PaRe: A Paper-Reviewer Matching Approach Using a Common Topic Space

Omer Anjum[*,1], **Hongyu Gong**[*,1], **Suma Bhat**[1], **Jinjun Xiong**[2], **Wen-Mei Hwu**[1]
[1]University of Illinois at Urbana-Champaign, USA
[2] IBM Thomas J. Watson Research Center, USA
{oanjum, hgong6, spbhat2, w-hwu}@illinois.edu, jinjun@us.ibm.com

Abstract

Finding the right reviewers to assess the quality of conference submissions is a time consuming process for conference organizers. Given the importance of this step, various automated reviewer-paper matching solutions have been proposed to alleviate the burden. Prior approaches, including bag-of-words models and probabilistic topic models have been inadequate to deal with the vocabulary mismatch and partial topic overlap between a paper submission and the reviewer's expertise. Our approach, the common topic model, jointly models the topics common to the submission and the reviewer's profile while relying on abstract topic vectors. Experiments and insightful evaluations on two datasets demonstrate that the proposed method achieves consistent improvements compared to available state-of-the-art implementations of paper-reviewer matching.

1 Introduction

The peer review mechanism constitutes the bedrock of today's academic research landscape spanning submissions to conferences, journals, and funding bodies across numerous disciplines. Matching a paper (or a proposal) to an expert in the topic presented in the paper requires the knowledge of diverse topics of both the submission as well as that of the reviewer's expertise in addition to knowing recent affiliations and co-authorship to resolve conflict of interest. Considering the scale of current conference submissions, performing the task of paper-reviewer matching manually incurs significant overheads to the program committee (PC) and calls for automating the process. Faced with record number of paper submissions essentially interdisciplinary in nature, the inadequacy of available reviewer matching systems to scale to the current needs is being expressed by many conference program committees. It is also notable that the approaches to address the challenges seem ad-hoc and non-scalable, as described in a few of the PC blogs; "Looking at the abstracts for many of the submissions it also quickly became clear that there was disparity in how authors chose topic keywords for submissions with many only using a single keyword and others using over half a dozen keywords. As such relying on the keywords for submissions became difficult. The combined effect of these problems made any automatic or semi-automatic assignment using HotCRP suboptimal...So, we chose to hand assign the papers." (Falsafi et al., 2018), and again in "Our plan was to rely on the Toronto Paper Matching System (TPMS) in allocating papers to reviewers. Unfortunately, this system didnt prove as useful as we had hoped for (it requires more extensive reviewer profiles for optimal performance than what we had available) and the work had to rely largely on the manual effort..." (ACL, 2019). Noting the urgent need to advance research to address this problem, we study this challenge of matching a paper with a reviewer from a list of potential reviewers for the purpose of assessing the quality of the submission.

Aside from the long precedence of research in the related area of expertise retrieval – that of expert finding and expert profiling (Balog et al., 2012), several recent attempts have been made to automate the process (Price and Flach, 2017). These include, the Toronto paper matching system (Laurent and Zemel, 2013), the IEEE INFOCOM review assignment system (Li and Hou, 2016), and the online reviewer recommendation system (Qian et al., 2018). Central to these systems is a module that performs the paper-reviewer assignment, which can be broken down into its matching and constraint satisfaction constituents. The constraint

*Omer Anjum and Hongyu Gong have equal contribution.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 518–528, Hong Kong, China, November 3–7, 2019.

satisfaction component typically handles the constraints that each paper be reviewed by at least a few reviewers, each reviewer be assigned no more than a few papers, and thaat reviewers not be assigned papers for which they have a conflict of interest. A second constituent is that of finding a reviewer from a list of reviewers based on the relevance of the person's expertise with the topic of the submission. This latter aspect will be the focus of our study.

Available approaches to solve this matching problem can be broadly classified into the following categories (Price and Flach, 2017): a) Feature-based matching, where a set of topic keywords are collected for each paper and each reviewer. The reviewers are then ranked in order of the number of keyword matches with the paper; b) Automatic feature construction with profile-based matching, where the relevance is decided by building automatic topic representations of both papers and reviewers; c) Bidding, a more recent method, involves giving the reviewers access to all the papers and asking them to bid on papers of their choice. The approaches used in this study are of the profile-based matching kind, where we rely on the use of abstract topic vectors and word embeddings (Mikolov et al., 2013b) to derive the semantic representations of the paper and the expertise area of the reviewer. This is a departure from the bag-of-words approach taken in related prior approaches, e.g. (Laurent and Zemel, 2013), relying on automatic topic extraction using keywords – a ranked list of terms taken from the paper and the reviewer's profile.

