Domain-oriented Language Modeling with Adaptive Hybrid Masking and Optimal Transport Alignment

Denghui Zhang[1], Zixuan Yuan[1], Yanchi Liu[2*], Hao liu[4],
Fuzhen Zhuang[3], Hui Xiong[1*], Haifeng Chen[2]
[1]Rutgers, The State University of New Jersey, [2]NEC Labs America, USA
[3]Institute of Artificial Intelligence, School of Computer Science, Beihang University, China
[4]The Hong Kong University of Science and Technology, China

ABSTRACT

Motivated by the success of pre-trained language models such as BERT in a broad range of natural language processing (NLP) tasks, recent research efforts have been made for adapting these models for different application domains. Along this line, existing domain-oriented models have primarily followed the vanilla BERT architecture and have a straightforward use of the domain corpus. However, domain-oriented tasks usually require accurate understanding of domain phrases, and such *fine-grained phrase-level knowledge* is hard to be captured by existing pre-training scheme. Also, the word co-occurrences guided semantic learning of pre-training models can be largely augmented by *entity-level association knowledge*. But meanwhile, by doing so there is a risk of introducing noise due to the lack of groundtruth word-level alignment. To address the above issues, we provide a generalized domain-oriented approach, which leverages auxiliary domain knowledge to improve the existing pre-training framework from two aspects. First, to preserve phrase knowledge effectively, we build a domain phrase pool as auxiliary training tool, meanwhile we introduce Adaptive Hybrid Masked Model to incorporate such knowledge. It integrates two learning modes, word learning and phrase learning, and allows them to switch between each other. Second, we introduce Cross Entity Alignment to leverage entity association as weak supervision to augment the semantic learning of pre-trained models. To alleviate the potential noise in this process, we introduce an interpretable *Optimal Transport based approach* to guide alignment learning. Experiments on four domain-oriented tasks demonstrate the superiority of our framework.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

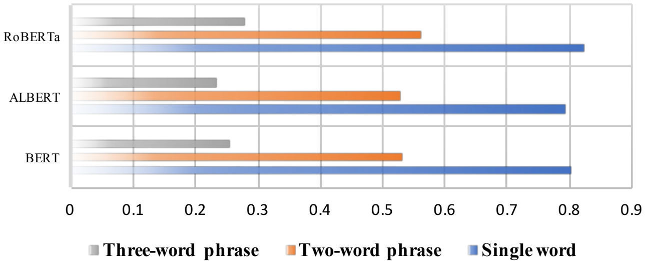


Figure 1: The single-word and phrase reconstruction accuracy of several existing language pre-training models.

KEYWORDS

Domain language modeling, pre-training, masked language model, optimal transport.

ACM Reference Format:
Denghui Zhang[1], Zixuan Yuan[1], Yanchi Liu[2*], Hao liu[4],, Fuzhen Zhuang[3], Hui Xiong[1*], Haifeng Chen[2] . 2021. Domain-oriented Language Modeling with Adaptive Hybrid Masking and Optimal Transport Alignment. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3447548.3467215

1 INTRODUCTION

Recent years have witnessed the great success of pre-trained language models (PLMs), such as BERT [7], in a broad range of natural language processing (NLP) tasks. Moreover, several domain-oriented PLMs have been proposed to adapt to specific domains [4, 8, 10]. For instance, BioBERT [13] and SciBERT [2] are pre-trained leveraging large-scale domain-specific corpora for biomedical and scientific domain tasks respectively. However, in the above models, the same pre-training scheme as BERT is reused straightforwardly, while insightful domain characteristics are largely overlooked. To this end, we raise a natural question: *for domain language pre-training, can we go further beyond the strategy of vanilla BERT + domain corpus by leveraging domain characteristics?* In this paper, we explore this question under e-commerce domain and present promising approaches that can also be generalized to other domains when auxiliary knowledge is available.

We first discuss the characteristics of domain-oriented tasks, and the limitations of current pre-training approaches, then present two major improving strategies, corresponding to leveraging two types of auxiliary domain knowledge smartly. On the one hand, understanding a great variety of *domain phrases* is critical to domain-oriented tasks. As shown in Table 1, the review aspect extraction

* Corresponding authors. This work was partially supported by the National Science Foundation through awards III-2006387, IIS 1814510, and IIS-2040799.

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467215

Table 1: An example of review aspect extraction, where correct answers (marked in color) are usually phrases.

Review: That included the extra Sony Sonic Stage software, the speakers and the subwoofer I got (that WAS worth the money), the bluetooth mouse for my supposedly bluetooth enabled computer, the extended life battery and the docking port. […]

task, widely used in the e-commerce domain, requires language models to understand domain phrases to extract the correct answers. However, such *phrase-level* domain knowledge is hard to be captured by Masked Language Model (MLM) [7] (i.e., the self-supervised task employed in most language pre-training models). Figure 1 depicts the language reconstruction performance of three existing language pre-training models on a public e-commerce corpus. As can be seen, the reconstruction accuracy drops drastically when the prediction length is increased from single word to multi-word phrase. We attribute this to the fact that MLM is a *word-oriented* task, i.e., it only reconstructs randomly masked words from the incomplete input however does not explicitly encourage any perception ability for domain phrases. Although later works [11, 32] propose to mask phrases instead of words in MLM to enable BERT for phrase perception, they have two major drawbacks: (i) *Overgeneralized phrase selection*, they use chunking [31] to randomly select phrases to mask, without considering the quality of phrases and the relatedness to specific domains. (ii) *Discard of word masking*, word masking helps to acquire word-level semantics essential for phrase learning, hence should be preserved in pre-training.

On the other hand, pre-trained language models are limited by corpus-level statistics such as co-occurrence, which can be mitigated by auxiliary domain knowledge. For instance, to learn that Android and iOS are semantically related, a large number of co-occurrences in similar contexts are required in the pre-training data. For domain-oriented learning, this can be mitigated by auxiliary knowledge, i.e., *entity association.* As shown in Table 2, when leveraging the "substitutable" association to pair the description texts of two product entities, Samsung galaxy and iPhone, we can augment the co-occurrence of some words/phrases (e.g, *5G network* vs *4G signal*; *Android* vs *iOS*) by learning the alignments of similar words across entities. However, the above intuition is challenging to fulfill in practice as it constitutes a *weakly supervised learning task.* In other words, only weak-supervision signals (i.e., entity-level alignments) are available, while the word-level groundtruth alignments across entities are hard to obtain. Hence, the aligning problem needs a *robust* learning algorithm to overcome the potential noises under the weak supervision. Moreover, the algorithm should also offer decent *interpretability* over the alignment for the ease of understanding and validation.

