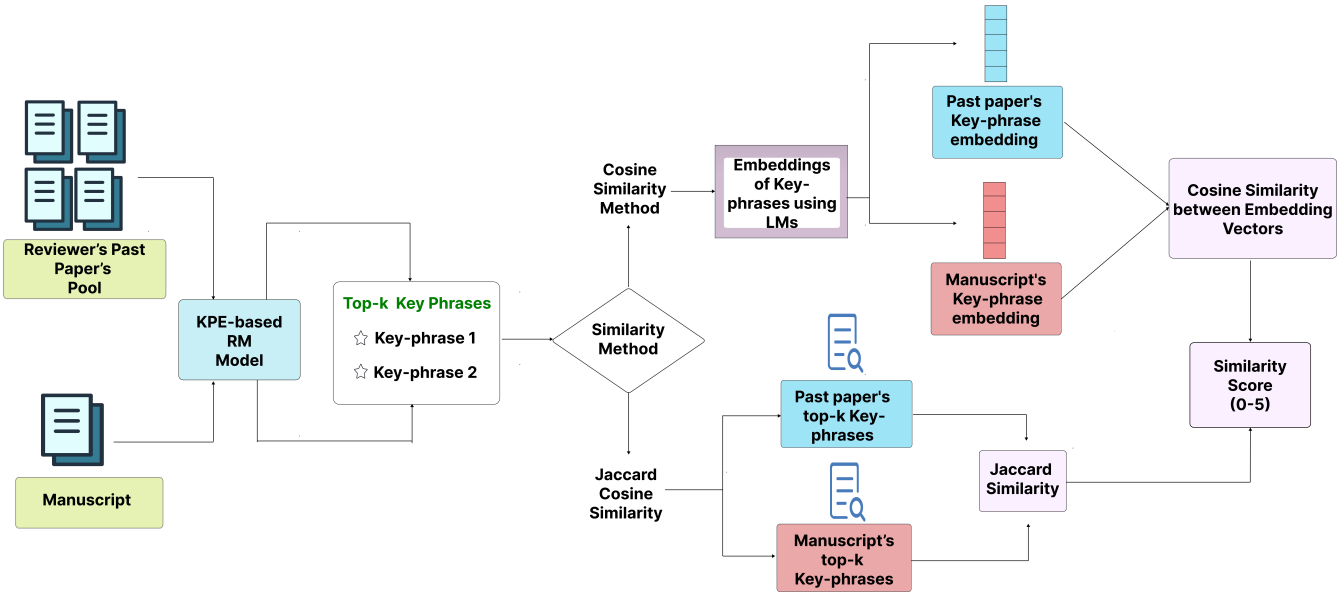

Figure 1: KPE model-based RM pipeline.

Total number of manuscripts: 463		
Characteristic	**Quantity**	**Value**
Open Access	# On semantic scholar	462
	# On arXiv	411
	# PDF available	457
Research Areas	# Computer science	459
Publication Year	% Before 2020	25
	% 2020 or later	75
	Total confidence rating received	477
	Mean manuscripts per reviewer	7.98

Total number of reviewers: 58		
Characteristic	**Quantity**	**Value**
Position	% PhD student	45
	% Faculty	28
	% Post-PhD (non-faculty)	12
Domain Expertise	# Computer science	459
Experience	Mean # publications	54
	Median # publications	20
	Total # publications	3112
	PDF available # publications	1506

Table 1: **CMU-RM23 Gold Standard Dataset**: Stats & domain experts demography; anonymity compliant as per dataset terms.

4.2 Evaluation Measure

An RM model's relevance score must correlate with reviewer confidence to be reliable. A robust evaluation measure should overcome the four issues of Kendall Loss—insensitivity to agreement degree, conflation of disagreement with agreement, over-penalization of minor discrepancies, and undue penalization of ties. We address this using standard correlation measures such as Pearson's r, Spearman's ρ, and Kendall's τ effectively as per following formulation:

$$^{3}\rho_{\Theta}(s,\epsilon) = \text{CORR}\left(\left\{\left(\frac{1}{m_r}\sum_{i=1}^{m_r} s_r^{(i)}, \frac{1}{m_r}\sum_{i} \epsilon_r^{(i)}\right)\right\}_{r=1}^{\mathcal{R}}\right) \quad (3)$$

