K-PLUG: KNOWLEDGE-INJECTED PRE-TRAINED LANGUAGE MODEL FOR NATURAL LANGUAGE UNDERSTANDING AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing pre-trained language models (PLMs) have demonstrated the effectiveness of self-supervised learning for a broad range of natural language processing (NLP) tasks. However, most of them are not explicitly aware of domain-specific knowledge, which is essential for downstream tasks in many domains, such as tasks in e-commerce scenarios. In this paper, we propose K-PLUG, a knowledge-injected pre-trained language model based on the encoder-decoder transformer that can be transferred to both natural language understanding and generation tasks. We verify our method in a diverse range of e-commerce scenarios that require domain-specific knowledge. Specifically, we propose five knowledge-aware self-supervised pre-training objectives to formulate the learning of domain-specific knowledge, including e-commerce domain-specific knowledge-bases, aspects of product entities, categories of product entities, and unique selling propositions of product entities. K-PLUG achieves new state-of-the-art results on a suite of domain-specific NLP tasks, including product knowledge base completion, abstractive product summarization, and multi-turn dialogue, significantly outperforms baselines across the board, which demonstrates that the proposed method effectively learns a diverse set of domain-specific knowledge for both language understanding and generation tasks. The code, data, and models will be publicly available[1].

1 INTRODUCTION

Pre-trained language models (PLMs), such as ELMo (Peters et al., 2018), GPT (Radford et al., 2018), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019), have made remarkable breakthroughs in many natural language understanding (NLU) tasks, including text classification, reading comprehension, and natural language inference. These models are trained on large-scale text corpora with self-supervision based on either bi-directional or auto-regressive pre-training. Equally promising performances have been achieved in natural language generation (NLG) tasks, such as machine translation and text summarization, by MASS (Song et al., 2019), UniLM (Dong et al., 2019), BART (Lewis et al., 2020), T5 (Raffel et al., 2019), PEGASUS (Zhang et al., 2020), and ProphetNet (Yan et al., 2020). In contrast, these approaches adopt Transformer-based sequence-to-sequence models to jointly pre-train for both the encoder and the decoder.

While these PLMs can learn rich semantic patterns from raw text data and thereby enhance downstream NLP applications, many of them do not explicitly model domain-specific knowledge. As a result, they may not be as sufficient for capturing human-curated or domain-specific knowledge that is necessary for tasks in a certain domain, such as tasks in e-commerce scenarios. In order to overcome this limitation, several recent studies have proposed to enrich PLMs with explicit knowledge, including knowledge base (KB) (Zhang et al., 2019; Peters et al., 2019; Xiong et al., 2020; Wang et al., 2019; 2020), lexical relation (Lauscher et al., 2019; Wang et al., 2020), word sense (Levine et al., 2020), part-of-speech tag (Ke et al., 2019), and sentiment polarity (Ke et al., 2019; Tian et al., 2020). However, these methods only integrate domain-specific knowledge into the encoder, and the decoding process in many NLG tasks benefits little from these knowledge.

[1]Our code is available at `https://github.com/ICLR21Anonymous/knowledge_pretrain`.

To mitigate this problem, we propose a **K**nowledge-injected **P**re-trained **L**anguage model that is suitable for both Natural Language **U**nderstanding and **G**eneration (K-PLUG). Different from existing knowledge-injected PLMs, K-PLUG integrates knowledge into pre-training for both the encoder and the decoder, and thus K-PLUG can be adopted to both downstream knowledge-driven NLU and NLG tasks. We verify the performance of the proposed method in various e-commerce scenarios. In the proposed K-PLUG, we formulate the learning of four types of domain-specific knowledge: e-commerce domain-specific knowledge-bases, aspects of product entities, categories of product entities, and unique selling propositions (USPs) (Garrett, 1961) of product entities. Specifically, e-commerce KB stores standardized product attribute information, product aspects are features that play a crucial role in understanding product information, product categories are the backbones for constructing taxonomies for organization, and USPs are the essence of what differentiates a product from its competitors. K-PLUG learns these types of knowledge into a unified PLM, enhancing performances for various language understanding and generation tasks.

