Review-Based Domain Disentanglement without Duplicate Users or Contexts for Cross-Domain Recommendation

Yoonhyuk Choi
Seoul National University
Seoul, Republic of Korea
younhyuk95@snu.ac.kr

Jiho Choi
Seoul National University
Seoul, Republic of Korea
jihochoi@snu.ac.kr

Taewook Ko
Seoul National University
Seoul, Republic of Korea
taewook.ko@snu.ac.kr

Hyungho Byun
Seoul National University
Seoul, Republic of Korea
notorioush2@snu.ac.kr

Chong-Kwon Kim
Korea Institute of Energy Technology
Naju, Republic of Korea
ckim@kentech.ac.kr

ABSTRACT

A cross-domain recommendation has shown promising results in solving data-sparsity and cold-start problems. Despite such progress, existing methods focus on domain-shareable information (overlapped users or same contexts) for a knowledge transfer, and they fail to generalize well without such requirements. To deal with these problems, we suggest utilizing review texts that are general to most e-commerce systems. Our model (named *SER*) uses three text analysis modules, guided by a single domain discriminator for disentangled representation learning. Here, we suggest a novel optimization strategy that can enhance the quality of domain disentanglement, and also debilitates detrimental information of a source domain. Also, we extend the encoding network from a single to multiple domains, which has proven to be powerful for review-based recommender systems. Extensive experiments and ablation studies demonstrate that our method is efficient, robust, and scalable compared to the state-of-the-art single and cross-domain recommendation methods.

CCS CONCEPTS

• **Information Systems** → **Recommender Systems**.

KEYWORDS

Cross-Domain Recommendation, Disentangled Representation Learning, Domain Adaptation, Textual Analysis

ACM Reference Format:
Yoonhyuk Choi, Jiho Choi, Taewook Ko, Hyungho Byun, and Chong-Kwon Kim. 2022. Review-Based Domain Disentanglement without Duplicate Users or Contexts for Cross-Domain Recommendation. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3511808.3557434

CIKM '22, October 17–21, 2022, Atlanta, GA, USA.

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557434

1 INTRODUCTION

With the rapid growth of e-commerce, recommender systems have become an obligatory tool for interconnecting customers with relevant items. Early schemes suffer from cold-start problems caused by data insufficiency. To tackle this problem, auxiliary information of data such as social relations [17], the trustworthiness of reviewers [1], item images [60] are exploited. Especially, textual or linguistic information such as reviews are commonly available, and many text-aided recommendation algorithms have been introduced [9, 13, 15, 50, 51, 54, 57, 66]. For example, some studies infer the preferences of users by applying natural language processing (NLP) techniques such as topic modeling [2, 38]. Recently, some apply convolutional neural network (CNN) on textual reviews [66], others [9, 15] further utilize attention mechanism to deal with important reviews. Many text analysis modules are considerable, but here, we employ a simple text analysis module [66] to focus on a cross-domain recommendation scenario.

Along with the text-based recommender systems, numerous cross-domain recommendation [18, 26, 62] and transfer learning [24, 58, 61] approaches have been introduced. They leverage the information learned from source domains to improve the quality of recommendation in a target domain. Some of them suggest a fine-tuning network, but they may suffer from catastrophic forgetting [10]. Context mappings [23, 32, 35, 62] are another branch of transfer learning, which map shareable information from a source to the target domain. However, these approaches require the same contexts or overlapped users, which may restrict their applicability severely [28]. In a real-world dataset, this kind of information might be scarce, and thus, their contributions might be insignificant, which still requires further investigation.

To solve such constraints, we focus on another branch of methods that are independent of specific users or contexts [7, 29, 64]. However, it is still challenging to achieve knowledge transfer between heterogeneous domains (e.g., totally different users or items), while debilitating noises. For one solution, domain adaptation (DA) [4, 20, 30, 31] minimizes source risk as well as H-divergence, capturing domain-indiscriminative information. Nonetheless, these methods are highly dependent on the consistency (maximum mean discrepancy) between domains. Though disentangled representation learning algorithms [6, 22, 34, 44] concurrently extract domain-specific, and domain-invariant knowledge to distill pertinent knowledge from multiple seemingly counterproductive domains, they also suffer from measuring theoretical bound and tractability of a log-partition function [19, 37].

In this paper, we suggest a novel method of disentangled representation learning that utilizes textual reviews common to most e-commerce systems. Here, we focus on debilitating noises of another domain and claim that this can be achieved by reinforcing the discrimination power of a domain discriminator. Specifically, using the connections between mutual information (MI) estimation and DA [27, 43, 48, 65], we integrate the role of the MI estimator with a single domain discriminator. Thus, the domain-shareable information is further utilized as an input for the domain discriminator, which can reinforce the discrimination power. In terms of adversarial training, a well-trained discriminator guides three types of FEs to secure robustness and improve the recommendation quality for both domains. We perform extensive experiments on the publicly available dataset to compare our model under single and cross-domain recommendation scenarios. The quantitative and qualitative analysis demonstrates the superiority of our suggestions. The contributions can be summarized as follows.

- We propose *SER* that adaptively disentangles features depending on the characteristics of the source and target domains, where the novel optimization strategy leads to a promising solution for a recommendation. Consequently, our approach is also applicable to heterogeneous scenarios, in case the source and target domain have less similar characteristics (e.g., non-overlapping users or contexts).
- Our approach is comprehensive that is closely integrated with text-based feature extraction. Unlike previous cross-domain recommendation schemes that require duplicate entities from heterogeneous domains, we focus on retrieving review information that is common for most domains.
- We perform extensive experiments to answer the important research questions described in Section 5. The results indicate the superiority and robustness of our proposed method.

2 RELATED WORK

In this section, we describe an overview of existing recommender systems; text-aided single-domain recommendations, extending to a cross-domain scenario, and applying domain adaptation to achieve disentangled representation learning, which can enhance the knowledge transfer between domains.

2.1 Text-aided Recommender System

Textual information is the most popular side information and many text-based methods [12, 13, 15, 66] have been proposed recently. Previous techniques simply integrate DNN-based feature extraction with MF for a rating prediction, while [66] utilizes two parallel CNNs, and [33] adopts Gated Recurrent Units for review analysis. Attention mechanisms are widely used also to pinpoint useful words and reviews [9, 15, 49, 53]. Even though prior works show the usefulness of textual information, the limitation of review information due to the limited size of training data, and the irrelevance of reviews toward target items have been raised also [47, 63]. To tackle this problem, we now introduce recent algorithms for cross-domain recommender systems which adopt supplementary domains.

2.2 Cross-Domain Recommendation (CDR)

CDR utilizes information from source domains to alleviate the cold-start problem in the target domain. Especially, most of them assume embedding-based or rating pattern-based transfer scenario [67]. In detail, early studies [16, 36] adopt a feature mapping technique that requires overlapped users. For example, RC-DFM [18] applies Stacked Denoising Autoencoder (SDAE) to each domain, where the learned knowledge of the same set of users is transferred from a source to the target domain. More recently, to overcome the requirements of overlapped users, there was an attempt [55, 64] that employs similar users for feature mapping. However, they implicate the limitation of debilitating noises from a source domain. In this paper, we adopt review texts that are common to most e-commerce systems without requiring overlapping users or contexts. Further, we suggest a novel disentangled representation learning for a knowledge transfer between domains which can reduce noises from a source domain. We now introduce some traditional methods for disentangled representation learning below.