Based on the above insights, we propose an enhanced domain-oriented framework for language pre-training. Our framework takes the mentioned domain characteristics into consideration, and introduces two approaches to tackle the challenges. First, to enable language pre-training with the perception ability for domain phrases, we propose an advanced alternative for Masked Language Model, namely, Adaptive Hybrid Masked Model (AHM). In contrast to MLM only masking and reconstructing single words, AHM introduces a new sampling scheme for masking quality phrases with the guidance of an external domain phrase pool, and meanwhile, a novel *phrase completeness regularization term* is proposed for sophisticated phrase reconstruction. Furthermore, since both word-level and phrase-level semantics are critical to language modeling, we unify the word and phrase learning modes via a loss-based parameter. It allows the adaptive switching between each other, ensuring a *smooth and progressive* learning process resembling the human cognition of language. Second, to exploit the rich co-occurrence signals hidden in entity associations, we formulate a new pre-training task, namely, Cross Entity Alignment (CEA). Specifically, CEA aims to learn the word-level alignment matrix of entity association based text pair (e.g., description pair) with only weak supervision, i.e., only knowing two entities are related but no word-level groundtruth alignments available. Moreover, we propose an alignment learning scheme leveraging Optimal Transport (OT) to train this task in a weakly-supervised fashion. At each round, the OT objective helps to find the pseudo optimal matching of similar words (or phrases) and returns a *sparse transport plan*, which reveals robust and interpretable alignments. The language model is further optimized with the guidance of the transport plan to minimize the Wasserstein Distance of the aligned entity contents, enabling the model to learn fine-grained semantic correlations.

Table 2: An example of relational text in the e-commerce domain, where product descriptions are connected by the "substitutable" product association.

Product	Description
Samsung Galaxy S10	OS: Andriod; 5G network; Dynamic AMOLED; …
iPhone XS	iOS; 4G signal; T-Mobile service; OLED screen; …

To validate the effectiveness of the proposed approach, we conduct extensive experiments in the e-commerce domain to compare our pre-training framework with state-of-the-art baselines. Specifically, we employ the pre-training corpus created from publicly available resources and fine-tune on four downstream tasks, i.e., Review-based Question Answering (RQA), Aspect Extraction (AE), Aspect Sentiment Classification (ASC), and Product Title Categorization (PTC). Quantitative results show that our method significantly outperforms BERT and other variants on all the tasks. Additionally, the visualization of OT-based approach reveals feasible alignment results despite the weak supervision, meanwhile, presenting convincing interpretability as the alignment vector is enforced to be sparse. Lastly, while we demonstrate the effectiveness of our approach in the e-commerce domain, the ideas of the framework can be generalized to broader domains since the aforementioned auxiliary knowledge is free of annotation cost. The domain phrase pool can be constructed from domain corpus. Entity association is broad and general, which is easy to obtain in main domains.

2 RELATED WORK

Pre-trained Language Models. Recently, the emergence of pre-trained language models (PLMs) [7, 23, 26] has brought natural language processing to a new era. Compared with traditional word embedding models [20], PLMs learn to represent words based on the entire input context to tackle polysemy, hence captures semantics more accurately. Following PLMs, many endeavors have been made for further optimization in terms of both architecture and training scheme [3, 15, 16, 32]. Along this line, SpanBERT [11] proposes to reconstruct randomly masked spans instead of single words.

However, the span consists of random continuous words and may not form phrases, thus fails to capture phrase-level knowledge effectively. ERNIE [31] integrates phrase-level masking and entity-level masking into BERT, which is closely related to our masking scheme. Differing from their work simply using chunking to get general phrases, we build high-quality domain phrase pool to assist learning domain-oriented phrase knowledge. Also, we propose a novel phrase regularization term over the reconstruction loss to encourage complete phrase learning. Moreover, we combine word and phrase learning cohesively according to their optimizing progress, achieving better performance than each single mode.

Domain-oriented PLMs. To adapt PLMs to specific domains, several domain-oriented BERTs such as BioBERT [13], SciBERT [2], and TweetBERT [25], have been proposed recently. BERT-PT [36] proposes to post-train BERT on a review corpus and obtains better performance on the task of review reading comprehension. Gururangan et al. [9] proposes an approach for post-training BERT on domain corpus as well as task corpus to obtain more performance gains on domain-specific tasks. DomBERT [37] proposes to select data from a mixed multi-domain corpus for the target domain, improving the diversity of domain language learning. More work along this line can be referred to [18, 28]. Similarly, incorporating domain knowledge has shown effectiveness in broader areas [14, 33, 38–41] such as representation learning. The above solutions have primarily leveraged domain corpus for pre-training in a straightforward way, without considering insightful domain characteristics and domain knowledge such as domain phrase and entity association. Our work is the first leveraging auxiliary domain knowledge to enhance domain-oriented pre-training.

3 PRELIMINARIES

In this section, we give a brief introduction to two essential concepts that are related to our work, namely, Masked Language Model and Optimal Transport.

Masked Language Model. Masked Language Model (MLM) [7] refers to the self-supervised pre-training task that have been applied in pre-trained language models (e.g., BERT, RoBERTa, etc.). It is considered as a fill-in-the-blank task, i.e., given an input sequence partially masked (15% tokens), it aims to predict those masked words using the embeddings generated by the language model:

$$p\left(X_m \middle| X_{\setminus M}\right) = \frac{\exp\left(\boldsymbol{W}_{\boldsymbol{m}}^{\top}\left[\mathcal{F}(X_{\setminus M};\theta)\right]_m\right)}{\sum\limits_{k\in\mathcal{V}}\exp\left(\boldsymbol{W}_{\boldsymbol{k}}^{\top}\left[\mathcal{F}(X_{\setminus M};\theta)\right]_m\right)}, \tag{1}$$

where $\mathcal{F}(;\theta)$ denotes the Transformer based language model. X is the full input sequence, M denotes the indices of all masked tokens in X, X_m indicates one of the tokens in M, $\setminus$ is set minus. $\left[\mathcal{F}\left(X_{\setminus M};\theta\right)\right]_m$ denotes the output vector corresponding to the masked token X_m and $W^\top$ denotes the softmax matrix with the same number of entries as the vocabulary $\mathcal{V}$.

Maximizing $p\left(X_m|X_{\setminus M}\right)$ enforces $\mathcal{F}(;\theta)$ to infer the meaning of masked words from their surroundings, in other words, preserving contextual semantics.

Optimal Transport and Wasserstein Distance. Optimal Transport (OT) studies the problem of transforming one probability distribution into another one (e.g., one group of embeddings to another) with the lowest cost. When considering the "cost" as distance, a commonly used distance metric for OT is Wasserstein Distance (WD) [34]. Formal definition is as follows[5]:

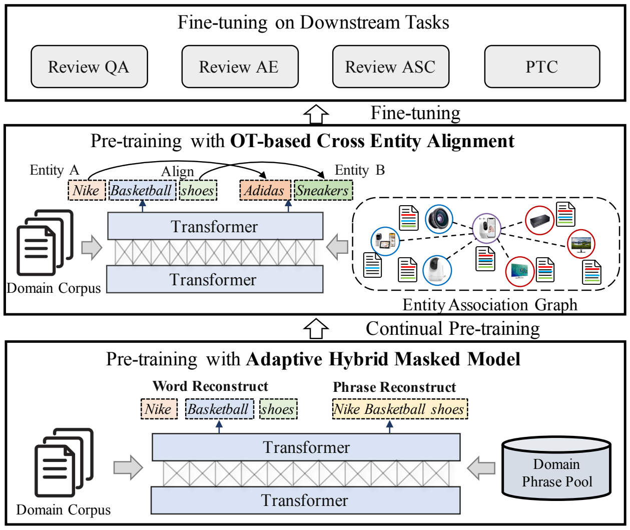


Figure 2: Framework overview.