To effectively learn these four types of valuable domain-specific knowledge in K-PLUG, we proposed five new pre-training objectives: knowledge-aware masked language model (KMLM), knowledge-aware masked sequence-to-sequence (KMS2S), product entity aspect boundary detection (PEABD), product entity category classification (PECC), and product entity aspect summary generation (PEASG). Among these objectives, KMLM and KMS2S learn to predict the masked single and multiple tokens, respectively, that are associated with domain-specific knowledge rather than general information; PEABD determines the boundaries between descriptions of different product aspects given full product information; PECC identifies the product category that each product belongs to; and PEASG generates a summary for each individual product aspect based on the entire product description.

After pre-training K-PLUG, we fine-tune it on three domain-specific NLP tasks, namely, e-commerce knowledge base completion, abstractive product summarization, and multi-turn dialogue. The results show that K-PLUG significantly outperforms comparative models on all these tasks.

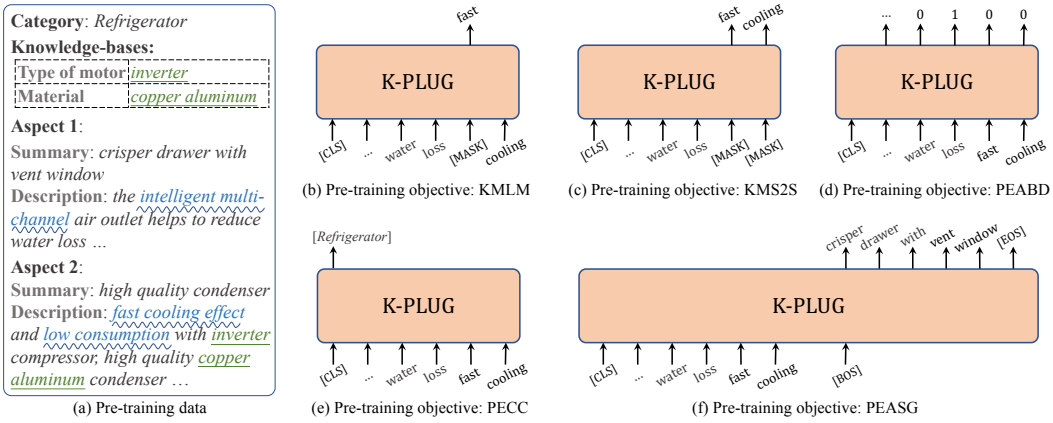
Our main contributions can be summarized as follows:

- We present K-PLUG that learns domain-specific knowledge for both the encoder and the decoder in a pre-training language model framework, which benefits both NLG and NLU tasks.
- We formulate the learning of four types of domain-specific knowledge in e-commerce scenarios: e-commerce domain-specific knowledge-bases, aspects of product entities, categories of product entities, and unique selling propositions of product entities, which provide critical information for many applications in the domain of e-commerce. Specifically, five self-supervised objectives are proposed to learn these four types of knowledge into a unified PLM.
- Our proposed model exhibits clear effectiveness in many downstream tasks in the e-commerce scenario, including e-commerce KB completion, abstractive product summarization, and multi-turn dialogue.

2 Related Work

2.1 PLMs in General

Unsupervised pre-training language model has been successfully applied to many NLP tasks. ELMo (Peters et al., 2018) learns the contextual representations based on a bidirectional LM. GPT (Radford et al., 2018) predicts tokens based on the context on the left-hand side. BERT (Devlin et al., 2019) adopts a bi-directional LM to predict the masked tokens. XLNet (Yang et al., 2019) predicts masked tokens in a permuted order through an autoregressive method. MASS (Song et al., 2019) pre-trains the sequence-to-sequence LM to recover a span of masked tokens. UniLM (Dong et al., 2019) combines bidirectional, unidirectional, and sequence-to-sequence LMs. T5 (Raffel et al., 2019) and BART (Lewis et al., 2020) present denoising sequence-to-sequence pre-training. PEGASUS (Zhang et al., 2020) pre-trains with gap-sentence generation objective. While human-curated or domain-specific knowledge is essential for downstream knowledge-driven tasks, these methods do not explicitly consider external knowledge like our proposed K-PLUG.



