K-PLUG: Knowledge-injected Pre-trained Language Model for Natural Language Understanding and Generation in E-Commerce

Song Xu\(^1\), Haoran Li\(^1\), Peng Yuan\(^1\), Yujia Wang\(^2\), Youzheng Wu\(^1\), Xiaodong He\(^1\), Ying Liu\(^3\), Bowen Zhou\(^1\)

\(^1\) JD AI Research
\(^2\) University of California, Berkeley
\(^3\) Renmin University of China

\{xusong28, lihaoran24, yuanpeng29\}@jd.com

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. Our code is available at https://github.com/xu-song/k-plug.

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., 2020), PEGASUS (Zhang et al., 2020), and ProphetNet (Qi 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), part-of-speech tag (Ke et al., 2020), and sentiment polarity (Ke et al., 2020; Tian et al., 2020). However, these methods only integrate knowledge into the encoder, and the decoding process in many NLG tasks benefits little from this knowledge.

To mitigate this problem, we propose a Knowledge-injected Pre-trained Language model that is suitable for both Natural Language Understanding and Generation (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) (Reeves, 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 are 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 knowledge in e-commerce scenarios: e-commerce 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., 2020) 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.

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), SentiLARE (Ke et al., 2020), 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 platform¹, 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. We adopt GELU activation (Hendrycks and Gimpel, 2016) as in GPT (Radford et al., 2018). We use Adam optimizer (Kingma and Ba, 2015) with a learning rate of $5\times10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, L2 weight decay of 0.01, learning rate warm-up over the first 10,000 steps and linear decay of the learning rate. The dropout probability is 0.1. The maximum sequence length is set to 512 tokens. Pre-training was performed with 4 Telsa V100 GPUs. The pre-training is done within 10 epochs, which takes around 10 days, and the fine-tuning takes up to 1 day. We use the beam search with a beam size of 5 for inference for the NLG tasks.

¹https://www.jd.com/
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., 2020a). 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 (Reeves, 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) to train the Transformer encoder 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 $x = (x_1, ..., x_M)$ 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 $\tilde{x} = (x_1, ..., [MASK], ..., x_M)$. MLM aims to model the conditional likelihood $P(x_m|\tilde{x})$, and the loss function is:

$$L_{MLM} = \log P(x_m|\tilde{x})$$  \hspace{1cm} (1)

The major difference from BERT is that our KMLM prioritizes knowledge tokens, which contain knowledge regarding product attributes and USPs, 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 $x = (x_1, ..., x_u, ..., x_M)$, a masked sequence $\tilde{x} = (x_1, ..., [MASK], ..., [MASK], ..., x_M)$ is produced by replacing the span $x_{us}$, ranging from $x_u$ to $x_v$, with the [MASK] token. MS2S aims to model $P(x_{us}|\tilde{x})$, which can be further factorized into a product $P(x_{us}|\tilde{x}) = \prod_{t=u}^v P(x_t|\tilde{x})$ according to the chain rule. The loss function is:

$$L_{MS2S} = \sum_{t=u}^v \log P(x_t|\tilde{x})$$  \hspace{1cm} (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., 2020a) 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 $x = (x_1, ..., x_M)$, the encoder of K-PLUG outputs a sequence $h = (h_1, ..., h_M)$, which is fed into a softmax layer, and generates a probability sequence $y$. The loss function is:

$$L_{PEABD} = -\sum_{t} y_t \log y_t$$  \hspace{1cm} (3)
where \( y \in \{[0, 1]\} \) 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 \( x = (x_1, ..., x_M) \), a softmax layer outputs the classification score, \( y \), based on the representation of the encoder classification token, [CLS]. The loss function maximizes the model’s probability of outputting the true product entity category as follows:

\[
L_{PECC} = -\hat{y} \log y
\]

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 naturally exists in our pre-training data. Since the aspect boundary detection task.