Definition 3.1. Let $\boldsymbol{\mu} \in \mathbf{P}(\mathbb{X}), \boldsymbol{\nu} \in \mathbf{P}(\mathbb{Y})$ denote two probability distributions, formulated as $\boldsymbol{\mu} = \sum_{i=1}^{m} \mathrm{u}_i \delta_{\boldsymbol{x}_i}$ and $\boldsymbol{\nu} = \sum_{j=1}^{n} \mathrm{v}_j \delta_{\boldsymbol{y}_j}$, with $\delta_{\boldsymbol{x}}$ as the Dirac function centered on $\boldsymbol{x}$. $\Gamma(\boldsymbol{\mu},\boldsymbol{\nu})$ denotes all the couplings (joint distributions) of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, with marginals $\boldsymbol{\mu}(\boldsymbol{x})$ and $\boldsymbol{\nu}(\boldsymbol{y})$. The optimal Wasserstein Distance between the two distributions $\boldsymbol{\mu},\boldsymbol{\nu}$ is defined as:

$$\begin{aligned}\mathcal{D}_w(\boldsymbol{\mu},\boldsymbol{\nu}) &= \inf_{\boldsymbol{\gamma}\in\Gamma(\boldsymbol{\mu},\boldsymbol{\nu})} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\boldsymbol{\gamma}}\ [c(\boldsymbol{x},\boldsymbol{y})] \\ &= \min_{\mathbf{T}\in\Gamma(\mathbf{u},\mathbf{v})} \langle \mathbf{T},\mathbf{C}\rangle = \min_{\mathbf{T}\in\Gamma(\mathbf{u},\mathbf{v})} \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{T}_{ij}\cdot c(\boldsymbol{x}_i,\boldsymbol{y}_j)\,,\end{aligned} \tag{2}$$

where $\Gamma(\mathbf{u},\mathbf{v}) = \{\mathbf{T}\in\mathbb{R}_{+}^{m\times n}|\mathbf{T}\mathbf{1}_n = \mathbf{u}, \mathbf{T}^\top\mathbf{1}_m = \mathbf{v}\}$, $\mathbf{1}_m$ denotes an m-dimensional all-one vector, the weight vectors $\mathbf{u} = \{\mathrm{u}_i\}_{i=1}^{m} \in \Delta_m$ and $\mathbf{v} = \{\mathrm{v}_i\}_{i=1}^{n} \in \Delta_n$ belong to the m- and n-dimensional simplex, respectively (*i.e.*, $\sum_{i=1}^{m}\mathrm{u}_i = \sum_{j=1}^{n}\mathrm{v}_j = 1$). And $c(\boldsymbol{x}_i,\boldsymbol{y}_j)$ is the cost function evaluating the distance between $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ (samples of the two distributions). Computing the optimal distance (1st line) is equivalent to solving the network-flow problem (2nd line) [17]. The calculated matrix $\mathbf{T}$ denotes the "transport plan", where each element $\mathbf{T}_{ij}$ represents the amount of mass shifted from u_i to v_j. We propose an Optimal Transport based approach for the cross entity alignment problem in Section 4.2.

4 METHODOLOGY

In this section, we provide an in-depth introduction to our enhanced framework for domain-oriented language pre-training. Figure 2 presents an overview of the framework, consisting of two major improvements, i.e., Adaptive Hybrid Masked Model (AHM) to replace MLM and a new weakly-supervised pre-training task, OT-based Cross Entity Alignment (CEA). The former leverages a domain corpus and a domain phrase pool to learn both word-level and phrase-level semantics, the latter utilizes the same corpus and an entity association graph to obtain text pairs for augmenting domain semantic learning. Moreover, we employ continual multi-task

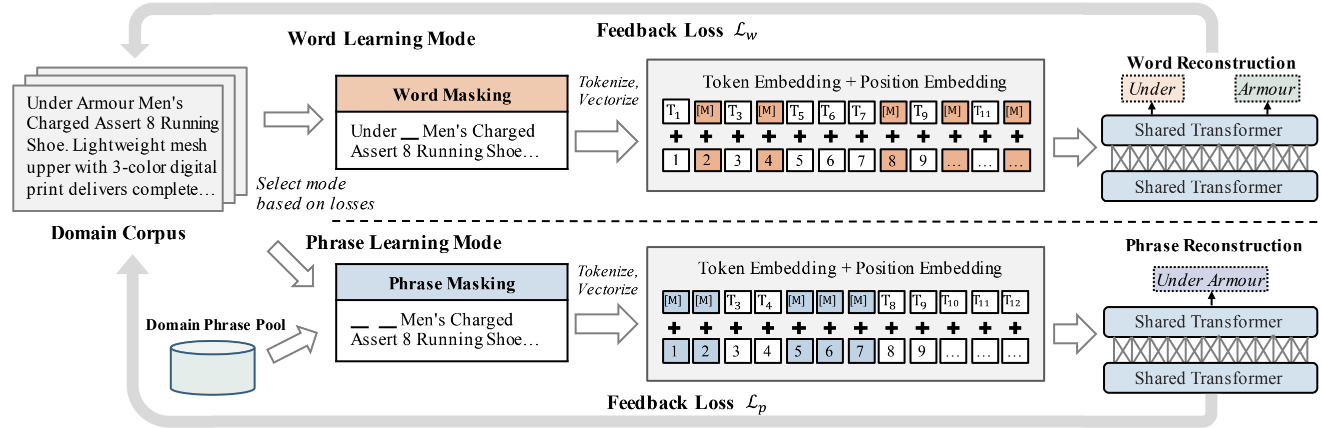

Word Learning Mode
Feedback Loss $\mathcal{L}_w$
Under Armour Men's Charged Assert 8 Running Shoe. Lightweight mesh upper with 3-color digital print delivers complete…
Domain Corpus
Word Masking
Under __ Men's Charged Assert 8 Running Shoe…
Tokenize, Vectorize
Token Embedding + Position Embedding
T1 [M] T3 [M] T5 T6 T7 [M] T9 [M] T11 [M]
1 2 3 4 5 6 7 8 9 … … …
Word Reconstruction
Under
Armour
Shared Transformer
Shared Transformer
Select mode based on losses
Phrase Learning Mode
Domain Phrase Pool
Phrase Masking
__ __ Men's Charged Assert 8 Running Shoe…
Tokenize, Vectorize
Token Embedding + Position Embedding
[M] [M] T3 T4 [M] [M] [M] T8 T9 T10 T11 T12
1 2 3 4 5 6 7 8 9 … … …
Phrase Reconstruction
Under Armour
Shared Transformer
Shared Transformer
Feedback Loss $\mathcal{L}_p$

Figure 3: Illustration of Adaptive Hybrid Masked Model. Based on the feedback losses, it adaptively switches between two learning modes, enabling the language model to learn word-level and phrase-level knowledge simultaneously.

pre-training [32] to jointly train AHM and CEA. Lastly, the model is fine-tuned to be deployed in domain-oriented applications.

4.1 Adaptive Hybrid Masked Model

In order to enhance the phrase perception ability of language model while meantime preserving its original word perception ability, we introduce a new masked language model, namely, Adaptive Hybrid Masked Model (AHM). Specifically, we set two learning modes in AHM, i.e., *word learning* and *phrase learning*, which in a nutshell, *masks then reconstructs* word units and phrase units, respectively. Moreover, we combine the two learning modes by adaptively switching between them, enabling the model to capture the word-level and phrase-level semantics *simultaneously* and *progressively*. Figure 3 provides an illustration of the model.