In general, we assume that a reviewer can be represented by a collection of the abstracts of her past publications (termed as the reviewer's profile) and a submission by its abstract. While attempting to match the paper with the reviewer via their profile representations, the obvious difference in the document lengths gives rise to a mismatch due to the small overlap in vocabulary and a consequent scarcity of shared contexts of these overlapping terms. This is because, while past publications of a reviewer may be sufficient to provide a reasonable context for a topical word, a submission abstract provides a very limited context for that topical word.

To alleviate this problem of mismatched 'profiles', we use the idea of a shared topic space between the submission and the reviewer's profile. In our experiments we compare our approach to match the profiles using abstract vectors with that using the hidden topic model (also a set of abstract topic vectors) (Gong et al., 2018). The two approaches primarily differ in the way the shared topic space is constructed, which we describe in Section 4. We also include other baseline comparisons where the matching is done on the basis of common topic words (keywords) and word- or document-embeddings.

This study makes the following contributions:
(1) Instead of relying on a collection of topic words (keywords chosen by the authors or experts), our approach relies on abstract topic vectors to represent the common topics shared by the submission and the reviewer.
(2) We propose a model that outperforms state-of-the-art approaches in the task of paper-reviewer matching on a benchmark dataset (Mimno and McCallum, 2007). Additionally, a field evaluation of our approach performed by the program committee of a tier-1 conference showed that it was highly useful.

2 Related Work

The paper-reviewer matching task lays the basis for the peer review process ubiquitous in academic conferences and journals. Existing automatic approaches can be broadly categorized into the following types according to the type of models for comparing documents: *feature-based models*, *probabilistic models*, *embedding-based models*, *graph models* and *neural network models*.

Feature-based models. A list of keywords which summarizes the topics of a submission is used as informative features in the matching process (Dumais and Nielsen, 1992; Basu et al., 1999). Automatic extraction of these features achieves higher efficiency and one commonly-used feature is a bag-of-words weighted by the words' TF-IDF scores (Jin et al., 2018; Tang et al., 2010; Li and Hou, 2016; Nguyen et al., 2018).

Probabilistic models. The Latent Dirichlet Allocation (LDA) model is the most commonly used probabilistic model in expertise matching, where each topic is represented as a distribution over a given vocabulary and each document is a mixture of hidden topics (Blei et al., 2003). The popular Toronto Paper Matching System (TPMS) (Laurent and Zemel, 2013) uses LDA to generate the similarity score between a reviewer and a submis-

sion (Li and Hou, 2016). One limitation of LDA is that it does not make use of the potential semantic relatedness between words in a topic because of its assumption that words are generated independently (Xie et al., 2015). Variants of LDA have been proposed to incorporate notions of semantic coherence for more effective topic modeling (Hu and Tsujii, 2016; Das et al., 2015; Xun et al., 2017). Beyond having probabilistic models for topics, Jin et al. sought to capture the temporal changes of reviewer interest as well as the stability of their interest trend with probabilistic modeling (Jin et al., 2017).

In addition to their inherent limiting assumptions such as independence of semantically related words, probabilistic models, including LDA, require a large corpus to accurately identify the topics and topic distribution in each document, which can be problematic when applied to short documents, such as abstracts.

Embedding-based models. Latent Semantic Indexing (LSI) proposes to represent a document as a single dense vector (Deerwester et al., 1990). The documents corresponding to reviewers and submissions can thus be transformed into their respective vector representations. The relevance of a reviewer to a given submission would then be measured using a distance metric in the vector space, such as the cosine similarity. Other approaches have used word or document embeddings as document representations in order to compare two documents.

Kou et al. derived topic vectors by treating each topic as a distribution of words (Kou et al., 2015). In comparison, the key improvement in our work is that the topics are derived based on word embeddings instead of word distributions. Moreover, we derive common topics for each submission-reviewer pair, and as a result, the topics can vary from pair to pair.

Another approach to capture similarity between documents is by the use of the Word Mover's Distance (WMD). It relies on the alignment of word pairs from two texts, and the textual dissimilarity is measured as the total distance between the vectors of the word pairs (Kusner et al., 2015). More recently, a hidden topic model has been used to compare two documents via extracted abstract topic vectors, which showed a strong performance in comparing document for semantic similarity (Gong et al., 2018).

Extending the models in this category, we propose the common topic model. Similar to the hidden topic model, we extract topic vectors using word embeddings and match documents at the topic level. The hidden topic model extracts topics purely relying on the reviewer profile, so the topic vectors can be regarded as a summary of the reviewers' research interest. In contrast, the common topic vectors are selected based on the knowledge of both the submission and the reviewer's profile, which are expected to capture the topical overlap between the two. As we will show in the qualitative evaluation, the hidden topic model is likely to miss some important topics when a reviewer has broad research interests, resulting in an underestimation of the paper-reviewer relevance. The common topic model is able to overcome this limitation by extracting topics with reference to submissions.

