Causality-aware Concept Extraction based on Knowledge-guided Prompting

Siyu Yuan♡, **Deqing Yang**♡*, **Jinxi Liu**♡, **Shuyu Tian**♡, **Jiaqing Liang**♡, **Yanghua Xiao**♠♣*, **Rui Xie**♢
♡School of Data Science, Fudan University, Shanghai, China
♠Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
♣Fudan-Aishu Cognitive Intelligence Joint Research Center, ♢Meituan, Beijing, China
{syyuan21,jxliu22, sytian21}@m.fudan.edu.cn,
{yangdeqing,liangjiaqing,shawyh}@fudan.edu.cn, rui.xie@meituan.com

Abstract

Concepts benefit natural language understanding but are far from complete in existing knowledge graphs (KGs). Recently, pre-trained language models (PLMs) have been widely used in text-based concept extraction (CE). However, PLMs tend to mine the co-occurrence associations from massive corpus as pre-trained knowledge rather than the real causal effect between tokens. As a result, the pre-trained knowledge confounds PLMs to extract biased concepts based on spurious co-occurrence correlations, inevitably resulting in low precision. In this paper, through the lens of a Structural Causal Model (SCM), we propose equipping the PLM-based extractor with a knowledge-guided prompt as an intervention to alleviate concept bias. The prompt adopts the topic of the given entity from the existing knowledge in KGs to mitigate the spurious co-occurrence correlations between entities and biased concepts. Our extensive experiments on representative multilingual KG datasets justify that our proposed prompt can effectively alleviate concept bias and improve the performance of PLM-based CE models. The code has been released on https://github.com/siyuyuan/KPCE.

1 Introduction

The concepts in knowledge graphs (KGs) enable machines to understand natural languages better, and thus benefit many downstream tasks, such as question answering (Han et al., 2020), commonsense reasoning (Zhong et al., 2021) and entity typing (Yuan et al., 2022). However, the concepts, especially the fine-grained ones, in existing KGs still need to be completed. For example, in the widely used Chinese KG *CN-DBpedia* (Xu et al., 2017), there are nearly 17 million entities but only 0.27 million concepts in total, and more than 20% entities even have no concepts. Although *Probase* (Wu et al., 2012) is a large-scale English KG, the fine-grained concepts with two or more modifiers in it only account for 30% (Li et al., 2021). We focus on extracting multi-grained concepts from texts to complete existing KGs.

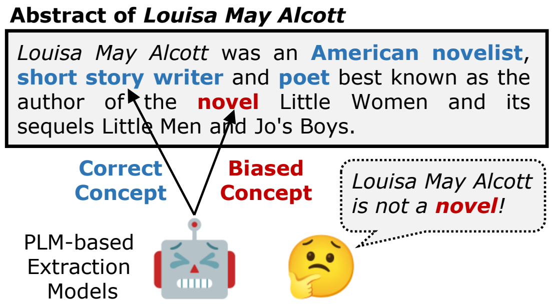

Figure 1: The example of concept bias. The PLM-based CE models are biased to extract *novel* mistakenly as the concept of *Louisa May Alcott* from the text.

Most of the existing text-based concept acquisition approaches adopt the extraction scheme, which can be divided into two categories: *1)* pattern-matching approaches (Auer et al., 2007; Wu et al., 2012; Xu et al., 2017), which can obtain high-quality concepts but only have low recall due to poor generalization; *2)* learning-based approaches (Luo et al., 2020; Ji et al., 2020; Yuan et al., 2021a), which employ pre-trained language models (PLMs) fine-tuned with labeled data to extract concepts.

However, an unignorable drawback of these learning-based approaches based on PLMs is **concept bias**. Concept bias means the concepts are extracted based on their contextual (co-occurrence) associations rather than the real causal effect between the entities and concepts, resulting in low extraction precision. For example, in Figure 1, PLMs tend to extract *novel* and *writer* together as concepts for the entity *Louisa May Alcott* even if we explicitly input the entity *Louisa May Alcott* to the model. Previous work demonstrates that causal inference is a promising technique for bias analy-

*Corresponding authors.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 9255–9272

sis (Lu et al., 2022). To analyze the reasons behind concept bias, we devise a Structural Causal Model (SCM) (Pearl, 2009) to investigate the causal effect in the PLM-based concept extraction (CE) system, and show that pre-trained knowledge in PLMs confounds PLMs to extract biased concepts. During the pre-training, the entities and biased concepts (*e.g.*, *Louisa May Alcott* and *novel*) often co-occur in many texts. Thus, PLMs tend to mine statistical associations from a massive corpus rather than the real causal effect between them (Li et al., 2022), which induces spurious co-occurrence correlations between entities (*i.e.*, *Louisa May Alcott*) and biased concepts (*i.e.*, *novel*). Since we cannot directly observe the prior distribution of pre-trained knowledge, the backdoor adjustment is intractable for our problem (Pearl, 2009). Alternatively, the frontdoor adjustment (Peters et al., 2017) can apply a mediator as an intervention to mitigate bias.

In this paper, we adopt language prompting (Gao et al., 2021; Li and Liang, 2021) as a mediator for the frontdoor adjustment to handle concept bias. We propose a novel **C**oncept **E**xtraction framework with **K**nowledge-guided **P**rompt, namely **KPCE** to extract concepts for given entities from text. Specifically, we construct a knowledge-guided prompt by obtaining the topic of the given entity (*e.g.*, *person* for *Louisa May Alcott*) from the knowledge in the existing KGs. Our proposed knowledge-guided prompt is independent of pre-trained knowledge and fulfills the frontdoor criterion. Thus, it can be used as a mediator to guide PLMs to focus on the right cause and alleviate spurious correlations. Although adopting our knowledge-guided prompt to construct the mediator is straightforward, it has been proven effective in addressing concept bias and improving the extraction performance of PLM-based extractors in the CE task.

In summary, our contributions include: *1)* To the best of our knowledge, we are the first to identify the concept bias problem in the PLM-based CE system. *2)* We define a Structural Causal Model to analyze the concept bias from a causal perspective and propose adopting a knowledge-guided prompt as a mediator to alleviate the bias via frontdoor adjustment. *3)* Experimental results demonstrate the effectiveness of the proposed knowledge-guided prompt, which significantly mitigates the bias and achieves a new state-of-the-art for CE task.

2 Related Work

Concept Acquisition Most of the existing text-based concept acquisition approaches adopt the extraction scheme, which can be divided into two categories: *1) Pattern-matching Approaches*: extract concepts from free texts with hand-crafted patterns (Auer et al., 2007; Wu et al., 2012; Xu et al., 2017). Although they can obtain high-quality concepts, they have low recall due to their poor generalization ability; *2) Learning-based Approaches*: mostly employ the PLM-based extraction models from other extraction tasks, such as the Named Entity Recognition (NER) models (Li et al., 2020; Luo et al., 2021; Lange et al., 2022) and Information Extraction models (Fang et al., 2021; Yuan et al., 2021a) in the CE task. Although they can extract many concepts from a large corpus, the concept bias cannot be well handled.