2.3 Disentangled Representation Learning

Recently, many efforts have been dedicated to capturing domain-shareable information. Especially, with the powerful mechanism of adversarial training, DA has been adopted for various fields; question answering tasks [46], and recommendation scenarios [5, 62]. Some adopt textual reviews under no user or context overlap for a cross-domain recommendation [59]. However, these approaches only focus on domain-shareable knowledge, while ignoring domain-specific ones. MMT [29] further captures domain-specific knowledge, but the disentanglement between domain-common features is not applied. Though DADA [44] introduced domain-agnostic learning, the domain discriminator is only utilized for the extraction of a domain-common feature. Since the domain-specific knowledge is solely guided using MI minimization [3], they have shown to implicate some defects [37]. In this paper, we clear the aforementioned limitations, adopting a framework for the simultaneous extraction of domain-specific and domain-common knowledge by integrating MI estimator with domain discriminator. Consequently, our model enhances the quality of disentangled representation learning without requiring overlapping users or contexts for a recommendation.

Table 1: Notations

Symbol	Explanation
D^s, D^t	Set of source and target domain dataset
$u, i, y_{u,i}, r_{u,i}$	Set of user, item, rating, and individual review
R_u, R_i	Set of aggregated reviews for user and item
p_u, q_i	Rating embedding of user and item
O	Extracted feature of R_u, R_i
I	Extracted feature of $r_{u,i}$
$\widehat{d}$	Predicted domain label of O
$\widehat{y}_E, \widehat{y}_I$	Predicted label of E and I
MI	Mutual information
N_s, N_t	Mini-batch training samples of s and t
ϕ	Word embedding function
$\oplus$	Concatenation operator
$\mathcal{L}$	Loss function

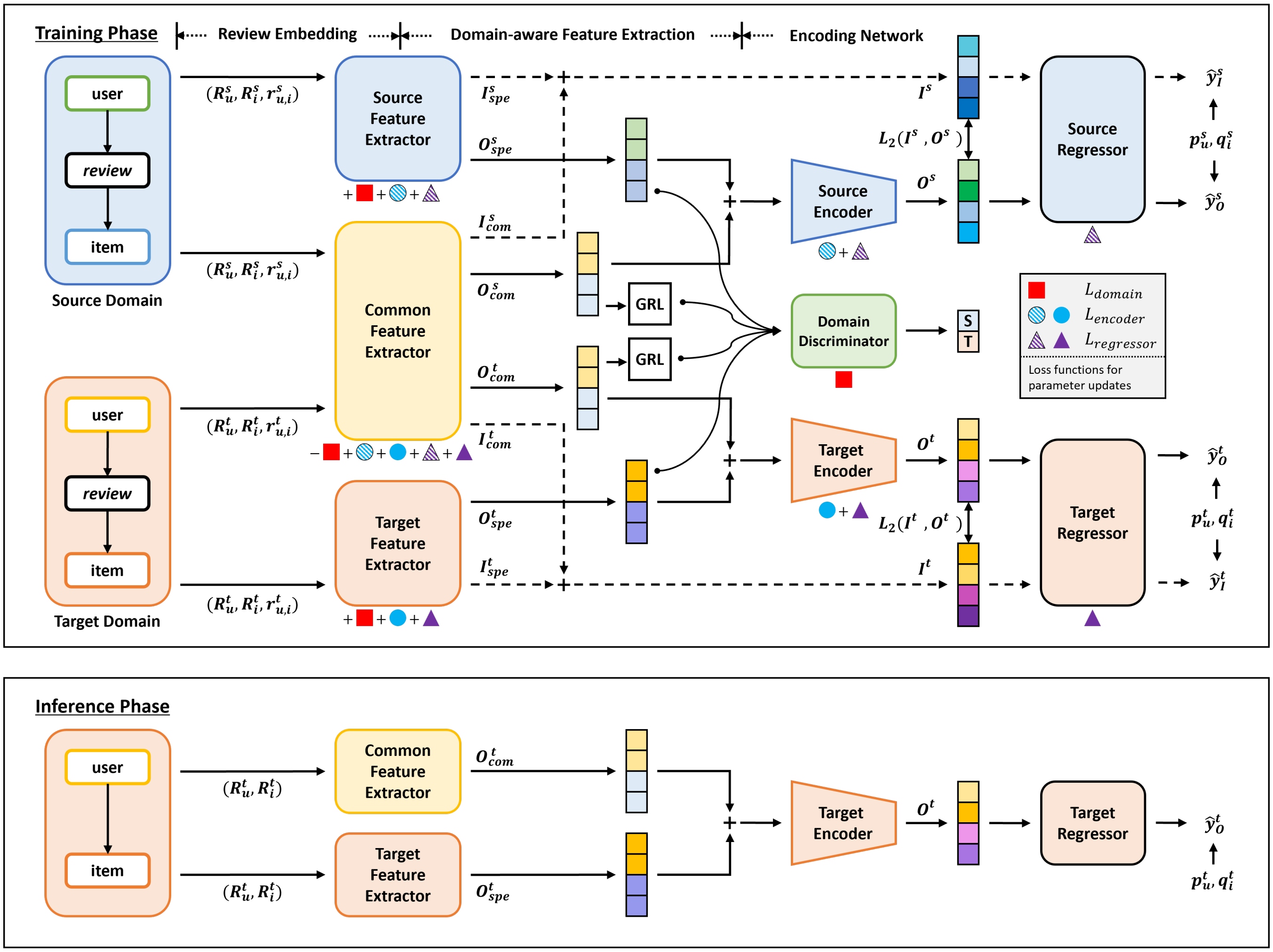

Figure 1: The overall framework of SER for a cross-domain recommendation scenario

3 PROBLEM FORMULATION

We formulate the cross-domain recommendation task as follows. Assume two datasets, D^s and D^t, be the data of the source and target domains. Each of them consists of tuples, $(u, i, y_{u,i}, r_{u,i})$ which represents an individual review $r_{u,i}$ provided by a user u for an item i with rating $y_{u,i}$. Our goal is to improve the recommendation quality in a target domain with the aid of a source domain. To achieve this, we suggest a domain-aware knowledge transfer module with encoding networks. The overall notations can be seen in Table 1.

4 OUR APPROACH

Figure 1 describes our model with the following key components:

- **Review embedding layer** Vectorize the aggregated and individual review using pre-trained word embedding function, which can be word2vec[1] [39] or GloVe[2] [45].
- **Domain-aware feature extraction** Integrated with domain discriminator, three feature extractors (FEs) capture domain-aware knowledge from vectorized aggregated reviews. Especially, the common FE acts as a knowledge transfer module and utilizes both source and target domain data.
- **Encoding network** The encoding networks generate a single vector from vectorized aggregated reviews by aligning them with the latent of an individual review.

We now illustrate the details of each module precisely.

4.1 Review Embedding Layer

Given the input data $(u, i, y_{u,i}, r_{u,i})$, we first ensemble the reviews of user and item. For example, given user u, all reviews written by her are treated as a single document R_u. The item reviews R_i can be retrieved similarly. However, for each training sample, we exclude an individual review $r_{u,i}$ that u has written after purchasing i since it cannot be used during the inference phase. Here, we do not consider a sequence (or time) of their purchasing histories.

[1]https://code.google.com/archive/p/word2vec
[2]https://nlp.stanford.edu/projects/glove

We utilize first n words in R and apply pre-trained word embedding function [39, 45] for vectorization. The words are mapped to c-dimensional vectors, and column-wisely concatenated to form document embedding $V = \phi(w_1) \oplus \phi(w_2) \oplus ... \oplus \phi(w_n)$, where ϕ and $\oplus$ is a embedding and concatenation operation.

4.2 Domain-aware Feature Extraction

As described in Figure 1, we utilize three types of feature extractors (FEs) to separate the domain-specific and domain-common knowledge. In this paper, each FE adopts the widely used text analysis method [66]. More specifically, using the document embedding V, three convolutional *FEs* produce outputs, followed by a non-linear activation function (e.g., ReLU) with a row-wise max-pooling layer. Finally, by concatenating the outputs of stacked convolution layers, we can derive a unified embedding O from user and item reviews (please refer to [66] for text convolutions). Similarly, we can retrieve the embedding of an individual review I using three FEs.