(a) Pre-training data (b) Pre-training objective: KMLM (c) Pre-training objective: KMS2S (d) Pre-training objective: PEABD (e) Pre-training objective: PECC (f) Pre-training objective: PEASG

Figure 1: Pre-training data consists of 25 million textual product descriptions depicting multiple product aspects. We define knowledge as *e-commerce knowledge-bases*, *aspects of product entities*, *categories of product entities*, and *unique selling propositions of product entities*. Pre-training objectives include knowledge-aware masked language model (KMLM), knowledge-aware masked sequence-to-sequence (KMS2S), product entity aspect boundary detection (PEABD), product entity category classification (PECC), and product entity aspect summary generation (PEASG).

2.2 Injecting Knowledge into PLMs

Recent work investigates how to incorporate knowledge into PLMs for NLU. ERNIE (Sun et al., 2019) enhances language representation with the entity/phrase-level masking. ERNIE (Zhang et al., 2019) identifies and links entity mentions in texts to their corresponding entities in KB. Similar to ERNIE (Zhang et al., 2019), KnowBERT (Peters et al., 2019) injects KBs into PLM. Xiong et al. (2020) leverages an entity replacement pre-training objective to learn better representations for entities. KEPLER (Wang et al., 2019) adopts the knowledge embedding objective in the pre-training. Besides, SKEP (Tian et al., 2020), SenseBERT (Levine et al., 2020), SentiLR (Ke et al., 2019), and K-ADAPTER (Wang et al., 2020) propose to integrate sentiment knowledge, word sense, sentiment polarity, and lexical relation into PLM, respectively. However, most of these studies are focused on integrating knowledge for language understanding task, work of utilizing domain-specific knowledge for pre-training for language generation tasks are limited. Inspired by these work, we construct K-PLUG that learns domain-specific knowledge into a PLM for both NLU and NLG tasks.

3 Knowledge-injected Pre-training

In this section, we explain the data used to pre-train K-PLUG, its model architecture, and our pre-training objectives.

3.1 Data Preparation

We collect the pre-training data from a mainstream Chinese e-commerce platform, which contains approximately 25 million textual product descriptions and covers 40 product categories. With an average length of 405 tokens, these product descriptions constitute a corpus with a size of 10B Chinese characters. Each product description consists of information on 10.7 product aspects on average, and each product aspect is accompanied with a summary highlighting its prominent features, as shown in Figure 1(a). Additionally, the e-commerce KB and USPs (further explained below) used in our pre-training data are as specified by the e-commerce platform and its online stores.

3.2 Model Architecture

K-PLUG adopts the standard sequence-to-sequence Transformer architecture (Vaswani et al., 2017), consisting of a 6-layer encoder and a 6-layer decoder as Song et al. (2019). We set the size of hidden

vectors as 768, and the number of self-attention heads as 12. More details about the experimental settings are in the appendix.

3.3 KNOWLEDGE FORMULATION AND PRE-TRAINING OBJECTIVES

We formulate the learning of four types of knowledge in a unified PLM: e-commerce KB, aspects of product entities, categories of product entities, and USPs of product entities. Specifically, **e-commerce KB** stores standardized product attribute information, *e.g.*, (*Material*: *Cotton*) and (*Collar Type*: *Pointed Collar*). It provides details about the products (Logan IV et al., 2017). **Aspects of product entities** are features of a product, such as the *sound quality* of a stereo speaker, etc. (Li et al., 2020). **Categories of product entities** such as *Clothing* and *Food* are widely used by e-commerce platforms to organize their products so to present structured offerings to their customers (Luo et al., 2020; Dong et al., 2020) **USPs of product entities** are the essence of what differentiates a product from its competitors Garrett (1961). For example, a stereo speaker's USP exhibiting its supreme sound quality could be "*crystal clear stereo sound*". An effective USP immediately motivates the purchasing behavior of potential buyers.