Graph models. All of the models mentioned above only assume access to the texts of submissions and reviewers' publications. Some works also make use of external information such as coauthorship to improve the matching performance. For instance, Liu et al. capture academic connections between reviewers using a graph model, and show that such information improves the matching quality. Each node in their graph model represents a reviewer (Liu et al., 2014). There is an edge between two nodes if the corresponding reviewers have co-authored papers, and the edge weight is the number of publications. This work also uses LDA to measure the similarity between the submission and the reviewer.

Neural network models. Dense vectors are learned by neural networks as the semantic representation of documents (Socher et al., 2011; Le and Mikolov, 2014; Lin et al., 2015; Lau and Baldwin, 2016). When it comes to the task of expertise matching, the reviewer-submission relevance can thus be measured by the similarity of the vector representations of their textual descriptions.

3 System Overview

The different stages of our system, together called PaRe, are briefly explained as below:

Data collection. At this stage, we collect previous publications from one or more tier-1 conferences in the same domain as the one to which reviewer-submission matching is applied. This data is used to create our pool of candidate reviewers and do-

main knowledge of the research area. The source of the data is Microsoft Academic Graph (MSG) (Sinha et al., 2015). All the abstracts of a reviewer are concatenated as one document, which is then used to profile the reviewer. Reviewers' profiles reflect their research topics, which are later used in the reviewer-submission matching process.

Data processing. Since our proposed model is based on word embeddings, we pre-train embeddings using CBOW model of word2vec on the collected publications (Mikolov et al., 2013a). The dense word representations are intended to capture domain-specific lexical semantics. The data collection and processing are detailed in Section 5.

Reviewer-submission matching. A common topic modeling approach is proposed in this work to match reviewers with submissions. The model compares the abstracts of submissions and reviewers' past abstracts during the matching process to decide the reviewer-submission relevance by finding their common research topics. The algorithm is described in Section 4.

4 Modeling

For the purpose of our study, we consider a reviewer's profile to be the concatenation of the abstracts from their previous publications. Let m be the number of words in the reviewer's profile and let the *normalized* word embeddings of these words be stacked as a reviewer matrix $\mathbf{R} \in \mathbb{R}^{\mathbf{d}\times\mathbf{m}}$, where d is the embedding dimension. Since the embedding is normalized, we have $\|\mathbf{R_i}\|_\mathbf{2} = \mathbf{1}$ for each column in $\mathbf{R}$. Next suppose that the submission is represented by an $n-$word sequence of its abstract. Similar to the case of the reviewer, we stack its normalized embeddings as a submission matrix $\mathbf{S} \in \mathbb{R}^{\mathbf{d}\times\mathbf{n}}$. Also we have for each column $\|\mathbf{S_j}\|_\mathbf{2} = \mathbf{1}, \forall \mathbf{1} \leq \mathbf{j} \leq \mathbf{n}$.

Common topic selection. Inspired by the compositionality of embeddings (Gong et al., 2017) and the hidden topic model in document matching (Gong et al., 2018), our intention is to extract topics from reviewer profiles and submissions to summarize their topical overlap. We would like to remind the reader that the topics extracted are neither words nor distributions, but only abstractions and constitute a set of numeric vectors that do not necessarily have a textual representation. To establish the connection between topics, reviewer profiles and submissions, we assume that the topic vectors can be written as a linear combination of the embeddings of component words in either the reviewer profiles or the submissions. This assumption is supported by the geometric property of word embeddings that the weighted sum of the component word embeddings have been shown to be a robust and efficient representation of sentences and documents (Mikolov et al., 2013b). Intuitively, the extracted common topics would be highly correlated with the subset of the words in the reviewer profile or that of the submission in terms of semantic similarity.

Let both the reviewer and the submission have K research topics, with each topic represented by a $d-$dimensional vector. This vector is an abstract topic vector and does not necessarily correspond to a specific word or a word distribution as in LDA (Blei et al., 2003). Suppose that these topic vectors of the reviewer are stacked as a matrix $\mathbf{P} \in \mathbb{R}^{d\times K}$, and those of the submission as $\mathbf{Q} \in \mathbb{R}^{d\times K}$. Therefore, these matrices can be represented as a linear combination of the underlying word vectors.

$$\begin{aligned} \mathbf{P} &= \mathbf{Ra}, \\ \mathbf{Q} &= \mathbf{Sb}, \end{aligned} \tag{1}$$

where $\mathbf{a} \in \mathbb{R}^{m\times K}, \mathbf{b} \in \mathbb{R}^{n\times K}$ are the coefficients in the linear combinations.