The above formulation ensures that the four limitations of Kendall Loss, as discussed in section 2.2, are handled. Pearson's r correlation will be < 1 for all the four cases described therein. To illustrate how rank correlation measures such as Spearman ρ and Kendall τ overcome the issues, we continue with the same case-wise examples. In the case of **disproportionate agreement**, we can see that the RM model will rank as per the **mean** predicted rating, thereby predicting the mean rating of Alice to be $(5+1)/2 = 3$, and that of Bob to be $(4+5)/2 = 4.5$, leading to the predicted rank of Bob $>$ Alice. However, the actual average rating of Alice is (5+4)/2 = 4.5, and Bob is (2+5)/2 = 3.5. Therefore, ideally, the rank should have been Alice $>$ Bob. Hence, *unlike Kendall Loss, the rank correlation measures will penalize the model due to misalignment with target ranking*. The second problem of **unidentified disagreement** will not occur since, again following the example in section 2.2, the actual ranking based on Alice's average rating (4.5) and Bob's average rating (2.25) will lead to the rank of Alice $>$ Bob, while the model will predict Alice's average rating to be 1.5 and Bob's rating to be 4.75, leading to just the opposite ranking. *Therefore, the disagreement will not be missed*. The third problem of penalization of **minor disagreement** will also not show up since, following the example in this case, *the model's predicted ranking will not be different than the actual ranking* of Alice $>$ Bob. Finally, the anomaly of **mishandling of tie** does not appear since, as per the first example, the actual ranking will be Alice $>$ Bob, while the model's predicted ranking will be Bob $>$ Alice (Alice's predicted avg. rating is 1 and that of

[3] See Appendix (https://tinyurl.com/5b3mz6zw) B for CORR(·)

Hyperparameter	Explanation	Notation/Values
Similarity Method	Applied similarity measure and the associated representation	Similarity measure: {cos (cosine), jaccard} Rep.: {SBERT, GloVe, KPE model's internal embedding }
Mode	Type of manuscript-reviewer cross similarity	{mean, max}
# Extracted Key Phrases (top-k)	The optimal number of key phrase to be extracted	$k = \{5, 15, 30\}$
Sub-word Elimination	Elimination of sub-words/phrases from top-k (to avoid redundancy)	nd: *no-discard* d: *discard*
Sub-word Elimination Order	The precedence order of the operations: *sub-word elimination* & *select top-k*	+: *sub-word elim.* $\rightarrow$ *select top-k* -: *select top-k* $\rightarrow$ *sub-word elim.*

Table 2: **KPE-RM Hyperparameters**: This leads to the ablation study of 672 KPE-RM model variants.

Bob's is 5) leading to a penalty.

4.3 SOTA RM Models Evaluated

We select six of the best-performing SOTA DR-RM models on the CMU-RM23 dataset (OpenReview-org 2022) for comparative analysis with fourteen SOTA KPE-RM models (see Table; Appendix). The SOTA KPE models were chosen based on their recent performance on scientific article datasets such as Inspec (Hulth 2003), Krapivin (Krapivin, Autaeu, and Marchese 2009), NUS (Nguyen and Kan 2007), and SemEval 2017 (Augenstein et al. 2017) and in terms of their architectural style. More specifically, we study the effect of the architecture (viz. term-statistics-based (Sparck Jones 1972; Campos et al. 2020; El-Beltagy and Rafea 2010), graph-traversal-based (Florescu and Caragea 2017; Wan and Xiao 2008; Mihalcea and Tarau 2004; Bougouin, Boudin, and Daille 2013; Boudin 2018), and deep neural pre-trained model-based (Schopf, Klimek, and Matthes 2022; Sun et al. 2021; Kulkarni et al. 2022; Grootendorst 2020; Xie et al. 2023; Kong et al. 2023)) and the corresponding key phrase extraction technique on the RM performance (Appendix A).