4.1.1 *Word Learning Mode*

In this mode, given an input sequence X^t (t denotes the t^{th} iteration), we first randomly sample words from X^t iteratively until the selected words constitute 15% of all tokens. Then we replace them with: (1) the [MASK] token 80% of the time, (2) a random token 10% of the time, (3) the original token 10% of the time. Next, we predict all the masked/perturbed tokens by feeding their embeddings of the language model to a shared softmax layer. Equivalently, we optimize the log-likelihood function below:

$$\mathcal{L}_w = -\log \prod_{m\in\mathcal{W}^t} p(X_m^t|X_{\setminus\mathcal{W}^t}^t), \tag{3}$$

where $\mathcal{W}^t$ denotes the indices of all the masked/perturbed tokens in X^t. X_m^t and $X_{\setminus\mathcal{W}^t}^t$ denotes the m^{th} masked token and perturbed input, respectively. $p(X_m^t|X_{\setminus\mathcal{W}^t}^t)$ follows the definition in Eq.(1). This mode resembles the original masking scheme in MLM except that we only mask whole words. It helps to learn preliminary word-level semantics, which is not only the basis of language understanding but also essential for phrase learning.

4.1.2 *Phrase Learning Mode*

In the phrase learning mode, we randomly mask consecutive tokens that constitutes *quality domain phrases* and train the language model to reconstruct them. First, given an input sequence X^t and a domain phrase pool $\mathcal{P}_D$ (comprising high-quality phrases and their quality scores)[1], following Algorithm 1, we detect domain phrases and sample to obtain 15% tokens. Then similar to the word mode, we replace the selected tokens with [MASK] token 80% of the time, a random token and the original token 10% of the time respectively. Next, we optimize the following loss function to reconstruct the masked phrases:

$$\mathcal{L}_P = -\Big(\log \prod_{m\in\mathcal{P}^t} p(X_m^t|X_{\setminus\mathcal{P}^t}^t) + \underbrace{\log \prod_{P\in\hat{\mathcal{P}}^t} r(X_P^t|X_{\setminus\hat{\mathcal{P}}^t}^t)}_{\textit{completeness regularization}}\Big), \tag{4}$$

$$r(X_P^t|X_{\setminus\hat{\mathcal{P}}^t}^t) = \frac{\exp\Big(\boldsymbol{C}_P^\top \mathbf{Avg}\big(\big[\mathcal{F}(X_{\setminus\hat{\mathcal{P}}^t}^t;\theta)\big]_P\big)\Big)}{\sum_{k\in\mathcal{V}_P}\exp\Big(\boldsymbol{C}_k^\top \mathbf{Avg}\big(\big[\mathcal{F}(X_{\setminus\hat{\mathcal{P}}^t}^t;\theta)\big]_P\big)\Big)}, \tag{5}$$

where the first term is defined the same way as Eq.(1) and (3) except that $\mathcal{P}^t$ denotes indices of all the masked tokens obtained via Algorithm 1. With the first term, we *reconstruct masked phrases by predicting their tokens*. Additionally, we propose an *completeness regularization term* (the second term) over the masked phrases to encourage complete phrase reconstruction, i.e., the model will get more rewards when an *entire phrase* is correctly predicted. As defined in Eq.(5), where $\hat{\mathcal{P}}^t$ also denotes the indices of masked tokens but grouped by phrases, P denotes one of the group in $\hat{\mathcal{P}}^t$, we first average all the token embeddings of a phrase to obtain the merged phrase feature (i.e., $\mathbf{Avg}\big(\big[\mathcal{F}(X_{\setminus M};\theta)\big]_P\big)$). Then we predict each complete phrase instead of the tokens in it using its merged feature along with a new phrase softmax matrix (i.e., $\boldsymbol{C}^\top$). $\mathcal{V}_P$ represents the set of all phrases in corpus.

4.1.3 *Adaptive Hybrid Learning*

As both word-level and phrase-level semantics are critical to language modeling, we combine the two learning modes via a dynamic parameter α based on the feedback losses of them. At each iteration, as shown in Figure 3, the model automatically selects the weaker mode according to the value of α.

[1] In this paper, we leverage AutoPhrase [30] to obtain domain phrase pool.

[2] Fulfill via a rule-based phrase matcher, https://spacy.io/usage/rule-based-matching

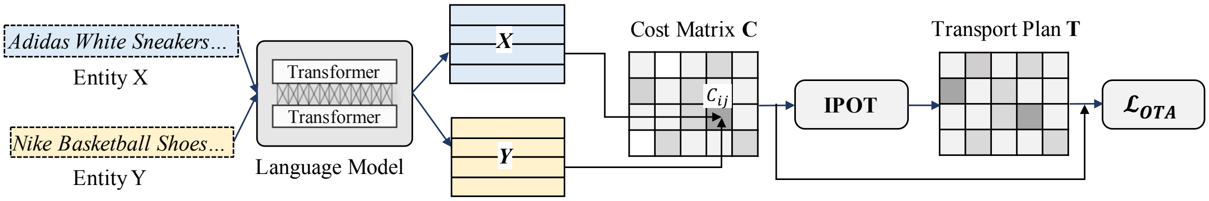


Figure 4: Illustration of the OT based approach for learning the word-level alignments for entity association based text pair.

Algorithm 1 Token sampling algorithm for the phrase mode.

```
Input: An sequence $X^t$; The domain phrase pool $\mathcal{P}_D$.
Output: Token indices of domain phrases, denoted by $\mathcal{P}^t$; Token
    indices grouped by phrases, denoted by $\hat{\mathcal{P}}^t$.
Detect phrases[2] in $X_t$ that intersect with $\mathcal{P}_D$, denoted by $\mathcal{P}_T$;
Retrieve their quality scores $\{s_i\}$ from $\mathcal{P}_D$;
Normalize all the scores by softmax, i.e.,
    $s_{n,i} = \exp(s_i)/\exp(\sum_j \exp(s_j))$;
Let count = 0, $\mathcal{P}^t = \emptyset$, $\hat{\mathcal{P}}^t = \emptyset$;
while count/num_token($X_t$)<15% do
    Sample a phrase $p$ from $\mathcal{P}_T$ based on the normalized
        scores, i.e., $\{s_{n,i}\}$;
    Add the indices of all tokens in $p$ into $\mathcal{P}^t$;
    Add the indices in $p$ as a list into $\hat{\mathcal{P}}^t$;
    count += 1;
end while
Return $\mathcal{P}^t$, $\hat{\mathcal{P}}^t$.
```