To summarize, the domain-specific FEs (source and target) generate O^s_{spe}, I^s_{spe} and O^t_{spe}, I^t_{spe} for each domain. Also, the domain-common FE extracts O^s_{com}, I^s_{com}, and O^t_{com}, I^t_{com} using inputs from source and target domain, respectively. Here, the superscript s, t acts as an identifier of two domains. Though we assume three FEs, they may contain similar information without any constraints. To achieve a better disentanglement, we suggest an optimization strategy from two perspectives: (1) for the extraction of domain-common knowledge, we employ domain adaptation to serve common FE as an intermediate agent for knowledge transfer. (2) for the disentanglement of domain-specific features, we suggest a novel constraint for adversarial training.

4.2.1 ***Extracting domain-common features.***

For a knowledge transfer between domains, we focus on domain adaptation (DA) [20] that has proven to be effective in case a source domain has a richer label than the target [4]. DA employs domain discriminator with Gradient Reversal Layer (GRL) to reduce the cross-domain discrepancy. The discriminator gives additional penalties (or loss) to common FE for capturing domain-discriminative information, which has proven to be effective for debilitating noises. Here, we adopt two layers of a fully-connected neural network as domain discriminator F_{disc}, which takes the output of common FE as below:

$$\widehat{d}^s_{com} = F_{disc}(O^s_{com}), \;\; \widehat{d}^t_{com} = F_{disc}(O^t_{com}). \tag{1}$$

Here, $\widehat{d}_{com}$ stands for predicted domain probability of common feature O_{com}, where the error can be calculated through binary cross-entropy loss between true label d_{com} as below:

$$\mathcal{L}^s_{com} = -\frac{1}{N_s}\sum_{s=1}^{N_s} log(1-\widehat{d}^s_{com}), \;\; \mathcal{L}^t_{com} = -\frac{1}{N_t}\sum_{t=1}^{N_t} log(\widehat{d}^t_{com}). \tag{2}$$

The true label d_{com} is a binary value, $\{0, 1\}$ for source and target domain. Here, N is the mini-batch training samples from two domains. GRL is positioned between common FE and domain discriminator (please refer to Figure 1), multiplies a negative constant during back-propagation. Consequently, the common FE is reinforced to capture domain-indiscriminative knowledge to fool discriminator. Nonetheless, DA itself implicates some limitations, which can be sensitive to domain divergence, and prohibitive applicability [11].

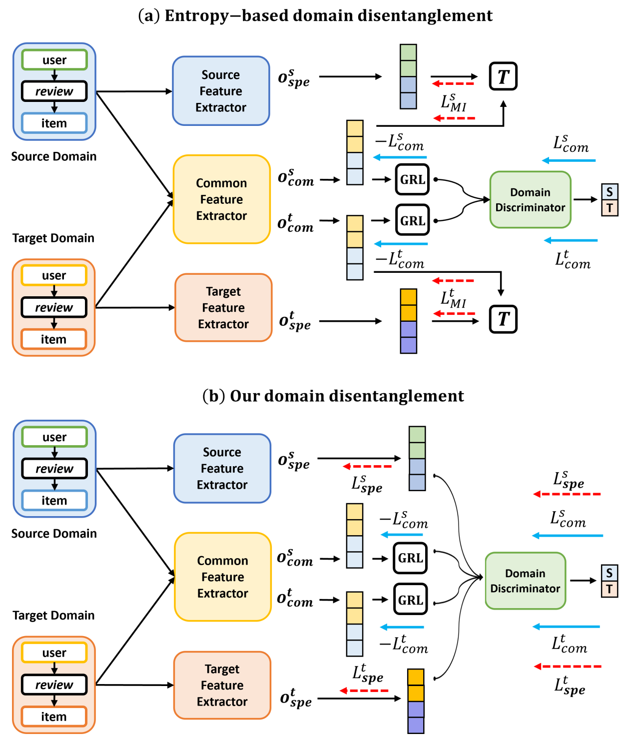

Figure 2: (a) Conventional domain disentanglement [44] using mutual information (MI) estimator T, and (b) ours integrating MI estimator with domain discriminator

4.2.2 ***Disentangling domain-specific features.***

To deal with the aforementioned problems, many approaches endeavored for the parallel extraction of domain-specific knowledge. One of the promising solutions is disentangled representation learning, which can secure pairwise independence between generated factors [22, 34, 41]. For example, DSN [6] optimizes the source, common, and target FEs by securing orthogonality between domain-specific and domain-common features. More recently, estimating mutual information (MI) between extracted features has been proposed [14, 44]. Specifically, IDEL [14] suggests a novel MI bound to retrieve disentangled representation of texts. DADA [44] captures domain-specific knowledge using MI estimator (see Figure 2-a). In terms of independence maximization, DADA aims to minimize MI between domain-specific O_{spe} and domain-common O_{com} features as follows:

$$\begin{aligned}\mathcal{L}^s_{MI} &= \frac{1}{N_s}\sum_{s=1}^{N_s} T(O^s_{spe}, O^s_{com}) - log(\frac{1}{N_s}\sum_{s=1}^{N_s} e^{T(O^s_{spe}, O'^s_{com})}),\\ \mathcal{L}^t_{MI} &= \frac{1}{N_t}\sum_{t=1}^{N_t} T(O^t_{spe}, O^t_{com}) - log(\frac{1}{N_t}\sum_{t=1}^{N_t} e^{T(O^t_{spe}, O'^t_{com})}).\end{aligned} \tag{3}$$

Since MI cannot be directly calculated, they adopt approximation theory MINE [3] with a neural network T:

$$T(x, z) = F_3(\sigma(F_1(x) + F_2(z))). \tag{4}$$

With two inputs x and z, a fully-connected layer F with non-linear activation σ estimates MI. Back into Equation 3, as an input of T, (O_{spe}, O_{com}) pairs are sampled from joint distribution, while the same number of (O_{spe}, O'_{com}) pairs are from marginals (please refer to [44] for more details). As described in Figure 2-a, domain-specific FEs are trained to minimize $\mathcal{L}^s_{MI}, \mathcal{L}^t_{MI}$ (red arrows), while GRL guides a common FE (blue arrows) through Equation 2. Nonetheless, MINE implicates the following limitations: a statistical drawback of measuring bounds [37], the estimators become intractable even for a known distribution as the dimension increases [19] and is highly dependent on the number of symbols and the size of dataset [42].

Using the observations [27, 43, 48, 65], minimizing $\mathcal{H}$-divergence is closely related to MI maximization. We integrate the MI estimator with a domain discriminator from the following perspectives: (1) maximizing domain loss (equal to MI maximization) for domain-common features, while (2) minimizing domain loss for domain-specific outputs. To achieve this, we exclude GRL between domain-specific FEs and domain discriminator (please refer to Figure 2-b), to guide these FEs to generate domain-discriminative knowledge. The advantages of this suggestion are as follows. Firstly, domain-specific features (O^s_{spe}, O^t_{spe}) are additionally utilized as an input of a domain discriminator, where they reinforce the discriminating power of a domain discriminator. Then, a discriminator well distinguishes the origin of the input vectors, leading three types of FEs better extract domain-aware knowledge. Secondly, our model can disentangle inter-domain outputs $(O^s_{spe} \leftrightarrow O^t_{spe})$, while the MI-based methods only separate intra-domain ones $(O^s_{spe} \leftrightarrow O^s_{com}$ or $O^t_{spe} \leftrightarrow O^t_{com})$. Lastly, the large values of MI can mislead the parameters, while the ground truth of our model lies between 0 and 1. Using Equation 1 and 2, we can define domain-specific losses as:

$$\begin{gathered}\widehat{d}^s_{spe} = F_{disc}(O^s_{spe}),\ \widehat{d}^t_{spe} = F_{disc}(O^t_{spe}),\\ \mathcal{L}^s_{spe} = -\frac{1}{N_s}\sum_{s=1}^{N_s} log(1-\widehat{d}^s_{spe}),\ \mathcal{L}^t_{spe} = -\frac{1}{N_t}\sum_{t=1}^{N_t} log(\widehat{d}^t_{spe}).\end{gathered} \tag{5}$$