Causality for Language Processing Several recent work studies causal inference combined with language models for natural language processing (NLP) (Schölkopf, 2022), such as controllable text generation (Hu and Li, 2021; Goyal et al., 2022) and counterfactual reasoning (Chen et al., 2022; Paranjape et al., 2022). In addition, causal inference can recognize spurious correlations via Structural Causal Model (SCM) (Pearl, 2009) for bias analysis and eliminate biases using causal intervention techniques (Weber et al., 2020; Lu et al., 2022). Therefore, there are also studies showing that causal inference is a promising technique to identify undesirable biases in the NLP dataset (Feder et al., 2022) pre-trained language models (PLMs) (Li et al., 2022). In this paper, we adopt causal inference to identify, understand, and alleviate concept bias in concept extraction.

Language Prompting Language prompting can distill knowledge from PLMs to improve the model performance in the downstream task. Language prompt construction methods can be divided into two categories (Liu et al., 2021a): *1) Hand-crafted Prompts*, which are created manually based on human insights into the tasks (Brown et al., 2020; Schick and Schütze, 2021; Schick and Schutze, 2021). Although they obtain high-quality results, how to construct optimal prompts for a certain downstream task is an intractable challenge; *2) Automated Constructed Prompts*, which are generated automatically from natural language phrases (Jiang et al., 2020; Yuan et al., 2021b) or vector space (Li

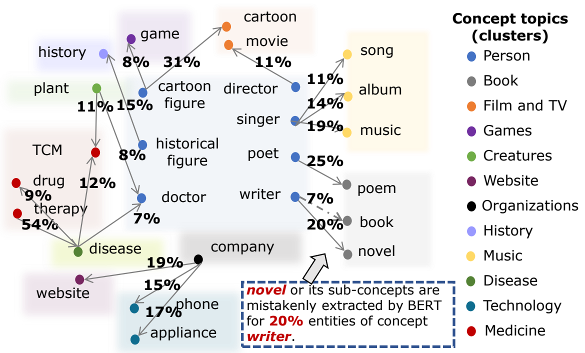

Figure 2: Concept bias map for the entities of popular concepts in CN-DBpedia (better viewed in color).

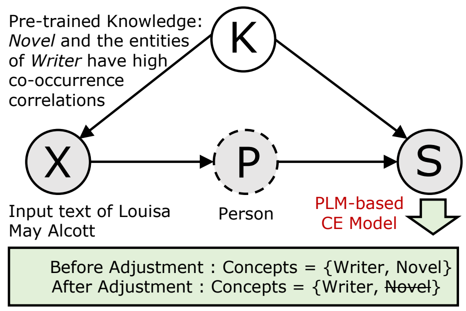

Figure 3: The proposed structural causal model (SCM). A hollow circle indicates the variable is latent, and a shaded circle indicates the variable is observed. Without causal intervention, the PLM-based CE model extract *Novel* due to the spurious correlation between the entities of *Writer* and *Novel* caused by the confounding variable K. The constructed mediating variable P can block the backdoor paths for $X \rightarrow S$ (opened by K) and help the model only extract the unbiased concept *Writer*.

and Liang, 2021; Liu et al., 2021b). Although previous work analyzes the prompt from a causal perspective (Cao et al., 2022), relatively little attention has been paid to adopting the prompt to alleviate the bias in the downstream task.

3 Concept Bias Analysis

In this section, we first formally define our task. Then we investigate the concept bias issued by PLMs in empirical studies. Finally, we devise a Structural Causal Model (SCM) to analyze the bias and alleviate it via causal inference.

3.1 Preliminary

Task Definition Our CE task addressed in this paper can be formulated as follows. Given an entity $E = \{e_1, e_2, \cdots, e_{|E|}\}$ and its relevant text $T = \{t_1, t_2, \cdots, t_{|T|}\}$ where e_i (or t_i) is a word token, our framework aims to extract one or multiple spans from T as the concept(s) of E.

Data Selection It must guarantee that the given text contains concepts. The abstract text of an entity expresses the concepts of the entity explicitly, which can be obtained from online encyclopedias or knowledge bases. In this paper, we take the abstract text of an entity as its relevant text T. The details of dataset construction will be introduced in § 5.1. Since we aim to extract concepts from T for E, it is reasonable to concatenate E and T to form the input text $X = \{E, T\}$.

3.2 Empirical Studies on Concept Bias

To demonstrate the presence of concept bias, we conduct empirical studies on the CN-DBpedia dataset (Xu et al., 2017). First, we randomly sample 1 million entities with their concepts from CN-DBpedia, and select the top 100 concepts with the most entities as the *typical concept* set. Then we randomly select 100 entities with their abstracts for each typical concept to construct the input texts and run a BERT-based extractor to extract concepts. Details of the extraction process will be introduced in § 4.2. We invite volunteers to assess whether the extracted concepts are biased. To quantify the degree of concept bias, we calculate the *bias rate* of concept A to another concept B. The bias rate is defined as the number of entities of A for which B or the sub-concepts of B are mistakenly extracted by the extractor, divided by the total number of entities of A.

The bias rates among 26 typical concepts are shown in Figure 2, where the concepts (dots) of the same topic are clustered in one rectangle. The construction of concept topics will be introduced in § 4.1. From the figure, we can conclude that concept bias is widespread in the PLM-based CE system and negatively affects the quality of the results. Previous studies have proven that causal inference can analyze bias via SCM and eliminate bias with causal intervention techniques (Cao et al., 2022). Next, we will analyze concept bias from a causal perspective.

3.3 The Causal Framework for Concept Bias Analysis

The Structural Causal Model We devise a Structural Causal Model (SCM) to identify the causal effect between the input text X of a given entity E and the concept span S that can be extracted from X. As shown in Figure 3, our CE task aims to extract one or multiple spans S from X as the concept(s) of E where the causal effect can be denoted as $X \rightarrow S$.

During the pre-training, the contextual embedding of one token depends on the ones that frequently appear nearby in the corpus. We extrapolate that the high co-occurrence between the entities of true concepts (*e.g.*, *writer*) and biased concepts (*e.g.*, *novel*) in the pre-trained knowledge induces spurious correlations between entities (*e.g.*, *Louisa May Alcott*) and biased concepts (*e.g.*, *novel*). Therefore, the PLM-based CE models can mistakenly extract biased concepts even if the entity is explicitly mentioned in X. The experiments in § 5.4 also prove our rationale. Based on the foregoing analysis, we define the pre-trained knowledge K from PLM-based extraction models as a confounder.