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

\[
L_{PEASG} = \sum_t \log P(y_t|x, y_{<t})
\]

Overall, the pre-training loss is the sum of above-mentioned loss functions:

\[
L_{K-PLUG} = L_{MLM} + L_{MS2S} + L_{PEABD} + L_{PECC} + L_{PEASG}
\]

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., 2020a), 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.

**Dataset.** We conduct experiments on the dataset of MEPAVE (Zhu et al., 2020b). 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.

**Model.** 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
<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>79.68</td>
<td>86.43</td>
<td>82.92</td>
</tr>
<tr>
<td>ScalingUp</td>
<td>65.48</td>
<td>93.78</td>
<td>77.12</td>
</tr>
<tr>
<td>BERT</td>
<td>78.27</td>
<td>88.62</td>
<td>83.12</td>
</tr>
<tr>
<td>JAVE</td>
<td>80.27</td>
<td>89.92</td>
<td>84.78</td>
</tr>
<tr>
<td>M-JAVE</td>
<td>83.49</td>
<td>90.94</td>
<td>87.17</td>
</tr>
<tr>
<td>C-PLUG</td>
<td>89.79</td>
<td>96.47</td>
<td>93.02</td>
</tr>
<tr>
<td>E-PLUG</td>
<td>89.91</td>
<td>96.75</td>
<td>93.20</td>
</tr>
<tr>
<td>K-PLUG</td>
<td>93.58</td>
<td>97.92</td>
<td>95.97</td>
</tr>
</tbody>
</table>

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. (2020b).

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** (Zhu et al., 2020b) is a joint attribute and value extraction model based on a pre-trained BERT.
- **M-JAVE** (Zhu et al., 2020b) is a multimodal JAVE model, which additionally utilizes product image information.

**Result.** 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. Similar to the text summarization task (Rush et al., 2015; Nallapati et al., 2016; Li et al., 2018b; Zhang et al., 2018a; Li et al., 2020b; Xu et al., 2020b), the input for this task is a textual product description, and the output is a condensed product summary.

**Dataset.** We perform experiments on the CEP-SUM dataset (Li et al., 2020a), 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. Notice that the task of abstractive product summarization and product entity aspect summary generation (PEASG) are partly different. The abstractive product summarization task aims to generate a complete and cohesive product summary given a detailed product description. Given a product aspect description, PEASG aims to produce an aspect summary that basically consists of condensed USPs. In addition, for abstractive product summarization task, the average length of the product summaries is 79, while the lengths of the product aspect summaries are less than 10 in general.

**Model.** 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 and 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., 2020a) is the state-of-the-art method for abstractive product summarization, taking both textual and visual product information as the input.

**Result.** Table 2 shows the experimental results, including ROUGE-1 (RG-1), ROUGE-2 (RG-2), and ROUGE-L (RG-L) F1 scores (Lin and 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
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. (2020a).

<table>
<thead>
<tr>
<th>Model</th>
<th>Home Applications</th>
<th>Clothing</th>
<th>Cases&amp;Bags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RG-1</td>
<td>RG-2</td>
<td>RG-L</td>
</tr>
<tr>
<td>LexRank</td>
<td>24.06</td>
<td>10.01</td>
<td>18.19</td>
</tr>
<tr>
<td>Seq2seq</td>
<td>21.57</td>
<td>7.18</td>
<td>17.61</td>
</tr>
<tr>
<td>MASS</td>
<td>26.19</td>
<td>8.02</td>
<td>18.73</td>
</tr>
<tr>
<td>Aspect MMPG*</td>
<td>34.36</td>
<td>12.52</td>
<td>22.35</td>
</tr>
<tr>
<td>E-PLUG</td>
<td>33.11</td>
<td>12.07</td>
<td>22.01</td>
</tr>
<tr>
<td>K-PLUG</td>
<td>33.56</td>
<td>12.50</td>
<td>22.15</td>
</tr>
</tbody>
</table>

Table 3: Human evaluation results (%). “Win” denotes that the generated summary of K-PLUG is better than E-PLUG.