Finally, integrating Equation 2 and 5, the overall domain loss function can be abbreviated as below:

$$\mathcal{L}_{dom} = a(\mathcal{L}^s_{com} + \mathcal{L}^s_{spe}) + (1-a)(\mathcal{L}^t_{com} + \mathcal{L}^t_{spe}). \tag{6}$$

$a = \frac{N_s}{N_s+N_t}$ controls the weight of source and target losses based on the size of two domains [4]. For training, the domain discriminator is updated using Equation 6. Then, GRL is applied for domain-common losses $-a\mathcal{L}^s_{com}, -(1-a)\mathcal{L}^t_{com}$ (blue arrows in Figure 2-b) to update common FE, while domain-specific losses $a\mathcal{L}^s_{spe}, (1-a)\mathcal{L}^t_{spe}$ (red arrows) without GRL update domain-specific FEs.

4.2.3 Theoretical analysis. Referring to the estimated bounds of semi-supervised DA [30], we can infer the theoretical bounds of DADA [44] and our SER. Since both of them captures domain-common knowledge through DA, the difference lies in domain disentanglement strategy and their performance. We first discuss the classification errors between true domain labels and predicted ones. DADA adopts dual representations of KL-divergence, which has proven [3] that a neural network T_θ with parameter θ satisfies the conditions below as the number of samples n goes to infinity:

$$|\widehat{I(X;Z)}_n - I_\theta(X,Z)| \leq \epsilon \tag{7}$$

Similarly, our SER adopts stochastic gradient descent with binary domain labels, which has proven to [21] converge under the arbitrary sampling scenario. Based on these results, both approaches well predict the true labels within ϵ. However, it is notable that the value of MI ranges from 0 to infinity. Referring to Equation 3, a large domain-specific loss might impede the stable convergence of three FEs. Comparatively, we adopt binary cross-entropy loss of Equation 6, and thus, the loss rarely diverges even for completely domain-indistinguishable features ($\widehat{d}^s_{spe} \approx \widehat{d}^t_{spe} \approx 0.5$). In Section 5.3, we scrutinize two methods that use Equation 3 and 5.

4.3 Encoding Network and Regressor

In this section, except for Equation 10 and 12, we exclude domain identifier d to simplify notation.

Encoding network This layer employs a transfer network [56] (MLP) to align the latent of aggregated reviews with an individual review, which is proposed by TransNet [8]. Using this strategy, FEs are trained to mimic the latent of an individual review that a user has written after purchasing a specific item (please refer to [8] for more details). Here, we extend this mechanism from a single to cross-domain recommendation scenario, while most text-aided schemes [9, 13, 15, 54, 66] underestimate this kind of information. Since we have obtained the latents of aggregated review O_{spe}, O_{com}, the encoded representation can be derived as below:

$$O = F_{enc}(O_{spe} + O_{com}) \tag{8}$$

, where F_{enc} is an encoding network with two layers. Similarly, the representation of an individual review can be retrieved as follows:

$$I = I_{spe} + I_{com}. \tag{9}$$

To align the two vectors $O \leftrightarrow I$ of Equation 8 and 9, for each domain $d \in \{s, t\}$, we adopt Euclidean distance as a loss function:

$$\mathcal{L}^d_{enc} = \frac{1}{N_d}\sum_{d=1}^{N_d} ||O^d - I^d||^2_2. \tag{10}$$

Regressor Finally, we can utilize two representations O, I to predict a recommendation score through regressor F_{reg} (MLP). Here, we further apply a widely used latent factor model [52], where p_u and q_i stand for the embedding of user and item:

$$\widehat{y}_I = F_{reg}(I) + p_u \cdot q_i^T,\ \widehat{y}_O = F_{reg}(O) + p_u \cdot q_i^T. \tag{11}$$

We utilize the prediction of an individual review $\widehat{y}_I$, which can guide the regressor precisely. The regression loss is defined in Equation 12, which is the Mean Squared Error (MSE) between two predicted scores $\widehat{y}_I, \widehat{y}_O$ and true label y in domain d as below:

$$\mathcal{L}^d_{reg} = \frac{1}{2N_d}\sum_{d=1}^{N_d}\left((\widehat{y}^d_I - y^d)^2 + (\widehat{y}^d_O - y^d)^2\right). \tag{12}$$

4.4 Optimization and Inference

Optimization We jointly optimize SER by minimizing the weighted sum of three losses defined in Equation 6, 10, 12:

$$\min_\theta \mathcal{L} = \alpha\mathcal{L}_{dom} + \beta(\mathcal{L}^s_{enc} + \mathcal{L}^t_{enc}) + \gamma(\mathcal{L}^s_{reg} + \mathcal{L}^t_{reg}) + \delta||\theta||. \tag{13}$$

In Figure 1, we describe how the gradients are propagated with three basic shapes (the horizontal stripe denotes a source loss). The hyper-parameters α, β, and γ balance the weight of three losses. We set $\alpha = 0.1$, $\beta = 0.05$, and $\gamma = 1$ through a grid search. θ denotes parameters of our model, where δ is a regularization term. We adopt early stopping under 300 iterations with Adam optimizer and learning ratio as $lr = 1e^{-4}$.

Inference During inference phase (lower side of Figure 1), SER recommends an item using aggregated reviews of user R_u^t and item R_i^t. Then, the common and target FEs are applied for R_u^t and R_i^t, retrieving their features O_{com}^t, O_{spe}^t, respectively. Finally, a target encoder generates meaningful feature O^t, followed by a regressor.

4.5 Computational Complexity Analysis

In this section, we analyze the computational cost of our model. Firstly, let us assume parameters of vanilla text analysis model DeepCoNN [66] as $A+B$, which adopts a single FE (A) and regressor (B). The running time can be approximated as $O((A+B) \cdot N_t \cdot e)$, where N_t is the size of training samples and e is a training epoch.

Our model further utilizes a domain discriminator with encoding networks. With the slight abuse of notation, let us assume the number of additional parameters as C. Also, we utilize the source domains that contain N_s samples, the computational complexity can be abbreviated as $O((3A+2B+C) \cdot (N_s+N_t) \cdot e)$. To summarize, we can infer that a cost linearly depends on the input size of the source domain N_s, which can be quite efficient.

5 EXPERIMENTS

We aim to answer the following research questions:

- **RQ1:** Can SER enhance the recommendation quality compared to the state-of-the-art approaches?
- **RQ2:** Does SER effectively disentangles features for a cross-domain recommendation?
- **RQ3:** How much does the domain-aware feature extractor and encoding network contribute to the overall quality?
- **RQ4:** Is a domain-common feature necessary for CDR? If so, how does the knowledge is shared across domains?

5.1 Dataset and Experimental Setup

Dataset description We systematically conduct investigations with publicly available dataset *Amazon*[3] and *Yelp*[4]. The target domain includes the following four categories of Amazon 5-core: *Office Products, Instant Video, Automotive, Patio Lawn and Garden*. We designate the source domain with more interactions [4]. Three categories from *Amazon*: *Baby, Kindle Store (KS), Toys and Games (TG)* and one another from *Yelp*. Here, *Yelp* data is used to show the effect of excluding duplicate users. The statistical details of the dataset are summarized in Table 2. Though some studies employ a source domain of seemingly relevant categories, it is natural to ask: *how can we choose the most relevant domain? if they are truly relevant, how does the knowledge transferred across domains?* Here, we focus on the second question rather than figuring out relevant domains. Multi-source adaptation is also considerable, but here, we leave it for future work.