We cannot directly observe the latent space of the PLMs, and thus the backdoor adjustment (Pearl, 2009) is not applicable in our case. Alternatively, we adopt the frontdoor adjustment (Peters et al., 2017) and design a mediator to mitigate the concept bias.

Causal Intervention To mitigate the concept bias, we construct a prompt P as a mediator for $X \rightarrow S$, and then the frontdoor adjustment can apply do-operation.

Specifically, to make the PLMs attend to the right cause and alleviate spurious co-occurrence correlation (*e.g.*, *novel* and *Louisa May Alcott*), we assign a topic as a knowledge-guided prompt P (*i.e.*, *person*) to the input text X (The detailed operation is elaborated in § 4.1). The topics obtained from KGs are independent of pre-trained knowledge, and thus P fulfills the frontdoor criterion.

For the causal effect $X \rightarrow P$, we can observe that $X \rightarrow P \rightarrow S \leftarrow K$ is a collider that blocks the association between P and K, and no backdoor path is available for $X \rightarrow P$. Therefore, we can directly rely on the conditional probability after applying the do-operator for X:

$$P(P = p|do(X = x)) = P(P = p|X = x). \quad (1)$$

Next, for the causal effect $P \rightarrow S$, $P \leftarrow X \leftarrow K \rightarrow S$ is a backdoor path from P to S, which we need to cut off. Since K is an unobserved variable, we can block the backdoor path through X:

$$P(S|do(P)) = \sum_x P(S|P, X = x)P(X = x). \quad (2)$$

Therefore, the underlying causal mechanism of our CE task is a combination of Eq.1 and Eq.2, which can be formulated as:

$$\begin{aligned} & P(S|do(X)) \\ &= \sum_p P(S|p, do(X))P(p|do(X)) \\ &= \sum_p P(S|do(P), do(X))P(p|do(X)) \\ &= \sum_p P(S|do(P))P(p|do(X)). \end{aligned} \quad (3)$$

The theoretical details of the frontdoor adjustment are introduced in Appendix A.

We make the assumption of strong ignorability, *i.e.*, there is only one confounder K between X and S. One assumption of the frontdoor criterion is that the only way the input text X influences S is through the mediator P. Thus, $X \rightarrow P \rightarrow S$ must be the only path. Otherwise, the front-door adjustment cannot stand. Notice that K already represents all the knowledge from pre-trained data in PLMs. Therefore, it is reasonable to use the strong ignorability assumption that it already includes all possible confounders.

Through the frontdoor adjustment, we can block the backdoor path from input text to concepts and alleviate spurious correlation caused by the confounder, *i.e.*, pre-trained knowledge. In practice, we can train a topic classifier to estimate Eq.1 (§ 4.1) and train a concept extractor on our training data to estimate Eq.2 (§ 4.2). Next, we will introduce the implementation of the frontdoor adjustment in detail.

4 Methodology

In this section, we present our CE framework KPCE and discuss how to perform prompting to alleviate concept bias. The overall framework of KPCE is illustrated in Figure 4, which consists of two major modules: *1) Prompt Constructor*: assigns the topic obtained from KGs for entities as a knowledge-guided prompt to estimate Eq.1; *2) Concept Extractor*: trains a BERT-based extractor with the constructed prompt to estimate Eq.2 and extract multi-grained concepts from the input text. Next, we will introduce the two modules of KPCE.

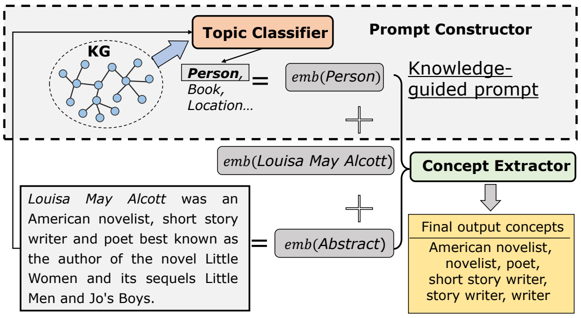

Figure 4: The overview of our CE framework.

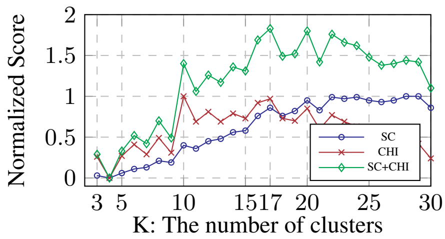

Figure 5: The scores of Silhouette Coefficient (SC) and Calinski Harabaz Index (CHI) under different cluster numbers. The scores are normalized with feature scaling for a fair comparison.

4.1 Prompt Constructor

Knowledge-guided Prompt Construction To reduce the concept bias, we use the topic of a given entity as a knowledge-guided prompt, which is identified based on the external knowledge of the existing KGs. Take *CN-DBpedia* (Xu et al., 2017) as an example [1]. We randomly sample one million entities from this KG and obtain their existing concepts. Then, we select the top 100 concepts having the most entities to constitute the *typical concept* set, which can cover more than 99.80% entities in the KG. Next, we use spectral clustering (Von Luxburg, 2007) with the adaptive K-means (Bhatia et al., 2004) algorithm to cluster these typical concepts into several groups, each of which corresponds to a topic. To achieve the spectral clustering, we use the following overlap coefficient (Vijaymeena and Kavitha, 2016) to measure the similarity between two concepts,

$$Overlap(c_1, c_2) = \frac{|ent(c_1) \cap ent(c_2)|}{min(|ent(c_1)|, |ent(c_2)|) + \delta} \tag{4}$$

where $ent(c_1)$ and $ent(c_2)$ are the entity sets of concept c_1 and concept c_2, respectively. We then construct a similarity matrix of typical concepts to achieve spectral clustering. To determine the best number of clusters, we calculate the Silhouette Coefficient (SC) (Aranganayagi and Thangavel, 2007) and the Calinski Harabaz Index (CHI) (Maulik and Bandyopadhyay, 2002) from 3 to 30 clusters. The scores are shown in Figure 5, from which we find that the best number of clusters is 17. As a result, we cluster the typical concepts into 17 groups and define a topic name for each group. The 17 typical topics and their corresponding concepts are listed in Appendix B.1

[1]In fact, the concepts of CN-DBpedia are inherited from Probase, so the typical topics are the same for CN-DBpedia and Probase.