Our goal is to find the common topics shared by a given reviewer and a given submission, to account for the overlap of their research areas. We consider a pair of topics from a reviewer and a submission respectively to constitute a pair of common topics if they are semantically similar. For example, if the reviewer's research areas are *machine learning* and *theory of computation*, and the submission is about *classification* in *natural language processing*, then (*machine learning, classification*) can be regarded as a pair of common topics, while the other pairs corresponding to the areas *theory of computation* and *natural language processing* are much less similar. We used cosine similarity to measure the semantic similarity of two topic vectors [*]. The similarity $\text{sim}(\mathbf{P}_k, \mathbf{Q}_k)$ between reviewer topic $\mathbf{P}_k$ and submission topic $\mathbf{Q}_k$ is shown below.

$$\text{sim}(\mathbf{P}_k, \mathbf{Q}_k) = \frac{\mathbf{P}_k^T \mathbf{Q}_k}{\|\mathbf{P}_k\| \cdot \|\mathbf{Q}_k\|}. \tag{2}$$

[*]We leave it to future work to experiment with other useful measures of semantic similarity.

For K pairs of topic vectors $\{\mathbf{P_k}, \mathbf{Q_k}\}_{k=1}^{K}$, their similarity is the sum of the pairwise similarities:

$$\text{sim}(\mathbf{P}, \mathbf{Q}) = \sum_{k=1}^{K} \text{sim}(\mathbf{P}_k, \mathbf{Q}_k). \quad (3)$$

This in turn translates to identifying the common research topics between the reviewer and the submission, i.e., we need to find K such pairs of topics that have the maximum similarity. Based on our discussions above, the approach of common topic extraction can be formulated as an optimization problem:

$$\begin{aligned} &\max_{\mathbf{a},\mathbf{b}} \text{sim}(\mathbf{P}, \mathbf{Q}) \\ \text{s.t.} \quad &\mathbf{P} = \mathbf{R}\mathbf{a}, \\ &\mathbf{Q} = \mathbf{S}\mathbf{b}, \\ &\mathbf{P}^T\mathbf{P} = \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \end{aligned} \quad (4)$$

The first two constraints are based on the linear assumption shown in Eq. 1. Without loss of generality, we add the third constraint that the topic vectors are orthogonal as shown in Eq. 4 to avoid generating multiple similar topic vectors. The closed-form solution to this optimization problem can be derived via singular value decomposition on the correlation matrix of $\mathbf{R}$ and $\mathbf{S}$ (Wegelin, 2000).

Let topic vectors $\mathbf{P}^*$ and $\mathbf{Q}^*$ be the optimal solution, both describing the common topics shared by the reviewer and the submission. In the following discussions, we use $\mathbf{P}^*$ as the common topic vectors.

Common topic scoring. To further quantify the reviewer-submission relevance, we need to evaluate how significant these common topics are for the reviewer and the submission respectively. Reusing the example where a reviewer's area are *machine learning* and *theory of computation*, we know that *machine learning* is the common topic between the reviewer and the submission. If the topic of *machine learning* were only a small part of the reviewer's publications, the reviewer may not be a good match for the submission since reviewer is more of an expert in *theory of computation* than in *machine learning*.

To evaluate how well the topics reflect a reviewer's expertise, we define the importance of common topics $\mathbf{P}^*$ for both the reviewer and the submission. Consider the vector of the $i-$th word in the reviewer's profile, $\mathbf{R}_i$, and the k-th topic vector $\mathbf{P}_k^*$. The relevance between $\mathbf{R}_i$ and $\mathbf{P}_k^*$ is defined as the their squared cosine similarity.

$$\text{rel}(\mathbf{R}_i, \mathbf{P}_k^*) = \cos^2(\mathbf{R}_i, \mathbf{P}_k^*) = (\mathbf{R}_i^T\mathbf{P}_k^*)^2. \quad (5)$$

Note that we do not use cosine similarity as is, since $\mathbf{R}_i$ and $\mathbf{P}_k^*$ might be negatively correlated and the cosine similarity can be negative. Instead, we use the square of the cosine similarity to reflect the strength of their correlation.

The relevance between word $\mathbf{R}_i$ and a set of topic vectors $\mathbf{P}^*$ is defined as the sum of the relevance between the word and each topic vector.

$$\text{rel}(\mathbf{R}_i, \mathbf{P}^*) = \sum_{k=1}^{K} \text{rel}(\mathbf{R}_i, \mathbf{P}_k^*). \quad (6)$$

We can think of word vector $\mathbf{R}_i$ to be projected along the K dimensions of a linear subspace spanned by topic vectors in $\mathbf{P}^*$. If $\mathbf{R}_i$ lies in this linear subspace, then it can be represented as a linear combination of the topic vectors. In this case, $\text{rel}(\mathbf{R}_i, \mathbf{P})$ achieves the maximum of 1. If the word vector is orthogonal to all topic vectors in $\mathbf{P}^*$, the relevance results in the minimum relevance of 0. Thus, the range of $\text{rel}(\mathbf{R}_i, \mathbf{P})$ is from 0 to 1.