4.4 KPE-RM Model Hyperparameters

We analyze the twenty most recent publications per reviewer and conduct ablation studies on fourteen KPE-RM variants with five hyperparameters: (i) similarity method, (ii) mode, (iii) number of key phrases (top-k, with k-values of 5, 15, and 30), (iv) application of sub-word elimination, and (v) the order of sub-word elimination (see Table 2). The similarity method pairs a metric-space representation with a similarity measure. We employ three representations: key phrase-based bag-of-words (BoW), static term embedding (GloVe (Pennington, Socher, and Manning 2014)), and contextual term embedding (SBERT (Reimers and Gurevych 2019) along with the KPE model's internal embedding), using Jaccard for BoW and cosine for embeddings. We compute pairwise similarity between manuscripts and reviewer profiles, examining the effects of two aggregation modes (mean and max) and assessing sub-word elimination—especially when applied before top-k selection—on RM performance.

5 Observation & Insights

In this section, we outline the comprehensive comparative analysis between the SOTA DR-RM and KPE-RM models. Our experiment was conducted on two setups - (i) title + abstract and (ii) full-text (i.e., title + abstract + body).[4]

5.1 Term-statistics based KPE-RM Models

Title + Abstract: We observe that term-statistics-based RM models are notably weaker than DR-based models in terms of Pearson r (except for BM25-RM, which also is a term-statistics-based model; see Table 3). ***This suggests that term-statistics as a feature is not enough***.

Full-text: We observe that ***KPE models*** such as KPMiner-RM (cos-SBERT/mean/15/nd) that are ***designed to harness the full-text utilizing the term-positional inductive bias perform better*** than title+abstract. However, even after providing additional content about the problem statement and methodology - aspects that can highly impact RM performance, there is no particular change overall (see Figure 2f.).[5] This indicates that ***there is no particular correlation between the core aspects relevant for RM and term frequency***.

5.2 Graph-traversal-based KPE-RM Models

Title + Abstract: We find that several graph-traversal-based models, such as PositionRank-RM (cos-SBERT & Jaccard variants), SingleRank-RM (cos-SBERT), and TextRank-RM (Jaccard), have comparable performance with the DR-RM models in terms of Pearson r and Spearman ρ. Interestingly, all these models outperform DNN-based models except for the best RM model - PatternRank-RM (Jaccard/mean/15/nd). This indicates that ***the embeddings generated are sub-optimal, thereby leading to a sub-optimal top-k list***. On the other hand, the superiority of PositionRank-RM is due to the incorporation of term-position information, whose benefit we also see in KPMiner for full-text.

Full-text: We find that graph-traversal-based models can utilize the full-text content better than term-statistics-based and DNN-based models. Specifically, we observe that TopicRank (cos-SBERT/mean/15/nd) and MultipartiteRank-RM (cos-SBERT/max/30/nd) have notably better results than all DR-RM model results on full-text. In fact, MultipartiteRank tops the chart in terms of all the correlations. We believe that this is due to the ranking of top-k based on the mapping of inter-connected key phrases (forming the context) to a common topic (key phrase cluster) set within the graph structure, which bears direct evidence that ***topic clustering-based topic-graph aids RM performance***. As evident, such a technique would implicitly require full-text content to perform well.

[4]Code: https://github.com/KDM-LAB/KPE-RM-AAAI-25

[5]See Table 7 in Appendix C.1 for more details.

Comparative Performance of KPE-RM models vs. DR-based RM models (Title + Abstract)								
	Similarity Method	**Mode**	**No. of Key-phrases**	**Discard sub-words**	**Pearson r**	**Spearman ρ**	**Kendall τ**	***Kendall Loss*****
KPE-based RM Models								
Term-statistics-based KPE-RM Models								
TF-IDF-RM	cos-SBERT	mean	30	nd	0.357	0.322	0.267	0.283
Yake-RM	jaccard	mean	30	nd	0.329	0.287	0.231	0.310
KPMiner-RM	cos-SBERT	mean	5	d	0.295	0.265	0.216	0.345
Graph-traversal based KPE-RM Models								
PositionRank-RM	cos-SBERT	max	30	nd	0.359	0.360	0.279	0.281
PositionRank-RM	jaccard	mean	30	nd	0.370	0.353	0.279	0.301
SingleRank-RM	cos-SBERT	max	15	nd	0.356	0.336	0.261	0.305
TextRank-RM	jaccard	mean	30	nd	0.352	0.333	0.272	0.313
TopicRank-RM	jaccard	mean	15	nd	0.304	0.324	0.258	0.332
MultipartiteRank-RM	jaccard	mean	15	nd	0.328	0.318	0.246	0.314
Deep Neural Network (DNN)-based KPE-RM Models								
PatternRank-RM (best RM)	jaccard	mean	15	nd	**0.433**	**0.423**	**0.339**	0.273
BERTKPE-RM	jaccard	max	30	nd	0.299	0.283	0.235	0.337
KeyBART-RM	jaccard	mean	15	nd	0.289	0.292	0.221	0.325
KeyBERT-RM	jaccard	mean	-5	d	0.363	0.367	<u>0.3</u>	0.297
One2Set-RM	cos-SBERT	max	-30	d	0.281	0.272	0.203	0.323
PromptRank-RM	jaccard	mean	30	nd	0.381	0.371	<u>0.297</u>	0.283
DR-based RM Models								
SPECTER2-RM (best DR-RM)					<u>0.425</u>	<u>0.417</u>	<u>0.296</u>	0.22
SPECTER2 + SciNCL-RM					<u>0.418</u>	<u>0.406</u>	0.289	0.207
SciNCL-RM					0.394	0.377	0.267	0.22
SPECTER-RM					0.385	0.334	0.237	0.272
SPECTER+MFR-RM					0.370	0.315	0.224	0.235
BM25-RM					0.152	0.248	0.177	0.359