Identifying Topic Prompt for Each Entity We adopt a topic classifier to assign the topic prompt to the input text X, which is one of the 17 typical topics in Table 6. To construct the training data, we randomly fetch 40,000 entities together with their abstract texts and existing concepts in the KG. According to the concept clustering results, we can assign each topic to the entities. We adopt transformer encoder (Vaswani et al., 2017) followed by a two-layer perception (MLP) (Gardner and Dorling, 1998) activated by ReLU, as our topic classifier [2]. We train the topic classifier to predict the topic prompt $P = \{p_1, p_2, \cdots, p_{|P|}\}$ for X, which is calculated as [3]:

$$P = \arg\max_i \big(P(P^i|X)\big), 1 \le i \le 17, \tag{5}$$

where P^i is the i-th topic among the 17 typical topics.

In our experiments, the topic classifier achieves more than 97.8% accuracy in 500 samples by human assessment. Through training the topic classifier, we can estimate Eq.1 to identify the causal effect $X \rightarrow P$.

4.2 Concept Extractor

Prompt-based BERT The concept extractor is a BERT equipped with our proposed prompt followed by a pointer network (Vinyals et al., 2015). The pointer network is adopted for extracting multi-grained concepts.

We first concatenate the token sequence with the tokens of P and X to constitute the input, *i.e.*, $\{$`[CLS]`P`[SEP]`X`[SEP]`$\}$, where `[CLS]` and `[SEP]` are the special tokens in BERT. With multi-headed self-attention operations over the above in-

[2]We do not employ the PLM-based topic classifier since it will bring a direct path from K to P in Figure 3.

[3]The detailed training operation of topic classifier can be found in Appendix B.1

put, the BERT outputs the final hidden state (matrix), *i.e.*, $\mathbf{H}^{N_L} \in \mathbb{R}^{(|P|+|X|+3)\times d'}$ where d' is the vector dimension and N_L is the total number of layers. Then the pointer network predicts the probability of a token being the start position and the end position of the extracted span. We use $\mathbf{p}^{start}, \mathbf{p}^{end} \in \mathbb{R}^{|P|+|X|+3}$ to denote the vectors storing the probabilities of all tokens to be the start position and end position, which are calculated as

$$[\mathbf{p}^{start}; \mathbf{p}^{end}] = \text{softmax}(\mathbf{H}^{N_L}\mathbf{W} + \mathbf{B}) \quad (6)$$

where $\mathbf{B} \in \mathbb{R}^{(|P|+|X|+3)\times 2}$ and $\mathbf{W} \in \mathbb{R}^{d'\times 2}$ and are both trainable parameters. We only consider the probabilities of the tokens in the abstract T. Given a span with x_i and x_j as the start token and the end token, its confidence score $cs_{ij} \in \mathbb{R}$ can be calculated as

$$cs_{ij} = p_i^{start} + p_j^{end}. \quad (7)$$

Accordingly, the model outputs a ranking list of candidate concepts (spans) with their confidence scores. We only reserve the concepts with confidence scores bigger than the selection threshold. An example to illustrate how to perform the pointer network is provided in Appendix B.2.

During training, the concept extractor is fed with the input texts with topic prompts and outputs the probability (confidence scores) of the spans, and thus can estimate the causal effect $P \to S$ in Eq.2.

Model Training We adopt the cross-entropy function $CE(\cdot)$ as the loss function of our model. Specifically, suppose that $\mathbf{y}_{start} \in \mathbb{N}^{|P|+|X|+3}$ (or $\mathbf{y}_{end} \in \mathbb{N}^{|P|+|X|+3}$) contains the real label (0 or 1) of each input token being the start (or end) position of a concept. Then, we have the following two training losses for the predictions:

$$\mathcal{L}_{start} = CE(\mathbf{p}^{start}, \mathbf{y}_{start}), \quad (8)$$

$$\mathcal{L}_{end} = CE(\mathbf{p}^{end}, \mathbf{y}_{end}). \quad (9)$$

Then, the overall training loss is

$$\mathcal{L} = \alpha\mathcal{L}_{start} + (1-\alpha)\mathcal{L}_{end} \quad (10)$$

where $\alpha \in (0, 1)$ is the control parameter. We use Adam (Kingma and Ba, 2015) to optimize $\mathcal{L}$.

5 Experiments

5.1 Datasets

CN-DBpedia From the latest version of Chinese KG CN-DBpedia (Xu et al., 2017) and Wikipedia, we randomly sample 100,000 instances to construct our sample pool. Each instance in the sample pool consists of an entity with its concept and abstract text [4]. Then, we sample 500 instances from the pool as our test set and divide the rest of the instances into the training set and validation set according to 9:1.

Probase We obtain the English sample pool of 50,000 instances from Probase (Wu et al., 2012) and Wikipedia. The training, validation and test set construction are the same as the Chinese dataset.

5.2 Evaluation Metrics

We compare KPCE with seven baselines, including a pattern-matching approach *i.e.*, Hearst pattern. Detailed information on baselines and some experiment settings is shown in Appendix C.1 and C.2. Some extracted concepts do not exist in the KG, and cannot be assessed automatically. Therefore, we invite the annotators to assess whether the extracted concepts are correct. The annotation detail is shown in Appendix C.3.

Please note that the extracted concepts may already have existed in the KG for the given entity, which we denote as *EC*s (existing concepts). However, our work expects to extract correct but new concepts (that do not exist in the KG) to complete the KGs, which we denote as *NC*s (new concepts). Therefore, we record the number of new concepts (NC #) and display the ratio of correct concepts (ECs and NCs) as precision (Prec.). Since it is difficult to know all the correct concepts in the input text, we report the relative recall (Recall$_R$). Specifically, suppose NCs # is the total number of new concepts extracted by all models. Then, the relative recall is calculated as NC # divided by NCs # [5]. Accordingly, the relative F1 (F1$_R$) can be calculated with Prec. and Recall$_R$. In addition, we also record the average length of new concepts (Len$_{NC}$) to investigate the effectiveness of the pointer network.

5.3 Overall Performance

We present the main results in Table 1. Generally, we have the following findings:

Our method outperforms previous baselines by large margins, including previous state-of-the-art (MRC-CE, Yuan et al., 2021a). However, the

[4] If one entity has multiple concepts, we randomly select one as the golden label.

[5] Please note that NCs # is counted based on all models in one comparison. Therefore, Recall$_R$ can be different for one model when the compared models change.

pattern-based approach still beats the learning-based ones in precision, envisioning a room for improvement. We find that KPCE achieves a more significant improvement in extracting new concepts, indicating that KPCE can be applied to achieve KG completion (§ 5.5). We also compare KPCE with its ablated variant and the results show that adding a knowledge-guided prompt can guide BERT to achieve accurate CE results.