Calculating $\boldsymbol{\alpha}$. We calculate α based on the relative loss reduction speed of the two modes. Specifically, at each iteration (assuming t^{th}), we first calculate a special variable for both modes to track their fitting progress, i.e., η_w^t and η_p^t. The larger η_w^t (η_p^t) is, the less sufficient the model is trained on the word (phrase) mode. Then α^{t+1} for next iteration is calculated as the rescaled ratio of η_w^t and η_p^t, i.e.,

$$\eta_w^t = \frac{\Delta_w^{t,t-1}}{\Delta_w^{t,1}} = \frac{\left[\mathcal{L}_w^{t-1} - \mathcal{L}_w^t\right]_+}{\mathcal{L}_w^1 - \mathcal{L}_w^t}, \quad \eta_p^t = \frac{\Delta_p^{t,t-1}}{\Delta_p^{t,1}} = \frac{\left[\mathcal{L}_p^{t-1} - \mathcal{L}_p^t\right]_+}{\mathcal{L}_p^1 - \mathcal{L}_p^t}, \quad (6)$$

$$\alpha^{t+1} = \tanh(\eta_w^{t+1}/\eta_p^{t+1}). \quad (7)$$

where $\mathcal{L}_w^t$ denotes the loss of the word learning mode and will only be updated if word mode is selected at the t-th iteration. Function $[x]_+$ is equivalent to $max(x, 0)$. $\Delta_w^{t,t-1}$ denotes the loss reduction of word mode between the current and last iteration. $\Delta_w^{t,1}$ denotes the total loss reduction. $\Delta_p^{t,t-1}$, $\Delta_p^{t,1}$, $\mathcal{L}_p^t$ represents the same variables in the phrase mode. Thus, η_w^{t+1} and η_p^{t+1} indicates the relative loss reduction speed of the two modes respectively, and the ratio them ($\eta_w^{t+1}/\eta_p^{t+1}$) reflects the relative importance of the word mode. The non-linear function tanh is used to rescale the ratio to [0,1].

Loss Function of AHM. The overall loss function of AHM is the combined losses of the two learning modes, with weights dynamically adjusted by α^t, i.e.,

$$\mathcal{L}_{\text{AHM}} = \frac{1}{|\mathcal{D}|} \sum_{X^t \in \mathcal{D}} \mathbb{I}(\alpha^t) \cdot \mathcal{L}_w + \mathbb{I}(1 - \alpha^t) \cdot \mathcal{L}_p, \quad (8)$$

$$\mathbb{I}(x) = \begin{cases} 1 & \text{if } x > 0.5, \\ 0 & \text{if } x \leq 0.5. \end{cases} \quad (9)$$

where $\mathcal{D}$ represents the training corpus. $\mathbb{I}$ denotes the indicator function defined in Eq.(9). As can be seen, when $\eta_w^{t+1} \gg \eta_p^{t+1}$, $\alpha^{t+1} \approx 1$, the word mode becomes dominating, and vice versa. In other words, α is able to control the model to switch to the weaker learning mode adaptively.

4.2 OT-based Cross Entity Alignment

To exploit the co-occurrence signals hidden in entity associations, we formulate a new pre-training task, i.e., Cross Entity Alignment (CEA), as defined below. We first exploit the entity association graph to extract a collection of *associated text pairs* from the domain corpus as training data. Next, an Optimal Transport (OT) based approach is introduced to train CEA effectively.

Definition 4.1. Given two paired entity contents denoted by word sequences $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$, Cross Entity Alignment aims to learn an word-level alignment matrix $\mathbf{A}$, where $\mathbf{A}_{i,j} \in [0, 1]$ indicates the correlation of x_i and y_j (s.t. $\sum_j \mathbf{A}_{i,j} = 1, \sum_i \mathbf{A}_{i,j} = 1$).

The task is challenging due to the lack of groundtruth alignment matrix $\mathbf{A}'$. A common solution to this problem involves designing advanced attention mechanisms to simulate soft alignment. However, the learned attention matrices are often too dense and lack interpretability, inducing less effective alignment learning. On the other hand, OT possesses *ideal sparsity* that makes it a good choice for cross-domain alignment problems [5]. Specifically, when solved exactly, OT yields a sparse solution $\mathbf{T}^* \in \mathbb{R}^{m \times n}$ containing $(2r - 1)$ non-zero elements at most, where $r = max(m, n)$, leading to a more interpretable and robust alignment. Hence, we propose an OT-based approach to the address CEA. Figure 4 presents an overview illustration of our Optimal Transport based approach for CEA. Concretely, we follow the below procedures to fulfill it.

Content Embeddings and Cost Matrix. Given the entity pair (X, Y), we first feed their content texts into the language model (Transformer) respectively to get the contextual embeddings, denoted by $\boldsymbol{X} = \{\boldsymbol{x_i}\}_{i=1}^m$ and $Y = \{\boldsymbol{y_j}\}_{j=1}^n$. Then we calculate a cost matrix $\mathbf{C} \in \mathbb{R}^{m \times n}$, where $\mathbf{C}_{ij}$ defines the cost (distance) of shifting one mass from $\boldsymbol{x_i}$ to $\boldsymbol{y}_j$, where we use cosine distance $c(\boldsymbol{x}_i, \boldsymbol{y}_j) = 1 - \frac{\boldsymbol{x}_i^\top \boldsymbol{y}_j}{||\boldsymbol{x}_i||_2 ||\boldsymbol{y}_j||_2}$ as the cost function.

Computing Transport Plan as Alignments. Next, by regarding the two set of content embeddings $\boldsymbol{X}, \boldsymbol{Y}$ as two probability distributions, we calculate the optimal transport plan $\mathbf{T}^* \in \mathbb{R}^{m \times n}$ of transforming one distribution to the other. Here $\mathbf{T}^*$ is obtained via substituting X, Y into Eq.(2), i.e.,

$$\mathbf{T}^* = \underset{\mathbf{T}\in\Gamma(\mathbf{u},\mathbf{v})}{\arg\min}\,\langle \mathbf{T}, \mathbf{C}\rangle = \underset{\mathbf{T}\in\Gamma(\mathbf{u},\mathbf{v})}{\arg\min} \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{T}_{ij}\cdot c(\boldsymbol{x}_i, \boldsymbol{y}_j), \qquad (10)$$

where each element $\mathbf{T}^*_{ij}$ in $\mathbf{T}^*$ denotes how much mass should be shifted from $\boldsymbol{x}_i$ to $\boldsymbol{y}_j$. To be noted, the value of $\mathbf{T}^*_{ij}$ can be automatically optimized smaller if $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are not very correlated, i.e., having a high cost value $\mathbf{C}_{ij}$. In other words, $\mathbf{T}^*$ actually reflect the strength of correlations between the word-level content pair across two products. Therefore, after jointly optimized with the language model, we use $\mathbf{T}^*$ as the approximation to the alignment matrix.

Efficient Solver: IPOT. Unfortunately, it is computational intractable [1, 29] to compute the exact minimization over $\mathbf{T}$. Hence, to ensure an efficient training on large neural networks of language models, we propose to apply the recent introduced Inexact Proximal point method for Optimal Transport (IPOT) algorithm [35] to compute the optimal transport plan $\mathbf{T}^*$. IPOT approximates the exact solution by iteratively solving the following optimization problem:

$$\mathbf{T}^{(t+1)} = \underset{\mathbf{T}\in\Gamma(\mathbf{u},\mathbf{v})}{\arg\min} \left\{ \langle \mathbf{T}, \mathbf{C}\rangle + \beta\cdot\mathcal{B}(\mathbf{T}, \mathbf{T}^{(t)}) \right\} \qquad (11)$$

where $\mathcal{B}(\mathbf{T}, \mathbf{T}^{(t)})$ is the proximity metric term used to penalizes solutions that are too distant from the latest approximation. We do not choose Sinkhorn algorithm [6] to solve the efficiency issue as it is too sensitive to the choice of the hyper-parameter ε in experiments.