Table 3: **Accuracy Chart**: Best-performing RM model-variants on <u>**Title + Abstract**</u> (manuscripts and reviewers' publications); top-3 scores are <u>underscored</u> (**bold** for best performer); **NT: *****Kendall Loss*** is shown to be unreliable in section 2.2 and established so empirically in section 5.5. For results on full-text see Table 7 in Appendix C.1.

5.3 DNN-based KPE-RM Models

Text+Abstract: PatternRank-RM (Jaccard/mean/15/nd), KeyBERT-RM (Jaccard/mean/-5/d), and PromptRank-RM (Jaccard/mean/30/nd) outperform SPECTER2-RM—the best DR-RM model—across various correlation measures, with PatternRank-RM topping the chart. This underscores that ***exploring alternative KPE-RM-based techniques*** is promising. However, embeddings from these models, as well as those from GloVe/SBERT, do not yield optimal RM results, indicating ***significant scope for improvement*** since even the best correlations remain only moderate.

Full-text: We find that *most DNN-RM models cannot effectively use the additional aspect content of full-texts*, except for PatternRank-RM (cos-SBERT/max/+ & - 15/d), which performs comparably to SPECTER2-RM in terms of Pearson r and Spearman ρ and outperforms it in Kendall τ.

5.4 Ablation Studies

Effect of Similarity Method. We find that GloVe-based cosine similarity method did not perform well compared to more contextual embedding-based methods like SBERT-based and KPE-models' internal embedding (Figure 2a.). Notably, the ***embedding-based similarity method (cos-SBERT)*** outperforms Jaccard when applied to full-texts, making it ***less suitable for RM on shorter texts***. On average, ***internal embedding boosts RM performance more than SBERT-generated embedding***.

Effect of Mode. We do not observe any notable difference in the RM performance if we switch mode from max to mean (see Figure 2b.).

Effect of # top-k Key Phrases. As expected, it is evident from Figure 2c., that if $k = 5$, KPE models suffer w.r.t recall, thereby leading to poor RM. However, we do not find any difference in effect between $k = 15$ and 30.

Effect of Sub-word Elimination. We find that sub-word elimination degrades the overall RM performance (see Figure 2d.). This is because several of the identified sub-words have very different semantics; hence, retaining them is better. This also suggests that we need ***more sophisticated sub-word elimination to handle redundancy***.

Effect of Order of Sub-word Elimination. We do not observe any notable difference in the RM performance if we switch the order of the elimination operation (see Figure 2e.).

Effect of Format. Although the best-performing models perform better on title + abstract, we observe that there is no overall effect of full-text on the studied models (see Figure 2f.). As mentioned earlier, this indicates the scope of improvement in utilizing RM-relevant aspect content.