Environmental settings and baselines Following previous studies, we divide the target domain dataset into three parts: 80 percent for training, 10 percent for validation, and another 10 percent for testing. We repeatedly consume a target domain w.r.t. the size of a source domain for one iteration. Early stopping is applied based on the validation score for 300 iterations. The word embedding is set to 100 with 100 convolution filters of size $\mathbb{R}^{5\times100}$. Now, we introduce state-of-the-art single and cross-domain methods below.

Table 2: Statistical details of the dataset

	Dataset	# users	# items	# reviews
Source	Baby	19,445	7,050	160,792
	Kindle Store (KS)	68,223	61,934	982,619
	Toys and Games (TG)	19,412	11,924	167,597
	Yelp	1.9 M	0.2 M	8.1 M
Target	Office Products	4,905	2,420	53,258
	Instant Video	5,130	1,685	37,126
	Automotive	2,928	1,835	20,473
	Patio Lawn and Garden	1,686	962	13,272

Single-Domain Approaches:

- **PMF** [40] is a classical probabilistic matrix factorization method, which only utilizes rating information.
- **NeuMF** [25] combines deep neural networks with a probabilistic model, and use rating information only.
- **DeepCoNN** [66] leverages review texts for rating prediction. They jointly encode the latent of user and item, respectively.
- **NARRE** [9] maintains the overall architecture of *DeepCoNN*, while employing attention mechanism. They firstly measure the usefulness of each review using attention scores.
- **AHN** [15] proposes a hierarchical attention mechanism: from a sentence to review-level attention for a better representation learning. It achieves state-of-the-art performance for a review-based single-domain recommendation.

Cross-Domain Approaches:

- **DANN** [20] proposes the seminal domain adversarial technique that extracts domain-common features from two different domains. Here, review texts are embedded as 5,000-dimensional feature vectors.
- **DAREC** [62] assume the same set of users between two domains. For each domain, the high-dimensional rating vectors are mapped to low-dimensional feature vectors using AutoEncoder, followed by a domain discriminator. They extract shareable rating patterns between two domains.
- **DDTCDR** [32] assumes that if two users have similar preferences in a source domain, it should be preserved in a target domain through an orthogonal mapping function.
- **RC-DFM** [18] suggest fusing review texts with rating information. With SDAE, they adequately preserve the latent features with rich semantic information. Under our experimental setting, we fine-tune the text convolution layer.
- **CATN** [64] extracts multiple aspects from reviews. For a knowledge transfer between domains, they assume an aspect correlation matrix with an attention mechanism.
- **MMT** [29] suggest component-wise transfer mechanism, while also preserving domain-specific modules. Here, we assume the text convolution layer as a knowledge transfer module and fine-tune the parameters in a target domain.

[3]http://jmcauley.ucsd.edu/data/amazon/
[4]https://www.yelp.com/dataset

Table 3: MSE (↓) on four target domain dataset. Bold and underline indicate 1^{st} and 2^{nd} best.

Target domain	Office Product				Instant Video				Automotive				Patio Lawn and Garden			
Source domain	Baby	KS	TG	Yelp	Baby	KS	TG	Yelp	Baby	KS	TG	Yelp	Baby	KS	TG	Yelp
PMF	1.085				1.129				1.162				1.177			
NeuMF	0.974				1.014				1.087				1.143			
DeepCoNN	0.902				0.949				0.979				1.128			
NARRE	0.863				0.914				0.887				1.108			
AHN	0.859				0.892				0.863				1.094			
DANN	0.966	0.939	0.943	1.118	0.986	0.946	0.987	1.147	0.946	0.881	0.945	1.183	1.129	1.189	1.199	1.395
DAREC	0.989	0.988	0.972	0.994	1.060	1.045	1.043	1.073	1.001	0.997	0.993	1.004	1.123	1.151	1.131	1.150
DDTCDR	0.954	0.947	0.926	0.965	0.974	0.981	0.967	0.988	0.961	0.959	0.954	0.969	1.109	1.111	1.105	1.133
RC-DFM	0.834	0.839	0.828	0.841	0.878	0.855	0.868	0.872	0.792	0.800	0.794	0.802	1.094	1.096	1.109	1.112
CATN	0.875	0.872	0.873	0.876	0.915	0.906	0.892	0.919	0.824	0.831	0.826	0.837	1.141	1.144	1.129	1.149
MMT	0.815	0.820	0.822	0.856	0.862	0.855	0.878	0.871	0.818	**0.798**	0.800	0.833	1.116	1.099	1.094	1.117
SER	**0.789**	**0.815**	**0.810**	**0.806**	**0.852**	**0.833**	**0.855**	**0.847**	**0.785**	**0.798**	**0.769**	**0.784**	**1.028**	**1.029**	**1.039**	**1.033**

Table 4: nDCG@5 (↑) on four target domain dataset. Bold and underline indicate 1^{st} and 2^{nd} best.

Target domain	Office Product				Instant Video				Automotive				Patio Lawn and Garden			
Source domain	Baby	KS	TG	Yelp	Baby	KS	TG	Yelp	Baby	KS	TG	Yelp	Baby	KS	TG	Yelp
PMF	0.737				0.759				0.764				0.770			
NeuMF	0.756				0.788				0.781				0.776			
DeepCoNN	0.856				0.840				0.816				0.842			
NARRE	0.861				0.872				0.851				0.844			
AHN	0.874				0.879				0.862				0.878			
DANN	0.843	0.847	0.840	0.829	0.851	0.849	0.846	0.835	0.844	0.858	0.836	0.831	0.830	0.823	0.818	0.812
DAREC	0.859	0.842	0.835	0.827	0.844	0.848	0.841	0.823	0.865	0.872	0.872	0.860	0.854	0.852	0.861	0.835
DDTCDR	0.854	0.853	0.860	0.847	0.852	0.858	0.849	0.840	0.877	0.874	0.881	0.865	0.846	0.851	0.849	0.839
RC-DFM	0.875	0.871	0.880	0.869	0.890	0.884	0.881	0.879	0.884	0.895	0.899	0.902	0.878	0.873	0.871	0.879
CATN	0.869	0.865	0.871	0.842	0.873	0.857	0.860	0.873	0.866	0.863	0.872	0.875	0.864	0.861	0.858	0.854
MMT	0.881	0.874	0.870	0.883	0.888	0.885	0.876	0.883	0.886	0.896	0.892	0.877	0.867	0.869	0.871	0.871
SER	**0.891**	**0.885**	**0.888**	**0.889**	**0.896**	**0.892**	**0.889**	**0.895**	**0.892**	**0.901**	**0.908**	**0.913**	**0.889**	**0.882**	**0.885**	**0.883**

5.2 Results and Discussion (RQ1)

For quantitative analysis, we assume two metrics: Mean Squared Error (MSE) for a rating prediction and normalized Discounted Cumulative Gain (nDCG@5) for ranking-based recommendations.

Utilizing additional domain generally enhances recommendation quality Table 3, 4 shows the performance of our model and previous methods based on MSE and nDCG@5 score. For single-domain methodologies, it is not surprising that rating-based PMF and NeuMF performed worse than other review-based methods, which indicates the usefulness of textual information. Among review-based methods, NARRE and AHN outperform DeepCoNN with an attention mechanism. Nonetheless, it is noticeable that textual-based CDR (RC-DFM, CATN, MMT) generally outperforms single-domain methods under our experimental scenario.

The quality of CDR can be degraded w.r.t. domain discrepancy To test a domain discrepancy, we adopt DANN and MMT, where they transfer knowledge independent of user overlapping. For DANN, the result varies significantly w.r.t. the selection of source domain. Rather, MMT secures stability through domain-specific information (rating). An important thing is that a CDR does not always achieve the best performance. In the case of *Patio Lawn and Garden*, AHN shows the best result among all baselines. It indicates that utilizing additional domains without debilitating noises can degrade the recommendation quality. Comparatively, SER manages to achieve stable performance, where the domain disentanglement alleviates noises of the auxiliary domain.