Loss Function of CEA. Lastly, we train the language model via optimizing the OT distance (i.e., Wasserstein distance) between the aligned content embeddings, with overall loss function defined as:

$$\mathcal{L}_{OTA}(X, Y) = \left\langle \mathbf{T}^*, \mathbf{C} \right\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{T}^*_{ij}\cdot c(\boldsymbol{x}_i, \boldsymbol{y}_j), \qquad (12)$$

$$\mathcal{L}_{OTA} = \frac{1}{|\mathcal{R}|} \sum_{(X,Y)\in\mathcal{R}} \mathcal{L}_{OTA}(X, Y) \qquad (13)$$

where $\mathcal{R}$ denotes the set of entity association based text pairs.

5 EXPERIMENTS

In this section, we conduct extensive experiments in the e-commerce domain to validate the effectiveness of the proposed framework. We first introduce the external and internal baselines compared in the paper. Next, we present the corpus as well as the auxiliary domain knowledge data used during pre-training. Besides, we elaborate the downstream tasks (definitions, datasets, performance metrics) for evaluating all the models. Lastly, we report the main performance comparison, ablation studies, case studies, and visualization of the OT-based alignments.

5.1 Baseline Models

External Baselines. In this paper, we compare our framework to following external baselines. (1) **BERT:** The vanilla BERT which is pre-trained on large-scale open-domain corpora by huggingface. (2) **BERT-PT** [36]: The vanilla BERT that is further **p**ost-**t**rained on review data. This can be considered as the domain-oriented vanilla BERT. (3) **BERT-NP:** The vanilla BERT using a different masking strategy, i.e., masks **n**oun **p**hrases instead of words. We contrast this method with another internal baseline (**DPM**) to reveal the effects of different phrase selection schemes. (4) **SpanBERT** [11]: An variant of BERT which masks spans of tokens instead of individual tokens. We compare with it to further validate the effect of different masking schemes. (5) **RoBERTa** [16]: A robustly optimized variant of BERT which deletes the Next Sentence Prediction task. (6) **ALBERT** [12]: A memory-efficient lite BERT that also high performances. To enable the above baselines (2)-(6) to be **domain-oriented**, like most existing work, we pre-train them on the same domain corpus as our method (except **BERT** for validating the effects of using domain corpus).

Internal Baselines. For ablation studies (validating the effects of each component in framework), we further compare with the following internal baselines: (1) **DPM:** The vanilla BERT that only masks **d**omain **p**hrases using our phrase pool, abandons word masking. (2) **DPM-R:** The vanilla BERT that only masks **d**omain **p**hrases and further employs the phrase **r**egularization term, abandons word masking. (3) **HM-R:** The vanilla BERT that **m**asks **d**omain **p**hrases and words in a **h**ybrid way (50%/50% of the time), employs the phrase **r**egularization term. (4) **AHM:** Adaptive Hybrid Masked Model, without leveraging entity association knowledge by Cross Entity Alignment. (5) **AHM+CEA:** The full version of our framework, combines AHM, OT-based CEA via continual multi-task learning. All internal baselines are pre-trained on the same domain corpus.

5.2 Domain-oriented Tasks and Metrics

We perform evaluations on four tasks of e-commerce. The definition, fine-tuning head and metric of each task is provided below.

Review Question Answering (Review QA). Given a question $q = \{q_i\}_{i=1}^{m}$ about a product and a related review snippet $r = \{r_i\}_{i=1}^{n}$, it aims to find the span $s = \{r_i\}_{i=s}^{e}$ from r that can answer q. We employ the same BERT fine-tuning head [7] as which on span-based QA to fine-tune this task, which maximizes the log-likelihoods of the correct start and end positions of the answer.

Review Aspect Extraction (Review AE). Given a review $r = \{r_i\}_{i=1}^{n}$, the task aims to find product aspects that reviewers have expressed opinions on. It is typically formalized as a sequence labeling task [36], in which each token is classified as one of $\{B, I, O\}$, and tokens between B and I are considered as extracted aspects. Following [36], we apply a dense layer and softmax layer on top of BERT output embeddings to predict the sequence labels.

Review Aspect Sentiment Classification (Review ASC). Given an aspect $a = \{a_i\}_{i=1}^{l}$ and the review sentence $r = \{r_i\}_{i=1}^{n}$ where a extracted from, this task aims to classify the sentiment polarity (positive, negative, or neutral) expressed on aspect a. For fine-tuning, following [36], both a and r are input into our framework, and we feed the `[CLS]` token to a dense layer and softmax layer to predict the polarity. Training loss is the cross entropy on the polarities.

Product Title Categorization (PTC). Given a product title $x = \{x_i\}_{i=1}^{n}$, the task aims to classify x using a predefined category collection C. Each title may belong to multiple categories, hence being a multi-label classification problem. We feed the embedding of `[CLS]` token to a dense layer and the multi-label classification head for fine-tuning.

Evaluation Metrics. For review QA, we adopt the standard evaluation script from SQuAD 1.1 [27] to report Precision, Recall, F_1

Table 3: The statistics of the pre-training datasets.

Resources	Volume	Size
Product Corpus	# titles and descriptions: 5,436,547	1.4 GB
Review Corpus	# product reviews: 9,636,112	2.3 GB
Domain Phrase Pool	# phrases: 536,332	–
Entity (Product) Association Graph	# product entities: 5,125,352 # entity associations: 6,484,325	–

Table 4: High-quality phrases of the e-commerce domain.

Category	Representative phrases
Automotive	jumper cables, cometic gasket, angel eyes, drink holder, static cling
Clothing, Shoes and Jewelry	high waisted jean, nike classic, removable tie, elegant victorian, vintage grey
Electronics	ipads tablets, SDHC memory card, memory bandwidth, auto switching
Office Products	decorative paper, heavy duty rubber, mailing labels, hybrid notebinder
Sports and Outdoors	basketball backboard, table tennis paddle, string oscillation, fishing tackles
Toys and Games	hulk hogan, augmented reality, teacup piggies, beam sabers, naruto uzumaki

scores, and Exact Match (EM). To evaluate review AE, we report Precision, Recall, and F_1 score. For review ASC, we report Macro-F_1 and Accuracy following [36]. Lastly, we adopt Accuracy (Acc), and Macro-F_1 to evaluate product title categorization.

5.3 Experimental Datasets

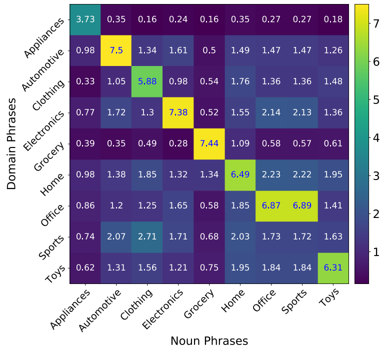
5.3.1 *Pre-training Resources*

In the paper, we collect and leverage a domain corpus and two domain knowledge datasets. Table 3 shows the datasets statistics and below presents the detailed collecting steps.

Domain Corpus. We extract millions of product titles, descriptions, and reviews from the Amazon Dataset[22] to build this corpus. The entire corpus consists of two sub-corpus, i.e., product corpus and review corpus. In the first corpus, each line corresponds to a product title and its description, while in the second, each line corresponds to a user comment on a specific product. The corpus serves as the foundation for language models to learn essential semantics of the e-commerce domain.