We propose and evaluate five novel self-supervised pre-training objectives to learn the above-mentioned four types of knowledge in the K-PLUG model (see Figure 1).

Knowledge-aware Masked Language Model (KMLM)

Inspired by BERT (Devlin et al., 2019), we adopt the masked language model (MLM) as one of our pre-training objectives, which learns to predict the masked tokens in the source sequence (*e.g.*, "The company is [MASK] at the foot of a hill."). Similar to BERT, we mask 15% of all tokens in a text sequence; 80% of the masked tokens are replaced with the [MASK] token, 10% with a random token, and 10% left unchanged. Particularly, given an original text sequence $\boldsymbol{x} = (x_1, ..., x_m, ..., x_M)$ with M tokens, a masked sequence is produced by masking x_m through one of the three ways explained above, *e.g.*, replacing x_m with [MASK] to create $\widetilde{\boldsymbol{x}} = (x_1, ..., [\mathrm{MASK}], ..., \mathrm{x_M})$. MLM aims to model the conditional likelihood $P(x_m|\widetilde{\boldsymbol{x}})$, and the loss function is:

$$L_{MLM} = \log P(x_m|\widetilde{\boldsymbol{x}}) \tag{1}$$

The major difference from BERT is that our KMLM prioritizes knowledge tokens when selecting positions to mask and, in the case that the knowledge tokens make up less than 15% of all tokens, randomly picks non-knowledge tokens to complete the masking.

Knowledge-aware Masked Sequence-to-Sequence (KMS2S)

K-PLUG inherits the strong ability of language generation from the masked sequence-to-sequence (MS2S) objective. The encoder takes a sentence with a masked fragment (several consecutive tokens) as the input, and the decoder predicts this masked fragment conditioned on the encoder representations (*e.g.*, "The company [MASK] [MASK] [MASK] the foot of a hill.").

Given a text sequence $\boldsymbol{x} = (x_1, ..., x_u, ..., x_v, ..., x_M)$, a masked sequence $\widetilde{\boldsymbol{x}} = (x_1, ..., [\mathrm{MASK}], ..., [\mathrm{MASK}], ..., \mathrm{x_M})$ is produced by replacing the span $\boldsymbol{x}_{u:v}$, ranging from x_u to x_v, with the [MASK] token. MS2S aims to model $P(\boldsymbol{x}_{u:v}|\widetilde{\boldsymbol{x}})$, which can be further factorized into a product $P(\boldsymbol{x}_{u:v}|\widetilde{\boldsymbol{x}}) = \prod_{t=u}^{v} P(x_t|\widetilde{\boldsymbol{x}})$ according to the chain rule. The loss function is:

$$L_{MS2S} = \sum_{t=u}^{v} \log P(x_t|\widetilde{\boldsymbol{x}}) \tag{2}$$

We set the length of the masked span as 30% of the length of the original text sequence. Similar to KMLM, KMS2S prioritizes the masking of text spans that cover knowledge tokens.

Product Entity Aspect Boundary Detection (PEABD)

A product description usually contains multiple product entity aspects. Existing work (Li et al., 2020) proves that product aspects influence the quality of product summaries from the views of importance, non-redundancy, and readability, which are not directly taken into account in language modeling. In order to train a model that understands product aspects, we leverage the PEABD objective to detect boundaries between the product entity aspects. It is essentially a sequence labeling task based on the representations of K-PLUG's top encoder layer.

Given a text sequence $\boldsymbol{x} = (x_1, ..., x_M)$, the encoder of K-PLUG outputs a sequence $\boldsymbol{h} = (h_1, ..., h_M)$, which is fed into a softmax layer, and generates a probability sequence $\boldsymbol{y}$. The loss function is:

$$L_{PEABD} = -\sum_t \hat{y}_t \log y_t \tag{3}$$

where $\boldsymbol{y} \in \{[0, 1]\}$ are the ground-truth labels for the aspect boundary detection task.