<table>
<thead>
<tr>
<th>KB</th>
<th>Aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Win/Lose/Tie</td>
<td>Kappa</td>
</tr>
<tr>
<td>32.67/11.00/56.33</td>
<td>0.515</td>
</tr>
<tr>
<td>Win/Lose/Tie</td>
<td>Kappa</td>
</tr>
<tr>
<td>32.67/12.00/55.33</td>
<td>0.441</td>
</tr>
<tr>
<td>Category USPs</td>
<td>Win/Lose/Tie</td>
</tr>
<tr>
<td>25.33/7.00/67.67</td>
<td>0.612</td>
</tr>
<tr>
<td>Win/Lose/Tie</td>
<td>Kappa</td>
</tr>
<tr>
<td>28.67/9.33/62.00</td>
<td>0.428</td>
</tr>
</tbody>
</table>

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.

Dataset. We conduct experiments on two datasets of JDDC (Chen et al., 2020) and ECD (Zhang et al., 2018b). 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, which are the goals of a dialogue, such as updating addresses, inquiring prices, etc., 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 preprocessing, 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.

Model. 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.
Table 4: 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).

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, CNN, BiLSTM, and BERT) and generative-based approaches. Other baselines are as follows.

- **SMN** (Wu et al., 2017) matches a response with each utterance in the context.
- **DUA** (Zhang et al., 2018b) is a deep utterance aggregation model based on the fine-grained context representations.
- **DAM** (Zhou et al., 2018) matches a response with the context based using dependency information based on self-attention and cross-attention.
- **IoI** (Tao et al., 2019) is a deep matching model by stacking multiple interactions blocks between utterance and response.
- **MSN** (Yuan et al., 2019) selects relevant context and generates better context representations with the selected context.

**Result.** Table 4 and 5 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.

**Human Evaluation.** We further perform a human evaluation on the JDDC dataset. We randomly choose 100 samples from the test set, and three annotators are involved to determine whether K-PLUG outperforms E-PLUG with respect to (1) relevance between the response and the contexts and (2) readability of the response. The results are shown in Table 6. We can see that the percentage of “Win”, which denotes that the results of K-PLUG is better than E-PLUG, is significantly larger than “Lose” (p-value < 0.01 for t-test).

**4.3 Ablation Studies**

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

**Result.** The ablation results are shown in Table 7. 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>
<thead>
<tr>
<th>Model</th>
<th>RG-L</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>19.47</td>
<td>9.94</td>
</tr>
<tr>
<td>BERT</td>
<td>19.90</td>
<td>10.27</td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>22.17</td>
<td>14.15</td>
</tr>
<tr>
<td>PG</td>
<td>23.62</td>
<td>14.27</td>
</tr>
<tr>
<td>C-PLUG</td>
<td>25.47</td>
<td>16.75</td>
</tr>
<tr>
<td>E-PLUG</td>
<td>25.93</td>
<td>17.12</td>
</tr>
<tr>
<td>K-PLUG</td>
<td><strong>26.60</strong></td>
<td><strong>17.80</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>R10@1</th>
<th>R10@2</th>
<th>R10@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>32.8</td>
<td>51.5</td>
<td>79.2</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>35.5</td>
<td>52.5</td>
<td>82.5</td>
</tr>
<tr>
<td>SMN</td>
<td>45.3</td>
<td>65.4</td>
<td>88.6</td>
</tr>
<tr>
<td>DUA</td>
<td>50.1</td>
<td>70.0</td>
<td>92.1</td>
</tr>
<tr>
<td>DAM</td>
<td>52.6</td>
<td>72.7</td>
<td>93.3</td>
</tr>
<tr>
<td>IoI-local</td>
<td>56.3</td>
<td>76.8</td>
<td>95.0</td>
</tr>
<tr>
<td>MSN</td>
<td>60.6</td>
<td>77.0</td>
<td>93.7</td>
</tr>
<tr>
<td>BERT</td>
<td>54.3</td>
<td>73.4</td>
<td>94.3</td>
</tr>
<tr>
<td>C-PLUG</td>
<td>62.7</td>
<td>76.8</td>
<td>95.0</td>
</tr>
<tr>
<td>E-PLUG</td>
<td>65.8</td>
<td>80.1</td>
<td>95.6</td>
</tr>
<tr>
<td>K-PLUG</td>
<td><strong>73.5</strong></td>
<td><strong>82.9</strong></td>
<td><strong>96.4</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Readability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Win/Lose/Tie</td>
<td>Kappa</td>
</tr>
<tr>
<td>29.00/21.00/50.00</td>
<td>0.428</td>
</tr>
</tbody>
</table>