Furthermore, we define the relevance between the reviewer and the topics as the average of the relevance between the words and the topics.

$$\text{rel}(\mathbf{R}, \mathbf{P}^*) = \frac{1}{m} \sum_{i=1}^{m} \text{rel}(\mathbf{R}_i, \mathbf{P}^*). \quad (7)$$

The reviewer-topic relevance $\text{rel}(\mathbf{R}, \mathbf{P})$ also ranges from 0 to 1.

Similarly, we measure the relevance between a submission and a set of common topics, $\text{rel}(\mathbf{S}, \mathbf{P}^*)$ by measuring the relevance between words in the submission and common topics. The submission-topic relevance reflects the importance of the common topics for a submission. We define the reviewer-submission matching score as a harmonic mean (f-measure) of the reviewer-topic and submission-topic relevance (Powers, 2015).

$$\text{rel}(\mathbf{R}, \mathbf{S}) = \frac{2 \cdot \text{rel}(\mathbf{R}, \mathbf{P}^*) \cdot \text{rel}(\mathbf{S}, \mathbf{P}^*)}{\text{rel}(\mathbf{R}, \mathbf{P}^*) + \text{rel}(\mathbf{S}, \mathbf{P}^*)}. \quad (8)$$

The reviewer-submission relevance is high when the common topic vectors $\mathbf{P}^*$ are highly relevant to both the reviewer and the submission.

It indicates that the submission has a substantial overlap with reviewer's research area, and that the reviewer is considered to be a good match for the submission.

5 Experiments and Results

In this section, we empirically compare our proposed common topic model approach against a variety of models in the task of expertise matching.

5.1 Dataset

For our experiments, we use the two datasets described below.

NIPS dataset. This is a benchmark dataset described in (Mimno and McCallum, 2007) and commonly used in the evaluation of expertise matching. It consists of 148 NIPS papers accepted in 2006 and abstracts from the publications of 364 reviewers. It includes annotations from 9 annotators on the relevance of 650 reviewer-paper pairs. Each pair is rated on a scale from 0 to 3. Here "0" means irrelevant, "1" means slightly relevant, "2" means relevant and "3" means very relevant.

A new dataset. Our proposed paper-reviewer matching system is applied to a tier-1 conference in the area of computer architecture. We created a new dataset for the evaluation of expertise matching from the submissions to this conference. We first collected a pool of 2284 candidate reviewers with publications in top conferences of computer architecture. A reviewer selection policy was adopted by the conference program committee to select reviewers still active in relevant areas. Reviewers were excluded if

1) they started publishing 40 years ago, but had no publications for the last ten years;

2) they did not have publications for the last ten years and have fewer than three papers before that.

The publications of these reviewers were collected from Microsoft Academic Graph (Sinha et al., 2015). Each reviewer had at least one publication, and some reviewers had as many as 34 publications. Again the abstracts were used as reviewers' profile.

We then used our proposed common topic model to assist the program committee of the conference on computer architecture, and recommended most relevant reviewers to all submissions in the conference. We randomly selected 20 submissions and with the help of the committee, we collected feedbacks from 33 reviewers on their relevance to the assigned submissions. These 33 reviewers were among the top reviewers recommended by our system for each of 20 submissions. The relevance was rated on a scale from 1 to 5, where a score of "1" meant that the paper was not relevant at all, "2" meant that the reviewer had passing familiarity with the topic of the submission, "3" meant that the reviewer knew the material well, "4" meant that the reviewer had a deep understanding of the submission, and "5" means that the reviewer was a champion candidate for the submission.

5.2 Baselines

We include previous approaches to paper-reviewer matching as our baselines.

- **APT 200**. Author-Person-Topic (Mimno and McCallum, 2007) is a generative probabilistic topic model which groups documents of an author into different clusters with the author's topic distribution. Clusters represent different areas of a reviewer's research.
- **Single Doc**. The Single Doc model is a probabilistic model which takes the idea of language modeling and estimates the likelihood that a submission is assigned to a reviewer given the reviewer's previous works (Mimno and McCallum, 2007).
- **Latent Dirichlet Allocation (LDA)**: LDA and its variants are the most popular topic models in expertise matching systems (Blei et al., 2003). LDA models assume that each document is a mixture of topics where each topic is a multinomial distribution over the words.
- **Hierarchical Dirichlet Process (HDP)**. HDP model is an extension of LDA (Teh et al., 2006). It is a non-parametric mixed-membership Bayesian model with variable number of topics. It is effective in choosing the number of topics to characterize a given corpus.
- **Random Walk with Restart (RWR)**. RWR is a graph model with sparsity constraints in expertise matching (Liu et al., 2014). It relies on LDA to capture reviewer-submission relevance and also takes diversity into consideration in the matching process.
- **Word Mover's Distance (WMD)**. WMD is a distance metric between two documents on the basis of pre-trained word embeddings (Kusner et al., 2015). It calculates the dissmilarity be-