5.5 Incompatibility of Kendall Loss

In section 2.2, we discuss the theoretical drawbacks of Kendall Loss. We also do not find any empirical equivalence between Kendall Loss-based leaderboard and that generated by the standard correlation measures. In fact, as we see

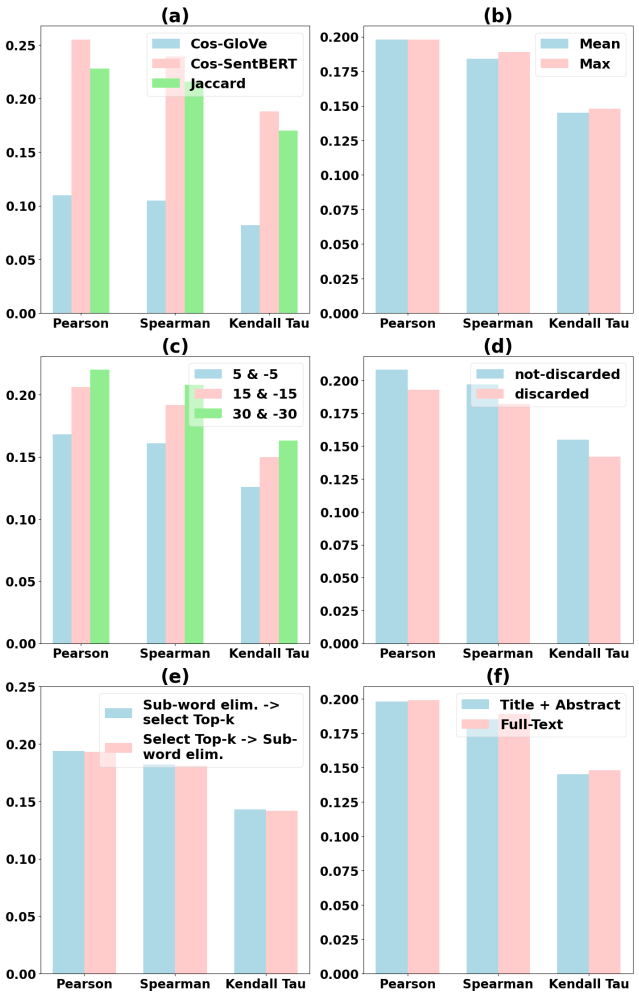

Figure 2: **Ablation Results:** Effect of similarity method (a), mode (b), top-k (c), sub-word elimination (d), sub-word elimination order (e), and format (f).

from Table 4, they have ***strong negative correlation*** w.r.t Spearman ρ and Kendall τ.

6 KPE Models cannot be Plug-n-played

In Table 5 we see disagreement between the leaderboards of the studied models w.r.t the RM task and that of the classical KPE task on four standard evaluation datasets comprising scholarly articles. This shows that top-performing KPE models cannot be plug-n-played for the RM task.

7 Related Work

A widely used term-statistics-based DR-RM model is Okapi-BM25-RM (Robertson and Walker 1994). An early DNN-RM model, ELMo-RM, relies on ELMo for document embeddings (Peters et al. 2018). In recent years, SPECTER-RM (Cohan et al. 2020) and its latest version, SPECTER2-RM (Singh et al. 2023b), generate embeddings of scientific papers using citation-aware fine-tuning on SciBERT.

Leaderboard Correlation: Kendall Loss based vs. Std. Corr. Measures based			
Inter-Corr.	**Pearson r-rank**	**Spearman ρ-rank**	**Kendall τ-rank**
Title + Abstract			
Spearman	-0.927	-0.794	-0.624
Kendall Tau	-0.782	-0.641	-0.471
Full-Text			
Spearman	-0.827	-0.839	-0.692
Kendall Tau	-0.649	-0.657	-0.492

Table 4: **Inadequacy of Kendall Loss**: Inter-correlation of Kendall Loss with std. corr. measures strongly negative.

Leaderboard Correlation: KPE Task (F1@10/M) vs. KPE-RM Task			
Inter-Corr.	**Pearson r-rank**	**Spearman ρ-rank**	**Kendall τ-rank**
Inspec Dataset			
Spearman ρ	-0.123	0.203	-0.098
Kendall τ	-0.070	0.198	-0.015
Krapivin Dataset			
Spearman ρ	-0.406	-0.382	-0.442
Kendall τ	-0.289	-0.289	-0.289
NUS Dataset			
Spearman ρ	-0.382	-0.370	-0.430
Kendall τ	-0.244	-0.244	-0.244
SemEval 2017 Dataset			
Spearman ρ	-0.067	0.083	-0.150
Kendall τ	-0.056	0.111	-0.111