Domain Phrase Pool. To build the e-commerce domain phrase pool, we extract one million domain phrases from the above corpus leveraging AutoPhrase[3], a high efficient phrase mining algorithm, which is able to generate a quality score for each phrase based on corpus-level statistics such as *popularity, concordance, informativeness,* and *completeness.* Moreover, we filter out phrases that have a score lower than 0.5 to keep *quality domain phrases.* Table 4 shows the top-ranked phrases from six product categories.

Entity Association Graph. We build this graph to store the product entity associations in the form of associated entity pairs. In the paper, we only consider the "substitutable" associations and use a shopping pattern based heuristic method [19] to extract corresponding product pairs with this relation. We exploit all entity pairs in this graph to extract the same amount of associated text (title, description) pairs from the product corpus for the task of CEA.



The overlap between the domain phrases selected by our scheme and the general phrases obtained by chunking from the same domain corpus. As can be seen, the overlap ratio (%) is relatively low, indicating our phrases are more domain-oriented.

Figure 5: The overlap of different phrase sampling schemes.

Figure 5 presents when sampling phrases on the same domain corpus, the overlap between the results by our phrase pool based scheme and the ones by chunking based scheme. Results are reported based on nine categories of the product corpus. Each entry represents the ratio of the overlapped phrases to the general chunking based phrases. As can be seen, the overlap ratio is relatively low across all the sub categories, indicating our phrase pool based scheme yields more domain-oriented phrases.

5.3.2 *Task-specific Datasets*

For review QA, we evaluate on a newly released Amazon QA dataset [21], consisting of 8,967 product-related QA pairs. For the task of review AE and review ASC, we employ the laptop dataset of SemEval 2014 Task 4 [24] which contains 3,845 review sentences, 3,012 annotated aspects and the sentiment polarities on them. For product title categorization, we create an evaluation dataset by extracting a subset of Amazon metadata, consisting of 10,039 product titles and 98 categories. The first three datasets above are publicly available from prior works and we will share the fourth dataset in future. For all the datasets, we divide them into training/validation/testing set with the ratio of 7:1:2.

5.4 Implementation Details

Pre-training details. All the models are initialized with the same pre-trained BERT (the `bert-base-uncased` by Huggingface, with 12 layers, 768 hidden dimensions, 12 heads, 110M parameters). We post-train all the models (except BERT) on the domain corpus for 20 epochs, with batch size 32 and learning rate 1e-5. For our framework, we adopt Continual Multi-task Learning [32] to combine AHM and CEA. Specifically, we first train AHM alone on the entire corpus for 10 epochs with the same batch size and learning rate. Then, we train AHM and CEA jointly on the product corpus (with instances reformatted as text pairs by entity associations) for another 10 epochs. In AHM, to initialize α and ensure a stable training, we fix α^t = 0.6 for t=1∼1,000 (word learning mode is easier and provides preliminary knowledge, hence we weigh it more for initial iterations). For training OT-based CEA, we set $\beta = 0.5$ in the IPOT algorithm. All the pre-training is performed on a computational cluster with 8 NVIDIA GTX-1080-Ti GPUs with 20 days duration.

Fine-tuning Details. In each task, we adopt the same task-specific architecture (task head) as aforementioned for all the models. We

[3] https://github.com/shangjingbo1226/AutoPhrase

Table 5: Performance comparison of baselines and our model on the e-commerce downstream tasks (%).

Models\Tasks		**Review QA**				**Review AE**			**Review ASC**		**PTC**	
		P.	*R.*	*F1*	*EM*	*P.*	*R.*	*F1*	*Acc.*	*Ma-F1*	*Acc.*	*Ma-F1*
External Baselines	**BERT**	58.91	62.18	60.50	40.22	83.15	84.66	83.90	86.01	62.87	78.76	76.83
	BERT-PT	60.28	62.25	61.25	41.23	84.33	84.09	84.21	86.43	64.96	80.41	78.99
	BERT-NP	61.39	64.57	62.94	43.35	85.23	85.71	85.47	85.79	63.21	81.19	79.32
	SpanBERT	62.52	64.77	63.63	43.94	85.67	86.22	85.94	86.56	65.13	81.36	80.11
	RoBERTa	62.76	63.98	63.36	44.12	85.82	86.51	86.16	86.71	65.34	82.57	81.25
	ALBERT	61.89	63.80	62.83	43.82	85.35	86.01	85.68	86.47	65.12	81.45	80.79
Internal Baselines	**DPM**	63.76	67.02	65.35	44.53	86.81	88.47	87.63	87.84	68.02	83.12	81.92
	DPM-R	64.47	68.11	66.24	44.98	87.29	89.24	88.25	88.45	69.18	84.18	82.34
	HM-R	64.69	68.24	66.42	45.25	87.38	89.32	88.34	88.69	69.21	84.30	82.40
	AHM	65.38	68.93	67.11	45.86	**88.12**	89.89	88.99	88.93	69.35	85.01	82.68
	AHM + CEA	**67.21**	**70.13**	**68.64**	**46.98**	88.05	**90.55**	**89.28**	**89.32**	**70.55**	**86.32**	**83.12**

choose the learning rate and epochs from {5e-6, 1e-5, 2e-5, 5e-5} and {2,3,4,5} respectively. For each task and each model, we pick the best learning rate and number of epochs on the development set and report the corresponding test results. We found the setting that works best across most tasks and models is 2 or 4 epochs and a learning rate of 2e-5. Results are reported as averages of 10 runs.

5.5 Experimental Results

5.5.1 *Main Results Analysis*

Table 5 presents the performance comparison of all the baselines and our framework on the four tasks. The key observations and conclusions are: (1) Our framework (**AHM, AHM+CEA**) easily outperforms all the external baselines by a large margin (**4.1%** in average), indicating the effectiveness of our general idea, i.e., leveraging auxiliary domain knowledge to enhance domain-oriented language modeling. (2) **BERT-PT** outperforms **BERT**, proving that for domain-oriented tasks, capturing domain semantics by pre-training on a domain corpus is necessary. (3) *The effects of different masking schemes*: **BERT-NP** and **SpanBERT** can perform better consistently than **BERT-PT**, indicating the advantage of phrase/span based masking strategy over word based masking strategy. (4) *The effects of different phrase selection schemes*: **DPM** achieves more improvements over **BERT-NT** and **SpanBERT**, certificating that the domain phrase pool based sampling outperforms general chunking based phrase sampling. We attribute this to that: the domain phrase pool, serving as a "supervisor", enables the language model to "focus" more on *domain-oriented phrases*, and these phrases have more effects over the downstream tasks.

5.5.2 *Ablation Studies*

The bottom of Table 5 shows the performance comparison of the internal baselines. As can be seen, (1) **DPM-R** outperforms **DPM**, validating the effectiveness of the proposed *phrase regularization term*. Compared with reconstructing phrases by tokens, it encourages complete phrase reconstruction, leads to a more accurate phrase perception learning. (2) **HM-R** uses hybrid masking in a straightforward way, achieves slightly better performances than **DPM-R**. Besides, **AHM** achieves more improvements on **DPM-R** than **HM-R**. This indicates that both word learning and phrase learning are essential for language models, and the adaptive hybrid learning method is a more solid way to combine them. (3) **AHM+CEA** further improves the performances by **0.5%~1.2%** over **AHM** on the four tasks, certificating the effectiveness of our idea of leveraging entity association knowledge to augment semantic learning. Moreover, the proposed OT-based alignment pre-train task can successfully exploit the hidden co-occurrence signals in entity association based text pairs.