Method	Number of Topics	P@5	P@10	P@5	P@10	P@5	P@10
		GT1		GT2		GT3	
Common Topic Model	5	53.7	43.0	58.5	49.1	63.7	**55.8**
	10	**56.6**	**44.6**	**63.2**	**50.4**	**67.2**	55.2
	20	54.4	43.4	59.2	49.5	63.6	54.7
Hidden Topic Model	10	43.4	41.1	46.4	45.4	61.8	46.3
	20	51.3	43.4	58.4	49.0	63.6	53.6
	30	47.5	40.0	49.6	44.0	56.3	49.4
APT200	200	41.18	29.71	-	-	-	-
Single Doc	-	44.71	27.35	-	-	-	-
LDA	50	41.3	38.4	51.2	45.0	55.4	51.5
	200	47.5	37.3	53.6	43.6	50.9	50.0
	300	46.2	38.8	52.0	46.8	50.0	45.2
HDP	-	45.5	38.0	48.0	44.5	55.4	50.0
RWR	-	45.3	43	-	-	-	-
Doc2Vec	-	51.7	41.1	59.2	46.8	64.5	51.1
WMD	-	36.1	32.4	42.2	36.8	46.8	41.8

Table 1: The mean precision of different baselines with optimal hyperparamters on the NIPS dataset. A reviewer is classified as relevant with a TREC score ≥ 2.

tween two documents, which is measured by the embedding distance of aligned words in these documents.

- **Hidden Topic Model**. This model proposes to learn hidden topic vectors to measure document similarity based on word embeddings (Gong et al., 2018).
- **Doc2Vec**. Doc2Vec is a neural network model which trains document embeddings to predict component words in the documents (Le and Mikolov, 2014). In expertise matching, the Doc2Vec model is pre-trained on the corpus consisting of reviewers' previous publications. We use the trained model to generate representations for reviewers and submissions respectively. The reviewer-submission relevance is quantified by the cosine similarity of their embeddings.

Setting. Since our model relies on word embeddings, we pre-train embeddings on all papers published in the NIPS conference until 2017 for the matching task in NIPS dataset. Similarly for our new dataset, we collected a corpus of publications until 2018 from top computer architecture conferences for embedding training. The embedding dimension was set as 100, and these word embeddings were also used in two embedding-based baselines: word mover's distance and hidden topic model. For a fair comparison, the corpora used for word embedding training were also used to train Doc2Vec model to generate document embeddings.

5.3 Results on NIPS Data

The NIPS dataset provides ground truth relevance for reviewer-submission pairs, and the relevance scales from 0 to 3. A score of 0 is assigned when that the reviewer is considered to be *not relevant* and a score of 3 is assigned when the reviewer is considered to be *highly relevant*. We set a relevance threshold of 2, and considered reviewers with a score equal to or higher than this threshold to be relevant reviewers to the given submission. In our matching system, we sorted reviewers in decreasing order of the predicted relevance score for a given submission.

Evaluation Metric. Precision at k (P@k) is a commonly-used ranking metric on NIPS dataset. P@k is defined to be the percentage of relevant reviewers in the top-k recommendations made by the model to a submission. It is likely that the top-k recommendations made by the model contain reviewers whose relevance information is not available in the ground truth. To address this issue, we first discard reviewers that do not have a relevance information prior to calculating P@k. In our experiments, we set k to be 5 and 10. We report the average P@k over all submissions in Table 1.

We note that not all submissions in NIPS dataset have the same number of relevant reviewers and a

failure to account for this discrepancy would negatively impact the performance of a system. For example, a submission with only one relevant reviewer would result in a P@5 no higher than 20% for any model. In order to take this discrepancy into consideration, we report the performance only on submissions with at least two relevant reviewers in the columns of "GT2", and on submissions with at least three relevant reviewers in column "GT3". In "GT1", we report the performance without making this distinction.

The reviewer-submission matching results of our model on the NIPS dataset are presented in Table 1 alongside those of our chosen baselines. We note that the results for APT 200 and Single Doc were only available for GT1 and we report them as such. Some approaches including Common Topic Model, Hidden Topic Model and LDA required a hyperparameter (number of topics) to be specified. We performed experiments on NIPS data with different number of topics in Table 1. As is shown, our proposed approach consistently outperforms the strong baselines. We also note that Hidden Topic Model and Doc2Vec are competitive approaches in expertise matching compared against probabilistic models.

Expertise Level	% of Reviewers Predicted by the System
$\geq$5	15.2
$\geq$4	63.6
$\geq$3	87.9
$\geq$2	100

Table 2: Percentage of reviewers in levels of expertise to the submissions recommended by our model.