Table 5: Experimental results for the multi-turn conversation task on the ECD dataset. The results in the first block are taken from Zhang et al. (2018b).

Table 6: Human evaluation results (%). “Win” denotes that the generated response of K-PLUG is better than E-PLUG.
### Table 7: Experimental results for ablation studies.

<table>
<thead>
<tr>
<th>Model</th>
<th>KB Completion F1</th>
<th>Home Applications RG-1</th>
<th>Clothing RG-1</th>
<th>Cases&amp;Bags RG-1</th>
<th>Cases&amp;Bags RG-2</th>
<th>Multi-Turn Conversation RG-L BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-PLUG</td>
<td>95.97</td>
<td>33.56</td>
<td>12.50</td>
<td>33.00</td>
<td>11.24</td>
<td>33.87</td>
</tr>
<tr>
<td>-KMLM</td>
<td>95.88</td>
<td>33.52</td>
<td>12.43</td>
<td>32.87</td>
<td>11.20</td>
<td>33.75</td>
</tr>
<tr>
<td>-KMS2S</td>
<td>95.76</td>
<td>33.13</td>
<td>12.14</td>
<td>32.12</td>
<td>10.97</td>
<td>33.74</td>
</tr>
<tr>
<td>-PEABD</td>
<td>95.89</td>
<td>33.26</td>
<td>12.30</td>
<td>32.96</td>
<td>11.14</td>
<td>33.69</td>
</tr>
<tr>
<td>-PECC</td>
<td>95.59</td>
<td>33.24</td>
<td>12.17</td>
<td>32.25</td>
<td>11.12</td>
<td>33.59</td>
</tr>
<tr>
<td>-PEASG</td>
<td>95.48</td>
<td>33.39</td>
<td>12.36</td>
<td>32.57</td>
<td>11.16</td>
<td>33.78</td>
</tr>
</tbody>
</table>

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.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600.

References


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. 2020. 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), pages 2724–2734.


Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qiangqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuyou Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qinpeng Zhao, Cong Yue, Xinrui Zhang, Zhengliao Yang, Kyle Richardson, and Zhenzhong Lan. 2020a. CLUE: A Chinese language understanding evaluation benchmark. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4762–4772.


Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu.


A Appendix

A.1 Case studies

We present some examples from the test set of each task, with comparisons of the ground-truth result and the outputs produced by the models of E-PLUG and K-PLUG.
Table 8: Case study for the e-commerce KB complete task. The product attribute and the corresponding attribute value is presented as [attribute value]product attribute. The K-PLUG model accurately complete the e-commerce KB, while the E-PLUG model sometimes fails. The translation texts are given below the original examples.
**Input Description**

Steel tube suction, copper coil design. Stainless steel. Quality you can see, multi-faceted, 310pa large suction, 17 m³ h⁻¹ suction, pressurized copper, large volume of fumes. Leaves no cooking smells. 