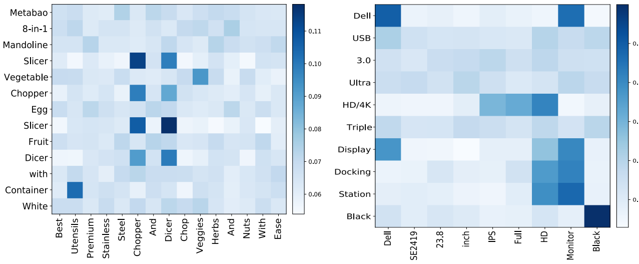
Table 6: Case studies of Aspect Extraction (AE). Given a review, it aims to extract specific product "aspects" that are discussed. Ground-truth answers are marked in color. For answers consisting of multi-word phrases, our model make more comprehensive predictions than BERT.

Review	**Model**	**Extracted Aspects**
We love the size of the screen, although it is still light-weight and very easy to tote around. The resolution is perfect for living room.	BERT-PT	screen, **resolution**
	Ours	size of the screen, **resolution**
That included the extra Sonic Stage software, the speakers and the subwoofer I got (that WAS worth the money), the bluetooth mouse for my supposedly bluetooth enabled computer, the extended life battery and the docking port.	BERT-PT	software, **bluetooth mouse**, battery
	Ours	Sonic Stage software, **bluetooth mouse**, battery, **docking port**

5.5.3 *Case Studies and Visualizations*

Table 6 shows a case study of the review aspect extraction task. We compare our model with BERT-PT, both are pre-trained on the same domain corpus, and employ the same fine-tuning architecture and task-specific dataset. As can be seen, for "aspects" that span multiple words, our model offers better predictions than BERT-PT in terms of the phrase completeness (*size of the screen* vs *screen*) . This indicates that our model indeed possesses fine phrase perception ability needed for *phrase-intensive tasks*.

Figure 6 presents the visualization of the optimal transport alignment for two product pairs, where darker color indicates stronger correlations. Example (a) is about `Mandoline Slicer` and `Steel Chopper`, example (b) is about a `Docking station` and `Dell Monitor`. As can be seen, in both examples, OT alignments are sparser and offers better interpretability, with meaningful word alignment pairs being discovered automatically (*Slicer* vs *Chopper*, *Vegetables* vs *Veggies*, *Monitor* vs *Display*). This certificates that the



(a) Two kitchen products. (b) Two electronic products.

Figure 6: Visualizing the optimal transport plan in two real examples.

OT-based alignment task can indeed benefit semantic learning by automatically correlating similar words/phrases across entity pairs.

6 CONCLUSION

In this paper, we introduced how to improve domain-oriented language modeling by leveraging auxiliary domain knowledge. Specifically, we proposed a generalized pre-training framework enhancing existing works from two perspectives. First, we developed Adaptive Hybrid Masked Model (AHM) to incorporate auxiliary domain phrase knowledge. Second, we designed Cross Entity Alignment (CEA) to leverage entity association as weak supervision for augmenting the semantic learning of pre-trained models. Without the loss of generalization, we performed the experimental validation on four downstream e-commerce tasks. The results showed that incorporating phrase knowledge via AHM can improve the performance on all the tasks, especially the phrase-intensive ones. Also, utilizing the entity association knowledge via CEA can further improve the performances and the learned alignments revealed meaningful semantic correlation across word pairs.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *ICML*. PMLR, 214–223.

[2] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *EMNLP*. 3606–3611.

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[4] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

[5] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. 2020. Graph optimal transport for cross-domain alignment. In *ICML*. PMLR, 1542–1553.

[6] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS* 26 (2013), 2292–2300.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[8] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

[9] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv preprint arXiv:2004.10964*.

[10] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

[11] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *TACL* 8 (2020), 64–77.

[12] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

[13] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[14] Manling Li, Denghui Zhang, Yantao Jia, Yuanzhuo Wang, and Xueqi Cheng. 2018. Link Prediction in Knowledge Graphs: A Hierarchy-Constrained Approach. *IEEE Transactions on Big Data* (2018).

[15] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *SIGKDD*. 1054–1064.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[17] Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. 2018. Differential properties of sinkhorn approximation for learning with wasserstein distance. *arXiv preprint arXiv:1805.11897*.

[18] Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain Adaptation with BERT-based Domain Classification and Data Selection. In *DeepLo*. 76–83.

[19] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *SIGKDD*. 785–794.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*. 3111–3119.

[21] John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The Effect of Natural Distribution Shift on Question Answering Models. *arXiv preprint arXiv:2004.14444*.

[22] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP*. 188–197.

[23] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *NAACL*. 2227–2237.

[24] Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *SemEval*.

[25] Mohiuddin Md Abdul Qudar and Vijay Mago. 2020. TweetBERT: A Pretrained Language Representation Model for Twitter Text Analysis. *arXiv preprint arXiv:2010.11091*.

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

[27] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*.

[28] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification. In *LREC*. 4933–4941.

[29] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. 2018. Improving GANs using optimal transport. *arXiv preprint arXiv:1803.05573*.

[30] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *TKDE* 30, 10 (2018), 1825–1837.

[31] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

[32] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding.. In *AAAI*. 8968–8975.

[33] Ying Sun, Fuzhen Zhuang, Hengshu Zhu, Qi Zhang, Qing He, and Hui Xiong. 2021. Market-oriented job skill valuation with cooperative composition neural network. *Nature communications* 12, 1 (2021), 1–12.

[34] Cédric Villani. 2008. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.

[35] Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. 2020. A fast proximal point method for computing exact wasserstein distance. In *UAI*. PMLR, 433–453.

[36] Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In *NAACL*. 2324–2335.

[37] Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2020. DomBERT: Domain-oriented Language Model for Aspect-based Sentiment Analysis. In *EMNLP*. 1725–1731.

[38] Zixuan Yuan, Hao Liu, Renjun Hu, Denghui Zhang, and Hui Xiong. 2021. Self-Supervised Prototype Representation Learning for Event-Based Corporate Profiling. *AAAI 2021* 35, 5 (May 2021), 4644–4652. https://ojs.aaai.org/index.php/AAAI/article/view/16594

[39] Zixuan Yuan, Hao Liu, Yanchi Liu, Denghui Zhang, Fei Yi, Nengjun Zhu, and Hui Xiong. 2020. Spatio-temporal dual graph attention network for query-poi matching. In *SIGIR*. 629–638.

[40] Denghui Zhang, Manling Li, Yantao Jia, Yuanzhuo Wang, and Xueqi Cheng. 2017. Efficient parallel translating embedding for knowledge graphs. In *WI*. 460–468.

[41] Denghui Zhang, Junming Liu, Hengshu Zhu, Yanchi Liu, Lichen Wang, Pengyang Wang, and Hui Xiong. 2019. Job2Vec: Job title benchmarking with collective multi-view representation learning. In *CIKM*. 2763–2771.