5.4 Results on the New Dataset

Our proposed approach has been used to assist in the paper-reviewer matching process in a tier-1 conference of computer architecture. We evaluated our approach on a new dataset constructed with reviewers' feedbacks on their assigned submissions. Based on the optimal number of topics on the NIPS dataset, we set the number of common topics to be 10 in this experiment.

We report the percentage of reviewers whose reported expertise level falls in the given range in Table 2. We note that all recommendations made by our system are reasonable considering that all reviewers had expertise levels no lower than 2. The majority (87.9%) of reviewers reported that they were familiar with the topics of the submissions assigned to them, and 63.6% of the reviewers had deep understanding of the submissions.

6 Error Analysis

We perform a qualitative analysis on NIPS dataset to analyze the difference of different algorithms on expertise matching. For the clarity of our discussion, we sample a submission whose abstract is shown in Table 4. We consider five models: common topic modeling (CT), hidden topic modeling (HT), LDA, Doc2Vec and WMD. We list reviewers who were considered as top candidates for this submission by the five models in Table 3. For the analyses, we used research topics from the publications of the reviewers as well as their relevance scores assigned by human annotators (i.e., their TREC scores in NIPS dataset). Reviewers are sorted in decreasing order of their relevance to the submission by five models. For example, rank 1 corresponds to the highest relevance. In Table 3, We also present the rank of each reviewer given by the models.

Common topic model. According to the common topic model, reviewer 3, 4 and 5 are included as its top 3 recommendations. But we note that it ranks reviewer 3 higher than reviewer 4 and 5. The relevance scoring of common topic model is based on the relevance between the common topic "Bayesian method" and reviewers' profile. Since reviewer 3 is more focused on Bayesian model, it's topic-reviewer relevance is higher than reviewer 4 and 5 who have broader research interests beyond Bayesian model and more publications. One limitation of common topic model reflected in this case is that it does not capture the authority and experience of reviewers.

Hidden topic model. It incorrectly considered reviewer 1 more relevant to the submission compared to reviewer 5. We note that reviewer 5 works on a broad set of research topics ranging from Bayesian model to active learning. Since the hidden topic model extracts reviewer's topics based on the topic importance without any knowledge of the submission, it is likely that Bayesian model was not selected into representative hidden topics, which results in low relevance of reviewer 5 to the given submission.

LDA model. LDA assigns higher relevance to reviewer 1 than reviewer 4. Reviewer 1 used

Reviewer	TREC	Research topics	CT	HT	LDA	Doc2Vec	WMD
1	1	Speech recognition with Bayesian approach, Neural network	6	2	4	1	7
2	0	Online learning, Sequential prediction, Bayes point machine	10	10	8	9	1
3	2	Bayesian network, Variational Bayes estimation, Mixture models	1	4	3	2	9
4	3	Variational method, Bayesian learning, Markov model	2	3	9	4	8
5	3	Bayesian learning, Variational method, Active learning	3	5	1	7	6

Table 3: Examples of reviewers and their relevance to the submission ranked by different algorithms.

Dirichlet Process (DP) mixture models are candidates for clustering applications where the number of clusters is unknown a priori. [...] The speedup is achieved by incorporating kd-trees into a variational Bayesian algorithm for DP mixture [...]

Table 4: An example of abstract from a submission.

Bayesian approach, whereas it was not his research focus according to his publications. Reviewer 4 had done extensive research in general graphical models including Bayesian model. We observed that LDA fails to capture the relevance between graph model and Bayesian model since it ignores the semantic similarity between words.

Doc2Vec model. Doc2Vec assigned the highest relevance to reviewer 1 among all reviewers. The document representation it generates for reviewer 1's profile is similar to the representation for the submission, possibly because the key word "Bayesian" and "mixture" in the submission also occurs frequently in the profile. It suggests that Doc2Vec model might be limited to lexical overlap.

WMD. Reviewer 2 is included as WMD's top recommendation, whereas the research focus of reviewer 2 is sequential prediction which is irrelevant to the submission. Moreover, actually relevant reviewers 4 and 5 were excluded from WMD's top recommendations. This may have resulted from WMD's word-level similarity measure. Reviewer 2's publications had some lexical overlap with the submission (e.g., words "Bayes", "algorithm" and "learning", which have high frequency in the submission). WMD tends to assign high relevance scores due to such lexical overlap.

7 Future Work

This study used a basic version of a reviewer's profile to be the concatenation of the abstracts from their previous publications. A concrete direction for future work would be to consider enhancements in representing reviewers' profiles. Such efforts could consider, for instance, the temporal variation of research interests in order to capture the relevance of a given reviewer to a given topic based on the recency of the contributions to a given area. Other efforts could involve the use of a variable number of research topics for each reviewer and exploring ways to render reviewer profiles human interpretable.