Product Entity Category Classification (PECC)

Product entity categories are the backbones for constructing taxonomies (Luo et al., 2020; Dong et al., 2020). Each product description document corresponds to one of the 40 categories included in our corpus, such as *Clothing*, *Bags*, *Home Appliances*, *Shoes*, *Foods*, etc. Identifying the product entity categories accurately is the prerequisite for creating an output that is consistent with the input.

Given a text sequence $\boldsymbol{x} = (x_1, ..., x_M)$, a softmax layer outputs the classification score, y, based on the representation of the classification token, [CLS]. The loss function maximizes the model's probability of outputting the true product entity category as follows:

$$L_{PECC} = -\hat{y} \log y \tag{4}$$

where $\hat{y}$ is the ground-truth product category.

Product Entity Aspect Summary Generation (PEASG)

Inspired by PEGASUS (Zhang et al., 2020), which proves that using a pre-training objective that more closely resembles the downstream task leads to better and faster fine-tuning performance, we propose a PEASG objective to generate a summary from the description of a product entity aspect. Unlike extracted gap-sentences generation in PEGASUS, our method constructs a more realistic summary generation task because the aspect summary naturally exists in our pre-training data.

Given a aspect description sequence $\boldsymbol{x} = (x_1, ..., x_M)$, and a aspect summary sequence $\boldsymbol{y} = (y_1, ..., y_T)$, PEASG aims to model the conditional likelihood $P(\boldsymbol{y}|\boldsymbol{x})$. The loss function is:

$$L_{PEASG} = \sum_t \log P(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) \tag{5}$$

4 Experiments and Results

4.1 Pre-trained Model Variants

To evaluate the effectiveness of pre-training with domain-specific data and with domain-specific knowledge separately, we implement pre-training experiments with two model variants: C-PLUG and E-PLUG, whose configurations are the same as that of K-PLUG.

- **C-PLUG** is a pre-trained language model with the original objectives of MLM and MS2S, trained with a general pre-training corpus, CLUE (Xu et al., 2020), which contains 30GB of raw text with around 8B Chinese words.
- **E-PLUG** is a pre-trained language model with the original objectives of MLM and MS2S, trained with our collected e-commerce domain-specific corpus.

4.2 Downstream Tasks

We fine-tune K-PLUG on three downstream tasks: e-commerce KB completion, abstractive product summarization, and multi-turn dialogue. The e-commerce KB completion task involves the prediction of product attributes and values given product information. The abstractive product summarization task requires the model to generate a product summary from textual product description. The multi-turn dialogue task aims to output the response by utilizing a multi-turn dialogue context. The domain-specific knowledge we defined in this paper is essential for these tasks.

4.2.1 E-COMMERCE KB COMPLETION

Task Definition. E-commerce KB provides abundant product information that is in the form of (*product entity*, *product attribute*, *attribute value*), such as (*pid#133443*, *Material*, *Copper Aluminum*). For the E-commerce KB completion task, the input is a textual product description for a given product, and the output is the product attribute values.

Datasets. We conduct experiments on the dataset of MEPAVE (Zhu et al., 2020). This dataset is collected from a major Chinese e-commerce platform, which consists of 87,194 instances annotated with the position of attribute values mentioned in the product descriptions. There are totally 26 types of product attributes such as *Material*, *Collar Type*, *Color*, etc. The training, validation, and testing sets contain 71,194/8,000/8,000 instances, respectively.

Models. We consider the e-commerce KB completion task as a sequence labeling task that tags the input word sequence $x = (x_1, ..., x_N)$ with the label sequence $y = (y_1, ..., y_N)$ in the BIO format. For example, for the input sentence "*A bright yellow collar*", the corresponding labels for "*bright*" and "*yellow*" are *Color-B* and *Color-I*, respectively, and *O* for the other tokens. For an input sequence, K-PLUG outputs an encoding representation sequence, and a linear classification layer with the softmax predicts the label for each input token based on the encoding representation.

Baselines.