Instead of overlapping users, review-based knowledge transfer further improves the recommendation quality For example, RC-DFM integrates textual data with rating information, which significantly outperforms DAREC and DDTCDR. However, the problem is that all these methods require duplicate users for a knowledge transfer. This can be quite restrictive for the selection of a source domain, and some argued [28] that user-based knowledge transfer has a limited impact. Even excluding duplicate users (selecting $Yelp$), their performance varies quite insignificant (please refer to [18] for more details) compared to the other methods (CATN, MMT, SER). Comparatively, MMT and SER (review-based transfer) improve their score even by selecting *Yelp* as a source, which fairly is different from *Amazon* dataset. Though CATN utilizes auxiliary reviews of another user, does not show outstanding performance in our experiments. Further, MMT shows lower performance than SER because of the limitation of transfer learning and the absence of review disentanglement. Using the above results, we show that SER surpasses the SOTA methods without duplicate users or contexts.

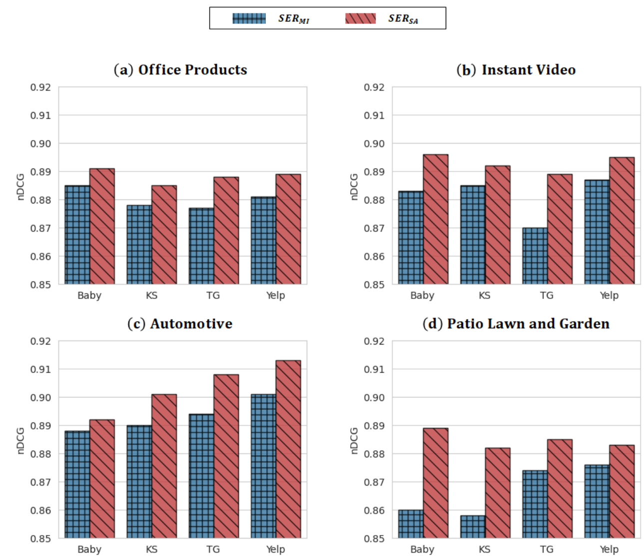

Figure 3: To show the influence of domain loss function, we describe nDCG@5 (↑) of two SER variants; SER with MI minimization SER_{MI}, and our proposed model SER_{SA}

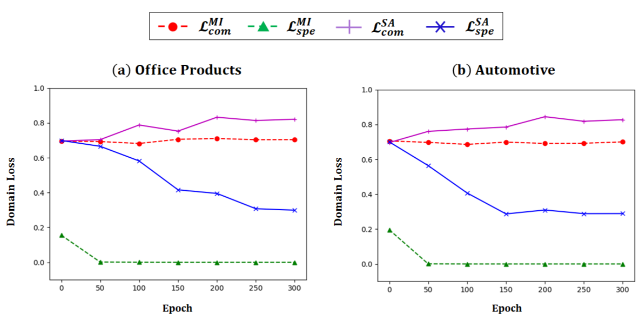

Figure 4: For two SER variants; SER_{MI} and SER_{SA}, we plot their domain losses *w.r.t.* training epoch

5.3 Analysis of Domain Discriminator (RQ2)

In this section, we compare state-of-the-art the domain disentanglement strategy [44] with our model, while numerous optimization methods can be considered [6, 22, 34]. We assume two different models; (1) SER_{MI} minimizes mutual information [44], while other modules remain the same as *SER*, and (2) our original SER_{SA}. To implement SER_{MI}, we utilize the publicly available source *code*[5]. Specifically, we train them to minimize the absolute value of estimated MI (please refer to Equation 3).

Analysis of the recommendation quality *w.r.t.* domain loss To scrutinize how the domain loss affects the overall performance, we measure the nDCG@5 of two variants using four target domain data, which is illustrated in Figure 3. Here, we notice that our SER_{SA} slightly outperforms SER_{MI} for all CDR scenarios. We insist that these results are closely related to domain losses. In Figure 4, using *Baby* as a source with two target domain datasets (*Office Products, Automotive*), we plot the variation of domain losses for SER_{MI} and SER_{SA} *w.r.t.* training epoch. Specifically, (1) for domain-common loss, both of them adopts adversarial training of Equation 2, $\mathcal{L}^{MI}_{com} = \mathcal{L}^{SA}_{com}=\mathcal{L}^{s}_{com} + \mathcal{L}^{t}_{com}$. (2) instead, their domain-specific loss ($\mathcal{L}^{MI}_{spe}$, $\mathcal{L}^{SA}_{spe}$) differs. SER_{MI} adopts Equation 3, while SER_{SA} utilizes Equation 5. Referring Figure 4, for SER_{MI}, $\mathcal{L}^{MI}_{com}$ (red line) rarely changes, while $\mathcal{L}^{MI}_{spe}$ (green) converges so fast. Comparatively, $\mathcal{L}^{SA}_{com}$ and $\mathcal{L}^{SA}_{spe}$ slowly converges, consistently pushing three FEs for domain disentanglement. This result indicates that integrating the MI estimator with a domain discriminator not only secures robustness but also achieves better performance.

[5]https://github.com/VisionLearningGroup/DAL

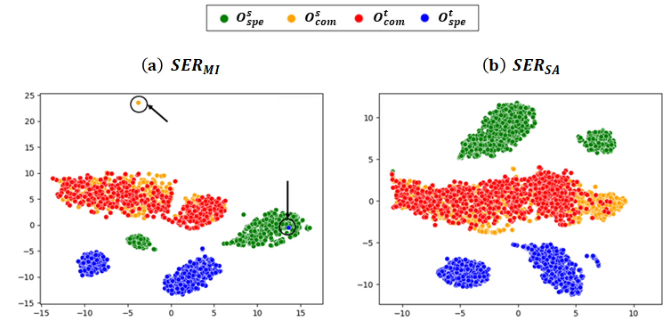

Figure 5: Visualization of disentangled representations for two SER variants; SER_{MI} and SER_{SA}. From the source and target domains, we randomly sample user and item pairs, and plotted their reviews using t-SNE

Visualization of disentangled representations In Figure 5, we visualize the disentangled representations of two *SER* variants. For an experiment, we pre-trained our model using *Baby* and *Patio Lawn and Garden* as source and target domain data. Here, we randomly select one thousand user-item pairs, and extract four types of features $O^s_{spe}, O^s_{com}, O^t_{com}, O^t_{spe}$ from their aggregated reviews. The visualization is conducted using t-distributed Stochastic Neighbor Embedding (t-SNE), which is called a nonlinear dimensionality reduction method for high dimensional representations. In detail, t-SNE has shown to reflect the similarity between data points, where we set the perplexity as 30.

Through Figure 5-a and b, we can see that both methods well capture domain-common knowledge (O^s_{com}, O^t_{com} are overlapped each other). Further, two methods seem to be well disentangle domain-specific features (O^s_{spe}, O^t_{spe}) from domain-common outputs. However, for a disentanglement between domain-specific features ($O^s_{spe} \leftrightarrow O^t_{spe}$), SER_{MI} has shown to implicate some defects. Firstly, in Figure 5-a, O^s_{spe} (cluster of green points) are positioned between O^t_{spe} (cluster of blue points). As mentioned above, t-SNE preserves a point-wise distance after dimensionality reduction, but the clusters derived from SER_{MI} fall apart from each other. Secondly, some points appear out of their original cluster (marked with a circle). We argue that the objective function of SER_{MI} only minimizes MI of each domain independently while disregarding discrimination between domain-specific features. Comparatively, in Figure 5-b, we notice that SER_{SA} well captures domain-specific knowledge, where the outputs are included in a dense region without outliers.