We notice that almost all models have higher extraction precision on the Chinese dataset than that on the English dataset. This is because the modifiers are usually placed before nouns in Chinese syntactic structure, and thus it is easier to identify these modifiers and extract them with the coarse-grained concepts together to form the fine-grained ones. However, for the English dataset, not only adjectives but also subordinate clauses modify coarse-grained concepts, and thus identifying these modifiers is more difficult.

Compared with learning-based baselines, KPCE can extract more fine-grained concepts. Although the Hearst pattern can also extract fine-grained concepts, it cannot simultaneously extract multi-grained concepts when a coarse-grained concept term is the subsequence of another fine-grained concept term. For example, in Figure 4, if Hearst Pattern extracts *American novelist* as a concept, it cannot extract *novelist* simultaneously. KPCE solves this problem well with the aid of the pointer network and achieves a much higher recall.

Model	NC #	Len_{NC}	Prec.	Recall_R	F1_R
Trained on CN-DBpedia					
Hearst	222	**5.95**	**95.24%**	21.66%	35.29%
FLAIR	64	3.09	95.71%	6.24%	11.72%
XLNet	47	2.66	88.48%	4.68%	8.90%
KVMN	254	4.03	64.45%	26.02%	37.08%
XLM-R	255	5.35	76.82%	24.78%	37.47%
BBF	26	4.34	88.28%	2.54%	4.93%
GACEN	346	3.58	84.89%	36.73%	51.27%
MRC-CE	323	5.33	92.12%	31.51%	46.96%
KPCE	**482**	5.52	94.20%	**44.38%**	**60.33%**
w/o P	338	5.21	72.07%	34.05%	46.25%
Trained on Probose					
Hearst	287	**2.43**	**89.04%**	17.10%	28.69%
FLAIR	140	1.68	84.31%	7.73%	14.16%
XLNet	342	1.51	79.30%	18.87%	30.49%
KVMN	403	1.97	47.39%	22.24%	30.27%
XLM-R	322	2.28	81.73%	17.77%	29.19%
BBC	154	1.68	81.13%	8.44%	15.30%
GACEN	486	1.75	76.93%	31.82%	45.02%
MRC-CE	598	2.23	88.59%	33.00%	48.09%
KPCE	**752**	2.31	88.69%	**46.83%**	**61.30%**
w/o P	691	2.26	78.64%	40.62%	53.57 %

Table 1: Concept extraction performance comparisons of 500 test samples. *w/o P* is the ablation variants of KPCE without the knowledge-guided prompt (P)

Concept_O	Concept_B	$\text{KPCE}_{w/o P}$	KPCE
writer	book	20%	7%
plant	doctor	8%	0%
illness	medicine	12%	6%
singer	music	19%	2%
poem	poet	25%	1%

Table 2: The bias rates issued by KPCE *w/o P* and KPCE in five selected concepts. $Concept_O$ is the original concept and $Concept_B$ is the biased concept.

5.4 Analysis

In response to the motivations of KPCE, we conduct detailed analyses to further understand KPCE and why it works.

How does KPCE alleviate the concept bias? As mentioned in § 3.2, the concept bias occurs primarily among 26 concepts in CN-DBpedia. To justify that KPCE can alleviate concept bias with the aid of prompts, we randomly select five concepts and run KPCE with its ablated variant to extract concepts for 100 entities randomly selected from each of the five concepts. Then we calculate the bias rates of each concept, and the results in Table 2 show that KPCE has a much lower bias rate than the vanilla BERT-based concept extractor. Thus, the knowledge-guided prompt can significantly mitigate the concept bias.

Furthermore, a case study for an entity *Korean alphabet* is shown in Table 3. We find that the proposed prompts can mitigate the spurious co-occurrence correlation between entities and biased concepts by decreasing the confidence scores of biased concepts (*i.e.*, *language* and *alphabet*) and increasing the scores of correct concepts (*i.e.*, *system* and *writing system*). Thus, the knowledge-guided prompt can significantly alleviate the concept bias and result in more accurate CE results.

How does the prompt affect the spurious co-occurrence correlations? To explore the rationale behind the prompt-based mediator, we focus on the attention distribution for the special token `[CLS]`, since it is an aggregate representation of the sequence and can capture the sentence-level semantic meaning (Devlin et al., 2019; Chang et al., 2022). Following previous work (Clark et al., 2019), we calculate the attention probabilities of

Topic: Technology. Entity: Korean alphabet. Abstract: The Korean alphabet is a writing system for the Korean language created by King Sejong the Great in 1443.			
Output Results			
KPCE $_{w/o P}$		KPCE	
Span	**C.S.**	**Span**	**C.S.**
language	0.238	system	0.240
alphabet	0.213	writing system	0.219
system	0.209	system for the Korean language	0.130

Table 3: A case to verify the effectiveness of the proposed prompts on addressing concept bias. We display an entity *Korean alphabet* with its top-3 extracted spans and the confidence scores (denoted as C.S.)

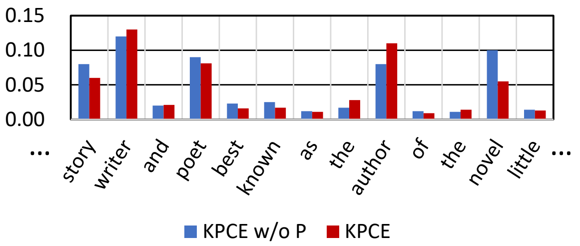

Figure 6: Visualization of the attention distribution of `[CLS]` to other tokens.

`[CLS]` to other tokens by averaging and normalizing the attention value in 12 attention heads in the last layers. The attention distributions of the KPCE and its ablation variant are visualized in Figure 6. We find that the tokens of *writer* and *novel* both have high attentions in the vanilla BERT-based concept extractor. However, after adopting our knowledge-guided prompt, the attention probabilities of *novel* is lower than before, and thus can help the model to reduce the spurious co-occurrence correlations derived from pre-trained knowledge.

What if other knowledge injection methods are adopted? We claim that the topics obtained from external KGs are better than the keyword-based topics from the text on guiding BERT to achieve our CE task. To justify it, we compare KPCE with another variant, namely KPCE $_{LDA}$, where the topics are the keywords obtained by running Latent Dirichlet Allocation (LDA) (Blei et al., 2001) over the abstracts of all entities. Besides, we also compare KPCE with ERNIE (Zhang et al., 2019), which implicitly learns the knowledge of entities during pre-training. The detail about LDA and ERNIE is shown in Appendix C.4. The comparison results are listed in Table 4. It shows that our design of the knowledge-guided prompt in KPCE exploits the value of external knowledge more thoroughly than the two remaining schemes, thus achieving better CE performance.