- **ScalingUp** Xu et al. (2019) adopts BiLSTM, CRF, and attention mechanism to extract attributes.
- **JAVE** is a joint attribute and value extraction model based on a pre-trained BERT.
- **M-JAVE** is a multimodal JAVE model, which additionally utilizes product image information.

Table 1: Experimental results with the F1 score for the e-commerce KB completion task. The results in the first block are taken from Zhu et al. (2020).

Model	F1
LSTM	82.92
ScalingUp	77.12
BERT	83.12
JAVE	84.78
M-JAVE	87.17
C-PLUG	93.02
E-PLUG	93.20
K-PLUG	**95.97**

Results.

Table 1 shows the experimental results. We observe that our K-PLUG performs better than baselines. C-PLUG achieves significantly better performance than BERT, which indicates that MS2S can also benefit the NLU task. E-PLUG outperforms C-PLUG, showing that training with domain-specific corpus is helpful. K-PLUG further exhibits a 2.51% improvement compared with E-PLUG. In short, we can conclude that the improvement is due to both the domain-specific pre-training data and knowledge-injected pre-training objectives.

4.2.2 ABSTRACTIVE PRODUCT SUMMARIZATION

Task Definition. Abstractive product summarization task aims to capture the most attractive information of a product that resonates with potential purchasers. The input for this task is a product description, and the output is a condensed product summary.

Datasets. We perform experiments on the dataset of Li et al. (2020), which contains 1.4 million instances collected from a major Chinese e-commerce platform, covering three categories of product: *Home Appliances*, *Clothing*, and *Cases & Bags*. Each instance in the dataset is a (product information, product summary) pair, and the product information contains an image, a title, and other product descriptions. In our work, we do not consider the visual information of products.

Models. Abstractive product summarization task is an NLG task that takes the product description as the input and product summary as the output.

Baselines.

- **LexRank** Erkan & Radev (2004) is a graph-based extractive summarization method.
- **Seq2seq** Bahdanau et al. (2015) is a standard seq2seq model with an attention mechanism.
- **Pointer-Generator (PG)** See et al. (2017) is a seq2seq model with a copying mechanism.
- **Aspect MMPG** Li et al. (2020) is the-state-of-the-art method for abstractive product summarization, taking both textual and visual product information as the input.

Table 2: Experimental results with the ROUGE score for the abstractive product summarization task. The results in bold are the best performances among the models taking only texts as the input, and * denotes the model taking both product images and texts as the input. The results in the first and second blocks are taken from Li et al. (2020).

Model	Home Applications			Clothing			Cases&Bags		
	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L
LexRank	24.06	10.01	18.19	26.87	9.01	17.76	27.09	9.87	18.03
Seq2seq	21.57	7.18	17.61	23.05	6.84	16.82	23.18	6.94	17.29
MASS	28.19	8.02	18.73	26.73	8.03	17.72	27.19	9.03	18.17
PG	31.11	10.93	21.11	29.11	9.24	19.92	31.31	10.27	21.79
Aspect MMPG*	34.36	12.52	22.35	31.93	11.09	21.54	33.78	12.51	22.43
C-PLUG	32.75	11.62	21.76	31.73	10.86	20.37	32.04	10.75	21.85
E-PLUG	33.11	12.07	22.01	32.61	11.03	20.98	32.37	11.14	21.98
K-PLUG	**33.56**	**12.50**	**22.15**	**33.00**	**11.24**	**21.43**	**33.87**	**11.83**	**22.35**

Results. Table 2 shows the experimental results, including ROUGE-1 (RG-1), ROUGE-2 (RG-2), and ROUGE-L (RG-L) F1 scores Lin & Hovy (2003). K-PLUG clearly performs better than other text-based methods. E-commerce knowledge plays a significant role in the abstractive product summarization task, and domain-specific pre-training data and knowledge-injected pre-training objectives both enhance the model. K-PLUG achieves comparable results with the multimodal model, Aspect MMPG. The work of Li et al. (2020) suggests that product images are essential for this task, and we will advance K-PLUG with multimodal information in the future.