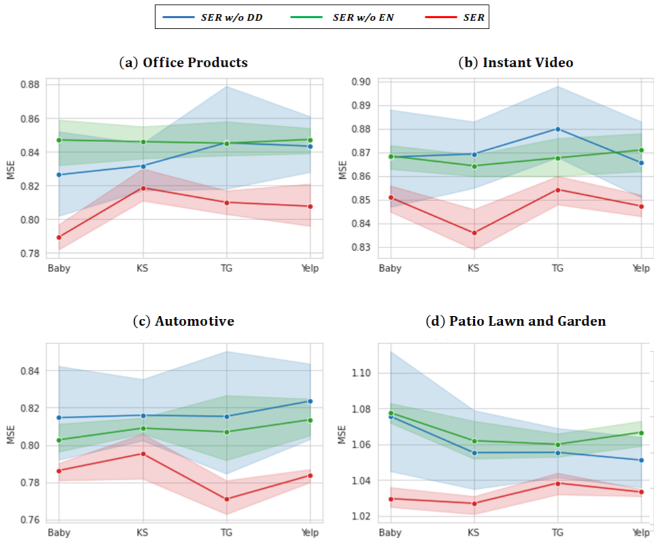

Figure 6: Ablation study (RQ3) of three SER variants. We adopt MSE (↓) score for a comparison

5.4 Ablation Study (RQ3)

To test the significance of domain-aware feature extractor and encoding network, we assume three variants: (1) *SER w/o DD*: excluding domain discriminator, (2) *SER w/o EN*: excluding encoding network, and (3) our original *SER*. Like previous studies, we conduct experiments under four target domain data. In Figure 6, for each domain, we plot the min, median, and max value of MSE scores under 5 iterations. We do not assume a model excluding both components, which is identical to the vanilla text analysis model [66].

Domain-aware feature extractor is fundamental for debilitating noises We discuss *SER w/o DD* (blue lines), removing domain discriminator from SER. In Figure 6, we can see that the performance of *SER w/o DD* shows higher variance compared to *SER w/o EN* (green lines), suggesting the domain discrepancy is critical for the overall performance in a target domain (the column-wise width denotes the variance of the results). For example, in Figure 6-b (*Instant Video*), we notice that the adoption of *Yelp* as a source domain improves the performance, whereas utilizing *Toys and Games* can degrade the recommendation quality.

Encoding network contributes to the recommendation quality significantly Compared to *SER w/o DD*, *SER w/o EN* shows relatively stable results, which can debilitate noises from a source domain. Though the domain-aware feature extractor effectively controls the mismatches between two different domains, we notice that excluding the encoding layer is critical for the overall performance. Specifically, for all datasets, it is notable that *SER w/o DD* slightly outperforms *SER w/o EN*.

5.5 Case Study (RQ4)

In this section, we scrutinize the interpretability of SER through a real-world dataset *Amazon*. In Figure 7, using *Automotive* as a target data, we describe the nDCG@5 of three methods; SER w/o domain-specific features O_{spe}, SER w/o domain-common features O_{com}, and our original *SER*. As can be seen, excluding a single type of information downgrades the performance. Here, we notice that domain-specific information contributes slightly more to the overall quality since *SER w/o* O_{com} outperforms *SER w/o* O_{spe} 1.2% on average. We now scrutinize this phenomenon below.

Figure 7: Case study (RQ4) of three SER variants. We adopt nDCG@5 (↑) score for a comparison

(a) user reviews

Review 1. Comes with Battery Cables, Air Pressure Hose and Connections. ... I like it and its worth the *money*.

Review 2. This is a good *quality* battery that comes with a good protective cover, is leak proof, and when I tested the voltage ...

(b) item (id: B00002243X) reviews

Review 1. I needed a jumper cables for my new car and these had good reviews and were at a good *price*. ... I would recommend ...

Review 2. They are high *quality*. They worked well for me and I had no issues with poor connections. Cables are well made.

Figure 8: Explainability analysis for user and item reviews

In Figure 8, we pre-trained a model using *Baby* for *Automotive*. Then, we apply common and target FEs for each review, retrieving domain-common O_{com}, and domain-specific O_{spe} outputs. We *italicized* the most similar words (*e.g., cosine similarity*) of domain-common feature O_{com}, and colorized most similar words of domain-specific ones O_{spe}. As can be seen, the italicized words further contain semantic meaning (*quality, price, and money*), while the colorized one contains the user's own preference (*cable and battery*). Throughout these results, we can infer that SER well disentangles domain knowledge since two features are complimentary, while none of them are negligible for the improvement of overall quality.

6 CONCLUSION

In this paper, we propose a novel method of utilizing review texts in multiple domains without requiring overlapping users or contexts. Our optimization strategies for domain disentanglement well achieve the knowledge transfer between domains, while also securing robustness. Further, we successfully extend a transfer network from a single to multiple domains, which can infer the latent of an individual review regarding domain-aware knowledge. The extensive experiments and ablation studies demonstrate the superiority and advantages of our method for the CDR scenario.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) (No. 2016R1A5A1012966, No. 2020R1A2C110168713), Institute of Information & communications Technology Planning & Evaluation (IITP) (No. 2021-0-02068 Artificial Intelligence Innovation Hub, No. RS-2022-00156287 Innovative Human Resource Development for Local Intellectualization support program) grant funded by the Korea government (MSIT).

REFERENCES

[1] Shankhadeep Banerjee, Samadrita Bhattacharyya, and Indranil Bose. 2017. Whose online reviews to trust? Understanding reviewer trustworthiness and its impact on business. *Decision Support Systems* 96 (2017), 17–26.

[2] Yang Bao, Hui Fang, and Jie Zhang. 2014. Topicmf: Simultaneously exploiting ratings and reviews for recommendation. In *Twenty-Eighth AAAI conference on artificial intelligence*.

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International Conference on Machine Learning*. PMLR, 531–540.

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79, 1 (2010), 151–175.

[5] Hamed Bonab, Mohammad Aliannejadi, Ali Vardasbi, Evangelos Kanoulas, and James Allan. 2021. Cross-Market Product Recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 110–119.

[6] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. *Advances in neural information processing systems* 29 (2016), 343–351.

[7] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. 2019. Learning disentangled semantic representation for domain adaptation. In *IJCAI: proceedings of the conference*, Vol. 2019. NIH Public Access, 2060.

[8] Rose Catherine and William Cohen. 2017. Transnets: Learning to transform for recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*. 288–296.

[9] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference*. 1583–1592.

[10] Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. 2019. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. (2019).

[11] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. 2019. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning*. PMLR, 1081–1090.

[12] Xu Chen, Yongfeng Zhang, and Zheng Qin. 2019. Dynamic Explainable Recommendation based on Neural Attentive Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 53–60.

[13] Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019. Co-attentive multi-task learning for explainable recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2137–2143.

[14] Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. 2020. Improving disentangled text representation learning with information-theoretic guidance. *arXiv preprint arXiv:2006.00693* (2020).

[15] Xin Dong, Jingchao Ni, Wei Cheng, Zhengzhang Chen, Bo Zong, Dongjin Song, Yanchi Liu, Haifeng Chen, and Gerard De Melo. 2020. Asymmetrical hierarchical networks with attentive interactions for interpretable review-based recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7667–7674.

[16] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*. 278–288.

[17] Bairan Fu, Wenming Zhang, Guangneng Hu, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2021. Dual Side Deep Context-aware Modulation for Social Recommendation. In *Proceedings of the Web Conference 2021*. 2524–2534.

[18] Wenjing Fu, Zhaohui Peng, Senzhang Wang, Yang Xu, and Jin Li. 2019. Deeply fusing reviews and contents for cold start users in cross-domain recommendation systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 94–101.