Model	NC #	Prec.	Recall$_R$	F1$_R$
Trained on CN-DBpedia				
KPCE	**482**	**94.20%**	**85.23%**	**89.49%**
KPCE $_{LDA}$	308	93.08%	82.13%	87.26%
ERNIE	302	93.86%	80.53%	86.69%
Trained on Probose				
KPCE	**752**	**88.69%**	**80.85%**	**84.59%**
KPCE $_{LDA}$	381	68.23%	61.45%	64.66%
ERNIE	286	77.96%	46.13%	57.97%

Table 4: Concept extraction results with different knowledge utilization.

Model	TS #	NC #	Prec.	Recall$_R$	F1$_R$
KPCE	0	62	82.66%	48.44%	61.08%
w/o P	0	55	69.62%	42.97%	53.14%
KPCE	300	**107**	**82.95%**	**83.59%**	**83.27%**
w/o P	300	89	81.65%	69.53%	75.10%

Table 5: Human evaluation on 100 CE results for Meituan entities. TS # is the number of training samples.

5.5 Applications

KG Completion We run KPCE for all entities existing in CN-DBpedia to complement new concepts. KPCE extracts 7,623,111 new concepts for 6 million entities. Thus, our framework can achieve a large-scale concept completion for existing KGs.

Domain Concept Acquisition We collect 117,489 Food & Delight entities with their descriptive texts from Meituan [6], and explore two application approaches. The first is to directly apply KPCE, and the second is to randomly select 300 samples as a small training set to fine-tune KPCE. The results in Table 5 show that: *1)* The transfer ability of KPCE is greatly improved with the aid of prompts; *2)* KPCE can extract high-quality concepts in the new domain only with a small portion of training samples. Furthermore, after running directly, KPCE extracts 81,800 new concepts with 82.66% precision. Thus, our knowledge-guided prompt can significantly improve the transfer ability of PLMs on the domain CE task.

6 Conclusion

In this paper, we identify the concept bias in the PLM-based CE system and devise a Structural

[6] http://www.meituan.com, a Chinese e-business platform.

Causal Model to analyze the bias. To alleviate concept bias, we propose a novel CE framework with knowledge-guided prompting to alleviate spurious co-occurrence correlation between entities and biased concepts. We conduct extensive experiments to justify that our prompt-based learning framework can significantly mitigate bias and has an excellent performance in concept acquisition.

7 Limitations

Although we have proven that our work can significantly alleviate concept bias and extract high-quality and new concepts, it also has some limitations. In this section, we analyze three limitations and hope to advance future work.

Model Novelty Although KPCE can effectively mitigate the spurious co-occurrence correlations between entities and biased concepts, the proposed framework is not entirely novel. The novelty of our work is to conduct the first thorough causal analysis that shows the spurious correlations between entities and biased concepts in the concept extraction task. After defining the problem and SCM of concept extraction in § 3.1, we propose a prompt-based approach to implement the interventions toward the SCM to elicit the unbiased knowledge from PLMs. Previous work in language prompting mostly guides the PLMs with prompts but is unaware of the cause-effect relations in its task, which may hinder the effectiveness of prompts. We hope our work can inspire future work to utilize language prompting from a causal perspective.

Topic Classification Although the topics obtained by clustering are mostly mutually exclusive, there are still cases where an entity can be classified into multiple topics. Therefore, considering only one topic for the entity excludes the correct concepts.

Threshold Selection We only reserve concepts with confidence scores bigger than the selection threshold (§ 4.2), which can hardly achieve a satisfactory balance of precision and recall. If we select a relatively big threshold, we can get more accurate concepts but may lose some correct ones. If the recall is preferred, precision might be hurt.

We suggest that future work consider these three limitations to achieve better performance in the CE task.

Acknowledgement

We would like to thank the anonymous reviewers for their valuable comments and suggestions for this work. This work is supported by the Chinese NSF Major Research Plan (No.92270121), Shanghai Science and Technology Innovation Action Plan (No.21511100401) and the Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902).

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

S Aranganayagi and Kuttiyannan Thangavel. 2007. Clustering categorical data using silhouette coefficient as a relocating measure. In *International conference on computational intelligence and multimedia applications (ICCIMA 2007)*, volume 2, pages 13–17. IEEE.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Sanjiv K Bhatia et al. 2004. Adaptive k-means clustering. In *FLAIRS conference*, pages 695–699.

D. M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent dirichlet allocation. *The Annals of Applied Statistics*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Boxi Cao, Hongyu Lin, Xianpei Han, Fangchao Liu, and Le Sun. 2022. Can prompt probe pretrained language models? understanding the invisible risks from a causal view. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5796–5808, Dublin, Ireland. Association for Computational Linguistics.

Haw-Shiuan Chang, Ruei-Yao Sun, Kathryn Ricci, and Andrew McCallum. 2022. Multi-cls bert: An efficient alternative to traditional ensembling. *arXiv preprint arXiv:2210.05043*.

Jiangjie Chen, Chun Gan, Sijie Cheng, Hao Zhou, Yanghua Xiao, and Lei Li. 2022. Unsupervised editing for counterfactual stories. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10473–10481.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Songtao Fang, Zhenya Huang, Ming He, Shiwei Tong, Xiaoqing Huang, Ye Liu, Jie Huang, and Qi Liu. 2021. Guided attention network for concept extraction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1449–1455. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Matt W Gardner and SR Dorling. 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636.

Navita Goyal, Roodram Paneri, Ayush Agarwal, Udit Kalani, Abhilasha Sancheti, and Niyati Chhaya. 2022. CaM-Gen: Causally aware metric-guided text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2047–2060, Dublin, Ireland. Association for Computational Linguistics.

Fred X Han, Di Niu, Haolan Chen, Weidong Guo, Shengli Yan, and Bowei Long. 2020. Meta-learning for query conceptualization at web scale. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3064–3073.

Zhiting Hu and Li Erran Li. 2021. A causal lens for controllable text generation. In *Advances in Neural Information Processing Systems*.

Jie Ji, Bairui Chen, and Hongcheng Jiang. 2020. Fully-connected lstm–crf on medical concept extraction. *International Journal of Machine Learning and Cybernetics*, 11(9):1971–1979.

Meng Jiang, Jingbo Shang, Taylor Cassidy, Xiang Ren, Lance M Kaplan, Timothy P Hanratty, and Jiawei Han. 2017. Metapad: Meta pattern discovery from massive text corpora. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 877–886.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2022. Clin-x: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain. *Bioinformatics*, 38(12):3267–3274.

Chenguang Li, Jiaqing Liang, Yanghua Xiao, and Haiyun Jiang. 2021. Towards fine-grained concept generation. *IEEE Transactions on Knowledge and Data Engineering*.

Lantian Li, Weizhi Xu, and Hui Yu. 2020. Character-level neural network model based on nadam optimization and its application in clinical concept extraction. *Neurocomputing*, 414:182–190.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022. How pre-trained language models capture factual knowledge? a causal-inspired analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1720–1732, Dublin, Ireland. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th*

International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4582–4597, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

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*.