Table 5: **Plug-n-play will not work**: Leaderboard w.r.t KPE task does not correlate with KPE-RM Leaderboard

Another DNN-RM model, SciNCL-RM (Scientific Neighbourhood Contrastive Learning), is based on the SciNCL model (Ostendorff et al. 2022) that uses controlled nearest-neighbor sampling over citation graph embeddings to create a positive and negative sample. An extension to this technique is SPECTER2+SciNCL-RM (OpenReview-org 2022), which combines citation graph information and the neighborhood contrastive learning approach. Very recently, SPECTER+MFR-RM has been proposed that uses a model called MFR (Lin et al. 2023) that captures the different aspects (e.g., *methods*, *experimental setup*, *results & analysis*, etc.) of the manuscript and past publications. However, as shown in this paper, all these models are far from acceptable.

8 Conclusion

In this paper, we first establish that conventional RM evaluation measures such as Precision@K and Kendall loss are inadequate. Instead, we recommend standard correlation measures. For the first time, we have done an extensive comparative analysis of six SOTA document-representation-based RM models with KPE-based RM models. We observe that KPE-RM models are comparable to the DR-RM models and can be a promising alternative direction. However, we also observe performing well in the KPE task does not lead to high RM accuracy, thereby needing KPE-RM models to be more RM-task-oriented.

Acknowledgements

This research has been supported with Cloud GPUs from Google's GPU Support Team and partially funded by IEEE Sensors Council.

References

ACL-org. 2022. Reviewer-paper matching for ACL. https://github.com/acl-org/reviewer-paper-matching.

Anjum, O.; Gong, H.; Bhat, S.; Hwu, W.-M.; and Xiong, J. 2019. PaRe: A Paper-Reviewer Matching Approach Using a Common Topic Space. In *EMNLP/IJCNLP (1)*, 518–528.

Augenstein, I.; Das, M.; Riedel, S.; Vikraman, L.; and McCallum, A. 2017. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. *CoRR*, abs/1704.02853.

Aziz, H.; Micha, E.; and Shah, N. 2023. Group Fairness in Peer Review. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, 2889–2891. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.

Boudin, F. 2018. Unsupervised Keyphrase Extraction with Multipartite Graphs. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 667–672. New Orleans, Louisiana: Association for Computational Linguistics.

Bougouin, A.; Boudin, F.; and Daille, B. 2013. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. In Mitkov, R.; and Park, J. C., eds., *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 543–551. Nagoya, Japan: Asian Federation of Natural Language Processing.

Campos, R.; Mangaravite, V.; Pasquali, A.; Jorge, A.; Nunes, C.; and Jatowt, A. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509: 257–289.

Charlin, L.; and Zemel, R. S. 2013. The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *Proceedings of the 30th International Conference on Machine Learning*, 90–98. PMLR.

Cohan, A.; Feldman, S.; Beltagy, I.; Downey, D.; and Weld, D. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2270–2282. Online: Association for Computational Linguistics.

Cortes, C.; and Lawrence, N. D. 2021. Inconsistency in conference peer review: revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*.

Cousins, C.; Payan, J.; and Zick, Y. 2023. Into the Unknown: Assigning Reviewers to Papers with Uncertain Affinities. In *International Symposium on Algorithmic Game Theory*, 179–197. Springer.

El-Beltagy, S. R.; and Rafea, A. 2010. KP-Miner: Participation in SemEval-2. In Erk, K.; and Strapparava, C., eds., *Proceedings of the 5th International Workshop on Semantic Evaluation*, 190–193. Uppsala, Sweden: Association for Computational Linguistics.

Florescu, C.; and Caragea, C. 2017. PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In Barzilay, R.; and Kan, M.-Y., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1105–1115. Vancouver, Canada: Association for Computational Linguistics.

Glazkova, A. V.; and Morozov, D. A. 2023. Applying Transformer-Based Text Summarization for Keyphrase Generation. *Lobachevskii Journal of Mathematics*, 44(1): 123–136.

Grootendorst, M. 2020. KeyBERT: Minimal keyword extraction with BERT.

Hulth, A. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 216–223.

Karimzadehgan, M.; Zhai, C.; and Belford, G. 2008. Multi-aspect expertise matching for review assignment. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, 1113–1122. New York, NY, USA: Association for Computing Machinery. ISBN 9781595939913.