[19] Marylou Gabrié, Andre Manoel, Clément Luneau, Jean Barbier, Nicolas Macris, Florent Krzakala, and Lenka Zdeborová. 2019. Entropy and mutual information in models of deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment* 2019, 12 (2019), 124014.

[20] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[21] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. 2019. SGD: General analysis and improved rates. In *International Conference on Machine Learning*. PMLR, 5200–5209.

[22] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*. Springer, 63–77.

[23] Lei Guo, Li Tang, Tong Chen, Lei Zhu, Quoc Viet Hung Nguyen, and Hongzhi Yin. 2021. DA-GCN: A Domain-aware Attentive Graph Convolution Network for Shared-account Cross-domain Sequential Recommendation. *arXiv preprint arXiv:2105.03300* (2021).

[24] Adeep Hande, Karthik Puranik, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. Domain identification of scientific articles using transfer learning and ensembles. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 88–97.

[25] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[26] Guangneng Hu, Yu Zhang, and Qiang Yang. 2018. Conet: Collaborative cross networks for cross-domain recommendation. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 667–676.

[27] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4893–4902.

[28] SeongKu Kang, Junyoung Hwang, Dongha Lee, and Hwanjo Yu. 2019. Semi-supervised learning for cross-domain recommendation to cold-start users. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1563–1572.

[29] Adit Krishnan, Mahashweta Das, Mangesh Bendre, Hao Yang, and Hari Sundaram. 2020. Transfer Learning via Contextual Invariants for One-to-Many Cross-Domain Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1081–1090.

[30] Bo Li, Yezhen Wang, Shanghang Zhang, Dongsheng Li, Kurt Keutzer, Trevor Darrell, and Han Zhao. 2021. Learning invariant representations and risks for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1104–1113.

[31] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. 2020. Maximum density divergence for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence* (2020).

[32] Pan Li and Alexander Tuzhilin. 2020. Ddtcdr: Deep dual transfer cross domain recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 331–339.

[33] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 345–354.

[34] Zejian Li, Yongchuan Tang, Wei Li, and Yongxing He. 2019. Learning disentangled representation with pairwise independence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4245–4252.

[35] Huiting Liu, Lingling Guo, Peipei Li, Peng Zhao, and Xindong Wu. 2021. Collaborative filtering with a deep adversarial and attention network for cross-domain recommendation. *Information Sciences* 565 (2021), 370–389.

[36] Tong Man, Huawei Shen, Xiaolong Jin, and Xueqi Cheng. 2017. Cross-Domain Recommendation: An Embedding and Mapping Approach.. In *IJCAI*, Vol. 17. 2464–2470.

[37] David McAllester and Karl Stratos. 2020. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 875–884.

[38] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. 165–172.

[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[40] Andriy Mnih and Russ R Salakhutdinov. 2008. Probabilistic matrix factorization. In *Advances in neural information processing systems*. 1257–1264.

[41] Preksha Nema, Alexandros Karatzoglou, and Filip Radlinski. 2021. Disentangling Preference Representations for Recommendation Critiquing with ß-VAE. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1356–1365.

[42] Stefano Panzeri, Riccardo Senatore, Marcelo A Montemurro, and Rasmus S Petersen. 2007. Correcting for the sampling bias problem in spike train information measures. *Journal of neurophysiology* 98, 3 (2007), 1064–1072.

[43] Changhwa Park, Jonghyun Lee, Jaeyoon Yoo, Minhoe Hur, and Sungroh Yoon. 2020. Joint contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2006.10297* (2020).

[44] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. 2019. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning*. PMLR, 5102–5112.

[45] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[46] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. *arXiv preprint arXiv:1810.03649* (2018).

[47] Noveen Sachdeva and Julian McAuley. 2020. How Useful are Reviews for Recommendation? A Critical Review and Potential Improvements. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in*

Information Retrieval. 1845–1848.

[48] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. 2019. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8050–8058.

[49] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the eleventh ACM conference on recommender systems*. 297–305.

[50] Shaoyun Shi, Weizhi Ma, Zhen Wang, Min Zhang, Kun Fang, Jingfang Xu, Yiqun Liu, and Shaoping Ma. 2021. WG4Rec: Modeling Textual Content with Word Graph for News Recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1651–1660.

[51] Jie Shuai, Kun Zhang, Le Wu, Peijie Sun, Richang Hong, Meng Wang, and Yong Li. 2022. A Review-aware Graph Contrastive Learning Framework for Recommendation. *arXiv preprint arXiv:2204.12063* (2022).

[52] Yunzhi Tan, Min Zhang, Yiqun Liu, and Shaoping Ma. 2016. Rating-boosted latent topics: Understanding users and items with ratings and reviews.. In *IJCAI*, Vol. 16. 2640–2646.

[53] Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Multi-pointer co-attention networks for recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2309–2318.

[54] Xi Wang, Iadh Ounis, and Craig Macdonald. 2021. Leveraging Review Properties for Effective Recommendation. In *Proceedings of the Web Conference 2021*. 2209–2219.

[55] Xinghua Wang, Zhaohui Peng, Senzhang Wang, S Yu Philip, Wenjing Fu, and Xiaoguang Hong. 2018. Cross-domain recommendation for cold-start users via neighborhood based feature mapping. In *International conference on database systems for advanced applications*. Springer, 158–165.

[56] Le Wu, Yonghui Yang, Lei Chen, Defu Lian, Richang Hong, and Meng Wang. 2020. Learning to transfer graph embeddings for inductive graph based recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1211–1220.

[57] Kun Xiong, Wenwen Ye, Xu Chen, Yongfeng Zhang, Wayne Xin Zhao, Binbin Hu, Zhiqiang Zhang, and Jun Zhou. 2021. Counterfactual Review-based Recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2231–2240.

[58] Xiangli Yang, Qing Liu, Rong Su, Ruiming Tang, Zhirong Liu, and Xiuqiang He. 2021. AutoFT: Automatic Fine-Tune for Parameters Transfer Learning in Click-Through Rate Prediction. *arXiv preprint arXiv:2106.04873* (2021).

[59] Wenhui Yu, Xiao Lin, Junfeng Ge, Wenwu Ou, and Zheng Qin. 2020. Semi-supervised collaborative filtering by text-enhanced domain adaptation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2136–2144.

[60] Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. Aesthetic-based clothing recommendation. In *Proceedings of the 2018 world wide web conference*. 649–658.

[61] Fajie Yuan, Xiangnan He, Alexandros Karatzoglou, and Liguang Zhang. 2020. Parameter-efficient transfer from sequential behaviors for user modeling and recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1469–1478.

[62] Feng Yuan, Lina Yao, and Boualem Benatallah. 2019. DARec: deep domain adaptation for cross-domain recommendation via transferring rating patterns. *arXiv preprint arXiv:1905.10760* (2019).

[63] Hansi Zeng, Zhichao Xu, and Qingyao Ai. 2021. A Zero Attentive Relevance Matching Networkfor Review Modeling in Recommendation System. *arXiv preprint arXiv:2101.06387* (2021).

[64] Cheng Zhao, Chenliang Li, Rong Xiao, Hongbo Deng, and Aixin Sun. 2020. CATN: Cross-Domain Recommendation for Cold-Start Users via Aspect Transfer Network. *arXiv preprint arXiv:2005.10549* (2020).

[65] Haiteng Zhao, Chang Ma, Qinyu Chen, and Zhihong Deng. 2021. Domain Adaptation via Maximizing Surrogate Mutual Information. *arXiv preprint arXiv:2110.12184* (2021).

[66] Lei Zheng, Vahid Noroozi, and Philip S Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 425–434.

[67] Feng Zhu, Yan Wang, Chaochao Chen, Jun Zhou, Longfei Li, and Guanfeng Liu. 2021. Cross-domain recommendation: challenges, progress, and prospects. *arXiv preprint arXiv:2103.01696* (2021).