Yujie Lu, Weixi Feng, Wanrong Zhu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. Neuro-symbolic causal language planning with commonsense prompting. *arXiv preprint arXiv:2206.02928*.

Xuan Luo, Yiping Yin, Yice Zhang, and Ruifeng Xu. 2021. A privacy knowledge transfer method for clinical concept extraction. In *International Conference on AI and Mobile Services*, pages 18–28. Springer.

Xusheng Luo, Luxin Liu, Yonghua Yang, Le Bo, Yuanpeng Cao, Jinghang Wu, Qiang Li, Keping Yang, and Kenny Q Zhu. 2020. Alicoco: Alibaba e-commerce cognitive concept net. In *Proceedings of the 2020 ACM SIGMOD international conference on management of data*, pages 313–327.

Ujjwal Maulik and Sanghamitra Bandyopadhyay. 2002. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on pattern analysis and machine intelligence*, 24(12):1650–1654.

Yuyang Nie, Yuanhe Tian, Yan Song, Xiang Ao, and Xiang Wan. 2020. Improving named entity recognition with attentive ensemble of syntactic information. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4231–4245, Online. Association for Computational Linguistics.

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. 2022. Retrieval-guided counterfactual generation for QA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1670–1686, Dublin, Ireland. Association for Computational Linguistics.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Timo Schick and Hinrich Schutze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Timo Schick and Hinrich Schütze. 2021. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bernhard Schölkopf. 2022. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 765–804.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

MK Vijaymeena and K Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2):19–28.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Noah Weber, Rachel Rudinger, and Benjamin Van Durme. 2020. Causal inference of script knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7583–7596, Online. Association for Computational Linguistics.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, pages 481–492.

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. Cn-dbpedia: A never-ending chinese knowledge extraction system. In *Proc. of IEA-AIE*, pages 428–438. Springer.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.

Siyu Yuan, Deqing Yang, Jiaqing Liang, Zhixu Li, Jinxi Liu, Jingyue Huang, and Yanghua Xiao. 2022. Generative entity typing with curriculum learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3061–3073, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siyu Yuan, Deqing Yang, Jiaqing Liang, Jilun Sun, Jingyue Huang, Kaiyan Cao, Yanghua Xiao, and Rui Xie. 2021a. Large-scale multi-granular concept extraction based on machine reading comprehension. In *International Semantic Web Conference*, pages 93–110. Springer.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021b. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

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

Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. 2021. Care: Commonsense-aware emotional response generation with latent concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14577–14585.

A Theoretical Details of Causal Framework

A.1 Preliminaries

SCM The Structural Causal Model (SCM) is associated with a graphical causal model to describe the relevant variables in a system and how they interact with each other. An SCM $G = \{V, f\}$ consists of a set of nodes representing variables V, and a set of edges between the nodes as functions f to describe the causal relations. Figure 3 shows the SCM that describes the PLM-based CE system. Here the input text X serves as the *treatment*, and the extracted concept span S is the *outcome*. In our SCM, pre-trained knowledge K is a cause of both X and S, and thus K is a *confounder*. A confounder can open *backdoor paths* (*i.e.*, $X \leftarrow K \rightarrow S$) and cause a spurious correlation between X and S. To control the confounding bias, intervention techniques with the do-operator can be applied to cut off backdoor paths.

Causal Intervention To identify the true causal effects of $X \rightarrow S$, we can adopt the causal intervention to fix the input $X = x$ and removes the correlation between X and its precedents, denoted as $do(X = x)$. In this way, the true causal effects of $X \rightarrow S$ can be represented as $P(S = s|do(X = x))$. The backdoor adjustment and the frontdoor adjustment are two operations to implement interventions and obtain $P(S = s|do(X = x))$.

Next, we will elaborate on the details of the two operations.

A.2 The Backdoor Adjustment

The backdoor adjustment is an essential tool for causal intervention. For our SCM, the pre-trained knowledge blocks the backdoor path between X and S, then the causal effect of $X = x$ on S can be calculated by:

$$\begin{aligned} &P(S = s|do(X = x)) \\ &= P_m(S = s|X = x) \\ &= \sum_k P_m(S = s|X = x, K = k)P_m(K = k) \\ &= \sum_k P(S = s|X = x, K = k)P(K = k), \end{aligned} \tag{11}$$

where P_m is the probability after applying the do-operator, and $P(K = k)$ needs to be estimated from data or priorly given. However, it is intractable to observe the pre-training data and obtain the prior distribution of the pre-trained knowledge. Therefore, the back adjustment is not applicable in our case.

A.3 The Frontdoor Adjustment

The frontdoor adjustment is a complementary approach to applying the intervention when we cannot identify any set of variables that obey the backdoor adjustment.

In our SCM, we aim to estimate the direct effect of X on S, while being unable to directly measure pre-trained knowledge K. Thus, we introduce a topic prompt P as a mediator, and then the frontdoor adjustment can adopt a two-step do-operation to mitigate bias.

Step 1 As illustrated in § 3.3, we first analyze the causal effect $X \rightarrow P$. Since the collider, *i.e.*, $X \rightarrow P \rightarrow S \leftarrow K$ blocks the association between P and K, there is no backdoor path from X to P. Thus we can obtain the conditional probability as (same as Eq.1):

$$P(P = p|do(X = x)) = P(P = p|X = x). \tag{12}$$

To explain Step 1, we take an entity *Louisa May Alcott* with her abstract as an example. We can assign the topic *person* as a prompt to make the PLM-based extractor alleviate spurious correlation between *Louisa May Alcott* and *novel*, and concentrate on extracting person-type concepts.

Step 2 In this step, we investigate the causal effect $P \rightarrow S$. $P \leftarrow X \leftarrow K \rightarrow S$ contains a backdoor from P to S. Since the data distribution of X can be observed, we can block the backdoor path through X:

$$\begin{aligned} &P(S = s|do(P = p)) \\ &= \sum_x P(S = s|X = x, P = p)P(X = x), \end{aligned} \tag{13}$$

where $P(X = x)$ can be obtained from the distribution of the input data, and $P(S = s|X = x, P = p)$ is the conditional probability of the extracted span given the abstract with a topic prompt, which can be estimated $P(S = s|X = x, P = p)$ by the concept extractor.