(Sacon’s European style kitchen range hood. Extra powerful ventilation. Intelligent touch control. Stainless steel. Superior quality with extra reliability. 310 Pa air pressure. 17 m³ h⁻¹ suction. Protection against head bumping. Limited noise of 53 dB. Highly powered. Captures large volume of fumes. Leaves no cooking smells.)

**Ground-truth**

Want to change your kitchen environment? Then try this **large suction** kitchen hood! **310pa** air pressure & **17 m³ h⁻¹** suction + large suction makes it easy to use.

(This kitchen range hood with extra suction and ventilation is ideal for a clean and fresh kitchen. Perfectly combining 310pa of air pressure and 17 m³ h⁻¹ of suction power, it filters out grease before you realize it. Never have to worry again about keeping your loved ones from cooking fumes.)

**E-PLUG**

This 4-star kitchen range hood, with high-quality stainless steel materials, provides smoke-free ventilation. The rust-free copper tank ensures quick boiling with stability. It features extra suction and ventilation for immediate filtering of fumes and grease. Its intelligent touch control panel makes it extremely easy to use.

**K-PLUG**

This 4-star kitchen range hood, with high-quality stainless steel materials, provides smoke-free ventilation. The rust-free copper tank ensures quick boiling with stability. It features extra suction and ventilation for immediate filtering of fumes and grease. Its intelligent touch control panel makes it extremely easy to use.

<table>
<thead>
<tr>
<th>Input Description</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>康宝8升四季控温型，经济适用，家用厨房卫生间淋浴洗澡器排式燃气热水器，厨房洗菜，冬天洗澡再也不会冻手，多达36段的配置。高浓度无氧铜水箱。LED数字大屏，封闭式燃烧系统。防患于未然，控温清晰可见，多重安全防护。无氧铜水箱，让每次沐浴都是全新享受。</td>
<td></td>
</tr>
<tr>
<td>(Canbo’s affordable 8L thermal control natural gas residential kitchen and bathroom water heater. Provides warm water for doing dishes and cleaning up in winter. High-quality accessories. High concentration oxygen-free copper tank. LED digital screen. Closed boiling system for safety. Easy temperature control. Multiple security checks. Oxygen-free copper tank. Let’s enjoy your showers.)</td>
<td></td>
</tr>
</tbody>
</table>

**Ground-truth**

Using high-purity oxygen-free copper tanks, it increases heat efficiency and thermal stability. Closed boiling technique, increased heat efficiency, and high-definition LED digital screen. Let you easily adjust the water temperature.

(Closed boiling technique increases efficiency and saves energy. Thermal control and LED digital screen for adjustable water temperature anytime.)

**E-PLUG**

This gas heater has a large volume and provides warm water year round. It adopts the closed boiling technique for faster warming up and shorter wait time in shower. It works in low pressure condition and is suitable for older neighborhoods.

**K-PLUG**

High-speed water heater, LED digital screen, high concentration oxygen-free copper tank. Equipped with LED digital screen for clear thermal control and temperature adjustment. Closed boiling technique speeds up warming and increases efficiency. High concentration oxygen-free copper tank for faster warming up.

**Table 9:** Case study for the abstractive product summarization task (*Home Applications* category). The K-PLUG model generates summaries describing more information about *e-commerce knowledge bases* and *unique selling propositions of product entities*.
input description

劲霸男士t恤，夏季新品，撞色时尚，舒适，花卉印花，短袖，黑色，商务休闲，圆领。夏季，时尚都市，短袖t恤，显垂感和耐穿性好，时尚圆领，简约大气，多色可选，莫代尔棉针织面料，柔滑触感，清凉舒爽，花卉印花结合，珠片绣花，时尚大方...

reference

选用莫代尔棉针织面料，手感柔滑，上身清凉舒爽，显垂感好，耐穿性强，胸前撞色花卉印花结合珠片绣花，时尚大方。经典圆领设计，自然突显型男风范。
(Made from modal knit fabric. Incredibly soft and elastic. Features a slightly looser fit. Natural draping with high durability. Decorated with floral print and glitter embroidery for a stylish and casual look. Finished with a classic crewneck, appealing to the modern sense of fashion)