4.2.3 MULTI-TURN DIALOGUE

Task Definition. The multi-turn dialogue task aims to output a response based on the multi-turn dialogue context (Shum et al., 2018). The input for this task is the dialogue context consisting of previous question answering, and the output is the response to the last question.

Datasets. We conduct experiments on two datasets of **JDDC** (Chen et al., 2020) and **ECD** (Zhang et al., 2018). **JDDC** is collected from the conversations between users and customer service staffs from a popular e-commerce website in China and contains 289 different intents from after-sales assistance. There are 1,024,196 multi-turn sessions and 20,451,337 utterances in total. The average number of turns for each session is 20, and the average tokens per utterance is about 7.4. After pre-processing, the training, validation, and testing sets include 1,522,859/5,000/5,000 (dialogue context, response) pairs, respectively. **ECD** is collected from another popular e-commerce website in China and covers over 5 types of conversations based on 20 commodities. Additionally, for each ground-truth response, negative responses are provided for discriminative learning. The training, validation, and testing sets include 1,000,000/10,000/10,000 (dialogue context, response) pairs, respectively.

Models. We test with two types of K-PLUG: retrieval-based K-PLUG on the ECD dataset and generative-based K-PLUG on the JDDC dataset. For the retrieval-based approach, we concatenate the dialogue context and use [SEP] token to separate context and response. The [CLS] representation is fed into the output layer for classification. The generative-based approach is a sequence-to-sequence model, which is the same as the model adopted in the abstractive product summarization task.

Baselines. The baselines also include both the retrieval-based (BM25, TF-IDF, CNN, BiLSTM, and BERT) and generative-based approaches. Other baselines are as follows.

- **Sequential Matching Network (SMN)** (Wu et al., 2017) matches a response with each utterance in the context.
- **Deep Utterance Aggregation (DUA)** (Zhang et al., 2018) is a deep utterance aggregation model based on the fine-grained context representations.

Results. Table 3 and 4 show the experimental results on the JDDC and ECD datasets, respectively. We report ROUGE-L (RG-L) F1, BLEU, and recall at position k in n candidates ($R_n@k$). We can observe that, both on the retrieval-based and generative-based tasks, K-PLUG achieves new state-of-the-art results, and e-commerce knowledge presents consistent improvements. K-PLUG is evidently superior to BERT, possibly due to BERT's lack of domain-specific knowledge for pre-training with the general MLM objective.

Table 3: Experimental results for the multi-turn conversation task on the JDDC dataset. The results in the first block are taken from Chen et al. (2020).

Model	RG-L	BLEU
BM25	19.47	9.94
BERT	19.90	10.27
SeqSeq	22.17	14.15
PG	23.62	14.27
C-PLUG	25.47	16.75
E-PLUG	25.93	17.12
K-PLUG	**26.60**	**17.80**

Table 4: Experimental results for the multi-turn conversation task on the ECD dataset. The results in the first block are taken from Zhang et al. (2018).

Model	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
TF-IDF	15.9	25.6	47.7
CNN	32.8	51.5	79.2
BiLSTM	35.5	52.5	82.5
SMN	45.3	65.4	88.6
DUA	50.1	70.0	92.1
BERT	54.3	73.4	94.3
C-PLUG	62.7	76.8	95.0
E-PLUG	65.8	80.1	95.6
K-PLUG	**73.5**	**82.9**	**96.4**

4.3 Ablation Studies

To better understand our model, we perform ablation experiments to study the effects of different pre-training objectives.

Results. The ablation results are shown in Table 5. We can conclude that the lack of any pre-training objective hurts performance across all the tasks. KMS2S is the most effective objective for the abstractive product summarization and generative conversation tasks since this objective is highly close to the essence of NLG. Product-aspect-related objectives, *i.e.*, PEABD and PEASG, contribute much to the abstractive product summarization task, which proves that this task requires comprehensively understanding the product description from the view of product aspects, going beyond individual tokens.

Table 5: Experimental results for ablation studies.