8 Conclusion

We proposed an automated reviewer-paper matching algorithm via jointly finding the common research topics between submissions and reviewers' publications. Our model is based on word embeddings and efficiently captures the reviewer-paper relevance. It is robust to cases of vocabulary mismatch and partial topic overlap between submissions and reviewers – factors that have posed problems for previous approaches. The common topic model showed strong empirical performance on a benchmark and a newly collected dataset.

Acknowledgments

This work is supported by the IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network. We thank the EMNLP anonymous reviewers for their constructive suggestions.

References

PC Chairs ACL. 2019. Whats new, different and challenging in acl 2019? http://acl2019pcblog.fileli.unipi.it/?p=156. Accessed: 2019-05-19.

Krisztian Balog, Yi Fang, Maarten de Rijke, Pavel Serdyukov, Luo Si, et al. 2012. Expertise retrieval. *Foundations and Trends® in Information Retrieval*, 6(2–3):127–256.

Chumki Basu, Haym Hirsh, William W. Cohen, and Craig Nevill-Manning. 1999. Recommending papers by mining the web. In *In: Proceedings of the*

IJCAI99 Workshop on Learning about Users, pages 1–11.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 795–804.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Susan T. Dumais and Jakob Nielsen. 1992. Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, pages 233–244, New York, NY, USA. ACM.

Babak Falsafi, Mario Drumond, and Mark Sutherland. 2018. Isca'18 review process reflections. https://www.sigarch.org/isca18-review-process-reflections/. Accessed: 2019-05-19.

Hongyu Gong, Suma Bhat, and Pramod Viswanath. 2017. Geometry of compositionality. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Hongyu Gong, Tarek Sakakini, Suma Bhat, and Jinjun Xiong. 2018. Document similarity for texts of varying lengths via hidden topics. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2341–2351.

Weihua Hu and Jun'ichi Tsujii. 2016. A latent concept topic model for robust topic inference using word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 380–386, Berlin, Germany. Association for Computational Linguistics.

Jian Jin, Qian Geng, Haikun Mou, and Chong Chen. 2018. Author-subject-topic model for reviewer recommendation. *Journal of Information Science*, 45(4):554–570.

Jian Jin, Qian Geng, Qian Zhao, and Lixue Zhang. 2017. Integrating the trend of research interest for reviewer assignment. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1233–1241. International World Wide Web Conferences Steering Committee.

Ngai Meng Kou, Nikos Mamoulis, Yuhong Li, Ye Li, Zhiguo Gong, et al. 2015. A topic-based reviewer assignment system. *Proceedings of the VLDB Endowment*, 8(12):1852–1855.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*.

Charlin Laurent and Richard S. Zemel. 2013. The toronto paper matching system: An automated paper-reviewer assignment system. In *Proceedings of 30th International Conference on Machine Learning*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Baochun Li and Y Thomas Hou. 2016. The new automated ieee infocom review assignment system. *IEEE Network*, 30(5):18–24.

Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li. 2015. Hierarchical recurrent neural network for document modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 899–907.

Xiang Liu, Torsten Suel, and Nasir Memon. 2014. A robust model for paper reviewer assignment. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 25–32, New York, NY, USA. ACM.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv: 1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

David Mimno and Andrew McCallum. 2007. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 500–509, New York, NY, USA. ACM.

Jennifer Nguyen, Germn Snchez-Hernndez, Nria Agell, Xari Rovira, and Cecilio Angulo. 2018. A decision support tool using order weighted averaging for conference review assignment. *Pattern Recogn. Lett.*, 105(C):114–120.

David MW Powers. 2015. What the f-measure doesn't measure: Features, flaws, fallacies and fixes. *arXiv preprint arXiv:1503.06410*.

Simon Price and Peter A Flach. 2017. Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, 60(3):7079.

Yujie Qian, Jie Tang, and Kan Wu. 2018. Weakly learning to match experts in online community. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pages 3841–3847. AAAI Press.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (ma) and applications. WWW '15 Companion.

Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136.

Wenbin Tang, Jie Tang, and Chenhao Tan. 2010. Expertise matching via constraint-based optimization. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 34–41, Washington, DC, USA. IEEE Computer Society.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Jacob A Wegelin. 2000. A survey of partial least squares (pls) methods, with emphasis on the two-block case. *Technical Report 371, Department of Statistics, University of Washington*.

Pengtao Xie, Diyi Yang, and Eric Xing. 2015. Incorporating word correlation knowledge into topic modeling. In *Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: human language technologies*, pages 725–734.

Guangxu Xun, Yaliang Li, Wayne Xin Zhao, Jing Gao, and Aidong Zhang. 2017. A correlated topic model using word embeddings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 4207–4213. AAAI Press.