E-PLUG

圆领的领口更好的将颈部曲线勾勒而出。水洗磨白的颜色平添了几分帅气与随性。让你和夏季的风一起随行。让休闲舒适的田园感油然而生。
(The crewneck design helps in elongating the neck. The light wash color gives an attractive yet effortless look. It dresses you up in a country aesthetic style, comfy and relaxed just like the summer breeze.)

K-PLUG

采用莫代尔棉针织面料，手感柔滑，清凉舒爽，花卉印花结合珠片绣花，时尚大方，简约圆领设计，轻松修饰颈部线条。
(Made from modal knit fabric. Incredibly soft and stretchy. Slightly looser fit. Flower print and glitter embroidery. Stylish and casual. The basic crewneck design easily helps in elongating the neck.)

input description

吉普，羽绒服，男。中长款，90绒，冬季外套，新品，连帽，加绒，加厚。保暖羽绒外套，黑色，白鸭绒，时尚都市，型男都这么穿。融合艺术细节，创造76年传奇，潮流趋势必备。温暖新升级...

reference

采用聚酯纤维面料，手感柔软，轻盈且透气性较好，穿在身上干爽舒适。内部以白鸭绒进行填充，充绒量较高，柔软蓬松，更有加厚修身的版型设计，保暖效果较好。为您抵御户外严寒。
(Made from polyester. Feel soft, lightweight, and breathable. Keeps you dry and comfortable. Filled primarily with white duck down, fluffy and light. Features a thick yet slim-fit design. Keeps you warm in cold climates.)

E-PLUG

这款羽绒服采用中长款的版型设计，修饰你的身材线条，而且还不乏优雅稳重气质。连帽的，加持增添青春学院风气息，衣上字母印花的点缀，俏皮又减龄。
(This down jacket features a mid-thigh length, keeping a stylish silhouette and giving you an elegant and mature look. The hood and letter print on the jacket make you look younger.)

K-PLUG

采用聚酯纤维面料制成，手感柔软，亲肤透气，内部以白鸭绒填充，蓬松度高，轻盈保暖，更有连帽设计，可以锁住人体的热量，为您抵御户外寒冷天气，带来舒适的穿着体验。
(Made from polyester. Soft and breathable. Filled primarily with white duck down, fluffy and light. Lightweight but warm. Features a hooded design. Locks in the heat and keeps out the cold. Comfortable to wear.)

Table 10: Case study for the abstractive product summarization task (Clothing category). The K-PLUG model generates summaries describing more information about e-commerce knowledge bases and unique selling propositions of product entities.
Table 11: Case study for the abstractive product summarization task (Bags & Cases category). The K-PLUG model generates summaries describing more information about e-commerce knowledge bases and unique selling propositions of product entities.
<table>
<thead>
<tr>
<th>Q1</th>
<th>Is it paper-peel walnut?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Yes, it is.</td>
</tr>
<tr>
<td>Q2</td>
<td>Is it of premium quality?</td>
</tr>
<tr>
<td>A2</td>
<td>What do you mean?</td>
</tr>
<tr>
<td>Q3</td>
<td>Aren’t these walnuts sized differently? I’m getting 5 kilos. Are you still here?</td>
</tr>
<tr>
<td>E-PLUG</td>
<td>This is paper-peel walnut, which you can peel easily with bare hands.</td>
</tr>
<tr>
<td>K-PLUG</td>
<td>They mostly have a diameter between 3-5 cm.</td>
</tr>
</tbody>
</table>

Table 12: Case study for the multi-turn dialogue task on the ECD dataset. The K-PLUG model produces more accurate responses for the questions related to e-commerce knowledge bases.