Kong, A.; Zhao, S.; Chen, H.; Li, Q.; Qin, Y.; Sun, R.; and Bai, X. 2023. PromptRank: Unsupervised Keyphrase Extraction Using Prompt. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9788–9801. Toronto, Canada: Association for Computational Linguistics.

Krapivin, M.; Autaeu, A.; and Marchese, M. 2009. Large Dataset for Keyphrases Extraction.

Kulkarni, M.; Mahata, D.; Arora, R.; and Bhowmik, R. 2022. Learning Rich Representation of Keyphrases from Text. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Findings of the Association for Computational Linguistics: NAACL 2022*, 891–906. Seattle, United States: Association for Computational Linguistics.

Li, X.; and Daoutis, M. 2021. Unsupervised Key-phrase Extraction and Clustering for Classification Scheme in Scientific Publications. arXiv:2101.09990.

Lin, X.; Wang, W.; Li, Y.; Feng, F.; Ng, S.-K.; and Chua, T.-S. 2023. A Multi-facet Paradigm to Bridge Large Language Model and Recommendation. arXiv:2310.06491.

Mihalcea, R.; and Tarau, P. 2004. TextRank: Bringing Order into Text. In Lin, D.; and Wu, D., eds., *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411. Barcelona, Spain: Association for Computational Linguistics.

Mimno, D. M.; and McCallum, A. 2007. Expertise modeling for matching papers with reviewers. In *KDD*, 500–509.

Nguyen, T. D.; and Kan, M.-Y. 2007. Keyphrase Extraction in Scientific Publications. In Goh, D. H.-L.; Cao, T. H.; Solvberg, I. T.; and Rasmussen, E., eds., *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, 317–326. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-77094-7.

OpenReview-org. 2022. Expertise modeling for the OpenReview matching system. https://github.com/openreview/openreview-expertise.

Ostendorff, M.; Rethmeier, N.; Augenstein, I.; Gipp, B.; and Rehm, G. 2022. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11670–11688. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics.

Qian, Y.; Tang, J.; and Wu, K. 2018. Weakly Learning to Match Experts in Online Community. In *IJCAI*, 3841–3847.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *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.

Robertson, S. E.; and Walker, S. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, 232–241. Springer.

Rodriguez, M. A.; and Bollen, J. 2008. An algorithm to determine peer-reviewers. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, 319–328. New York, NY, USA: Association for Computing Machinery. ISBN 9781595939913.

Schopf, T.; Klimek, S.; and Matthes, F. 2022. PatternRank: Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction. In *KDIR*, 243–248.

Singh, A.; D'Arcy, M.; Cohan, A.; Downey, D.; and Feldman, S. 2023a. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Singh, A.; D'Arcy, M.; Cohan, A.; Downey, D.; and Feldman, S. 2023b. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5548–5566. Singapore: Association for Computational Linguistics.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1): 11–21.

Stelmakh, I.; Shah, N. B.; and Singh, A. 2019. PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, 828–856. PMLR.

Stelmakh, I.; Wieting, J.; Neubig, G.; and Shah, N. B. 2023. A Gold Standard Dataset for the Reviewer Assignment Problem. arXiv:2303.16750.

Sun, S.; Liu, Z.; Xiong, C.; Liu, Z.; and Bao, J. 2021. Capturing Global Informativeness in Open Domain Keyphrase Extraction. arXiv:2004.13639.

Wan, X.; and Xiao, J. 2008. CollabRank: Towards a Collaborative Approach to Single-Document Keyphrase Extraction. In Scott, D.; and Uszkoreit, H., eds., *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 969–976. Manchester, UK: Coling 2008 Organizing Committee.

Xie, B.; Wei, X.; Yang, B.; Lin, H.; Xie, J.; Wang, X.; Zhang, M.; and Su, J. 2023. WR-ONE2SET: Towards Well-Calibrated Keyphrase Generation. arXiv:2211.06862.

Zhang, Y.; Shen, Y.; Chen, X.; Jin, B.; and Han, J. 2023. " Why Should I Review This Paper?" Unifying Semantic, Topic, and Citation Factors for Paper-Reviewer Matching. *arXiv preprint arXiv:2310.14483*.