Model	KB Completion	Abstractive Product Summarization						Multi-Turn Conversation	
		Home Applications		Clothing		Cases&Bags			
	F1	RG-1	RG-2	RG-1	RG-2	RG-1	RG-2	RG-L	BLEU
K-PLUG	**95.97**	**33.56**	**12.50**	**33.00**	**11.24**	**33.87**	**11.83**	**26.60**	**17.80**
-KMLM	95.88	33.52	12.43	32.87	11.20	33.75	11.70	26.43	17.62
-KMS2S	95.76	33.13	12.14	32.12	10.97	33.74	11.43	25.82	16.97
-PEABD	95.89	33.26	12.30	32.96	11.14	33.69	11.17	26.07	17.58
-PECC	95.59	33.24	12.17	32.25	11.12	33.59	11.18	26.02	17.16
-PEASG	95.48	33.39	12.36	32.57	11.16	33.78	11.45	26.12	17.38

5 Conclusion

We present a knowledge-injected pre-trained model (K-PLUG) that is a powerful domain-specific language model trained on a large-scale e-commerce corpus designed to capture e-commerce knowledge, including e-commerce KB, product aspects, product categories, and USPs. The pre-training framework combines masked language model and masked seq2seq with novel objectives formulated as product aspect boundary detection, product aspect summary generation, and product category classification tasks. Our proposed model demonstrates strong performances on both natural language understanding and generation downstream tasks, including e-commerce KB completion, abstractive product summarization, and multi-turn dialogue.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, (ICLR)*, 2015.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. The jddc corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pp. 459–466, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 4171–4186, 2019.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, (NeurIPS)*, pp. 13042–13054, 2019.

Xin Luna Dong, Xiang He, Andrey Kan, Xian Li, Yan Liang, Jun Ma, Yifan Ethan Xu, Chenwei Zhang, Tong Zhao, Gabriel Blanco Saldana, et al. AutoKnow: Self-driving knowledge collection for products of thousands of types. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 2724–2734, 2020.

Günes Erkan and Dragomir R Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 22:457–479, 2004.

Thomas M Garrett. Reality in advertising. *America*, 1961.

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. SentiLR: Linguistic knowledge enhanced language representation for sentiment analysis. *arXiv preprint arXiv:1911.02493*, 2019.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. Informing unsupervised pretraining with external linguistic knowledge. *arXiv preprint arXiv:1909.02339*, 2019.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4656–4667, 2020.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7871–7880, 2020.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. Aspect-aware multimodal summarization for chinese e-commerce products. In *Proceedings of the Thirty-Forth AAAI Conference on Artificial Intelligence (AAAI)*, pp. 8188–8195, 2020.

Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 150–157, 2003.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pre-training approach. *arXiv preprint arXiv:1907.11692*, 2019.

Robert L Logan IV, Samuel Humeau, and Sameer Singh. Multimodal attribute extraction. *arXiv preprint arXiv:1711.11118*, 2017.

Xusheng Luo, Luxin Liu, Yonghua Yang, Le Bo, Yuanpeng Cao, Jinghang Wu, Qiang Li, Keping Yang, and Kenny Q Zhu. AliCoCo: Alibaba e-commerce cognitive concept net. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pp. 313–327, 2020.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 2227–2237, 2018.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 43–54, 2019.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1073–1083, 2017.

Heung-Yeung Shum, Xiao-dong He, and Di Li. From Eliza to Xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19:10–26, 2018.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 5926–5936, 2019.

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

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4067–4076, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. K-Adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*, 2020.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *arXiv preprint arXiv:1911.06136*, 2019.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 496–505, 2017.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *8th International Conference on Learning Representations (ICLR)*, 2020.

Huimin Xu, Wenting Wang, Xin Mao, Xinyu Jiang, and Man Lan. Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5214–5223, 2019.

Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai Sun, Yechen Xu, et al. CLUE: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*, 2020.

Yu Yan, Weizhen Qi, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. ProphetNet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*, 2020.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNET: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 5753–5763, 2019.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1441–1451, 2019.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pp. 3740–3752, 2018.

Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. Multimodal joint attribute prediction and value extraction for e-commerce product. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.