Now we can chain the two steps to obtain the causal effect $X \rightarrow S$:

$$
\begin{aligned}
& P(S|do(X)) \\
& = \sum_p P(S|p, do(X))P(p|do(X)) \\
& = \sum_p P(S|do(P), do(X))P(p|do(X)) \\
& = \sum_p P(S|do(P))P(p|do(X)). \qquad (14)
\end{aligned}
$$

B Detailed Information about KPCE

Topic	Corresponding Concept Examples
Person	politicians, teachers, doctors
Book	romance novels, art books, fantasy novels
Location	towns, villages, scenic spots
Film and TV	movies, animation, TV dramas
Language	idioms, medical terms, cultural terms
Game	electronic games, web games, mobile games
Creature	plants, animals, bacteria
Food	Indian cuisine, Japanese cuisine
Website	government websites, corporate websites
Music	singles, songs, pop music
Software	application software, system software
Folklore	social folklore, belief folklore
Organization	companies, brands, universities
History	ancient history, modern history
Disease	digestive system disease, oral disease
Technology	technology products, electrical appliances
Medicine	Chinese medicine, Western medicine

Table 6: 17 typical topics and their corresponding concept examples.

B.1 Identifying Topic for Each Entity

The 17 typical topics and their corresponding concepts are listed in Table 6. We predict the topic of the entity as one of the 17 typical topics using a transformer encoder-based topic classifier. We randomly fetch 40,000 entities together with their existing concepts in the KG. According to the concept clustering results, we can assign each topic to the entities. Specifically, we concatenate E and X as input to the classifier. With multi-headed self-attention operation over the input token sequence, the classifier takes the final hidden state (vector) of a token `[CLS]`, i.e., $\mathbf{h}^{N_L}_{\texttt{[CLS]}} \in \mathbb{R}^{d''}$, to compute the topic probability distribution $P(T^i|E, X) \in \mathbb{R}^{17}$, where N_L is the total number of layers and d'' is the vector dimension. Then, we identify the topic with the highest probability T as the topic of X,

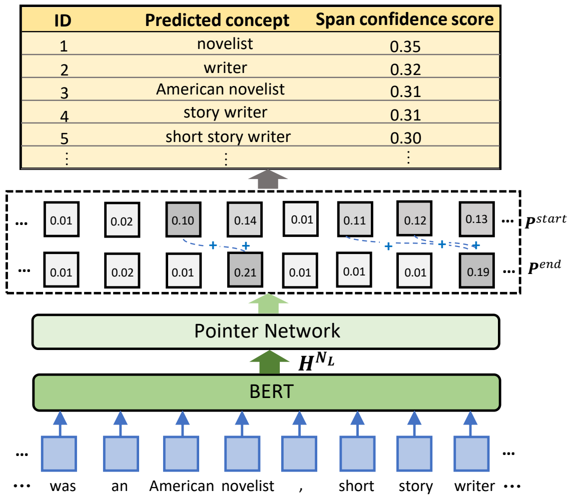

Figure 7: An example to illustrate how to perform the pointer network.

which is calculated as follows,

$$
\begin{aligned}
\mathbf{H}^0 &= \mathbf{E}\mathbf{W}^0 + \mathbf{B}^0, && (15)\\
\mathbf{H}^l &= \mathit{encoder}(\mathbf{H}^{l-1}), 1 \leq l \leq N_L, && (16)\\
P(T) &= \mathrm{softmax}(\mathbf{h}^{N_L}_{\texttt{[CLS]}}\mathbf{W}^L), && (17)\\
T &= \arg\max_i \left(P(T^i)\right), 1 \leq i \leq 17 && (18)
\end{aligned}
$$

where $\mathbf{E} \in \mathbb{R}^{(|\mathbf{E}|+|\mathbf{X}|)\times d}$ is the random initial embedding matrix of all input tokens and d is the embedding size. $\mathbf{H}^l \in \mathbb{R}^{(|\mathbf{E}|+|\mathbf{X}|)\times d''}$ is the hidden matrix of the l-th layer. $\mathbf{h}^{N_L}_{\texttt{[CLS]}}$ is obtained from $\mathbf{H}^{N_L}$. Furthermore, $\mathbf{W}^0 \in \mathbb{R}^{d\times d''}$, $\mathbf{B}^0 \in \mathbb{R}^{(|\mathbf{E}|+|\mathbf{X}|)\times d''}$ and $\mathbf{W}^L \in \mathbb{R}^{d''\times 17}$ are both training parameters.

B.2 An Example for Point Network

As mentioned in § 4.2, we adopt a point network to achieve multi-grained concept extraction (Yuan et al., 2021a). The model generates a ranking list of candidate concepts (spans) along with their confidence scores, and outputs the concepts with confidence scores bigger than the selection threshold. Note that one span may be output repeatedly as the same subsequence of multiple extracted concepts through an appropriate selection threshold.

For example, as shown in Figure 7, *writer* is extracted multiple times as the subsequence of three different granular concepts when the confidence score threshold is set to 0.30. Therefore, the point network enables our framework to extract multi-grained concepts.

C Experiment Detail

C.1 Baselines

We compare KPCE with seven baselines. Most of the compared models are the extraction models feasible for extraction tasks, including Named Entity Recognition (NER), Relation Extraction (RE), and Open Information Extraction (Open IE). In addition, we also compare the pattern-based approach. However, we do not compare ontology extension models and generation models, since both do not meet our scenario. Since entity typing models cannot find new concepts, they are also excluded from our comparison. Please note that, except MRC-CE, other baselines applied in concept extraction cannot extract multi-grained concepts.

- **Hearst** (Jiang et al., 2017): With specific handwritten rules, this baseline can extract concepts from free texts. We design 5 Hearst patterns listed in Table 7 where we translate the Chinese patterns for the Chinese dataset into English.

- **FLAIR** (Akbik et al., 2019): It is a novel NLP framework that combines different words and document embeddings to achieve excellent results. FLAIR can also be employed for concept extraction since it can extract spans from the text.

- **XLNet** (Yang et al., 2020): With the capability of modeling bi-directional contexts, this model can extract clinical concepts effectively.

- **KVMN** (Nie et al., 2020): As a sequence labeling model, KVMN is proposed to handle NER by leveraging different types of syntactic information through the attentive ensemble.

- **XLM-R** (Conneau et al., 2020; Lange et al., 2022): It is a Transformer-based multilingual masked language model incorporating XLM (Conneau and Lample, 2019) and RoBERTa (Liu et al., 2019), which has proven to be effective in extracting concepts.

- **BBF** (Luo et al., 2021): It is an advanced version of BERT built with Bi-LSTM and CRF. With optimal token embeddings, it can extract high-quality medical and clinical concepts.

- **GACEN** (Fang et al., 2021): The model incorporates topic information into feature representations and adopts a neural network to pre-train a soft matching module to capture semantically similar tokens.

- **MRC-CE** (Yuan et al., 2021a): MRC-CE handles the concept extraction problem as a Machine Reading Comprehension (MRC) task built with an MRC model based on BERT. It can find abundant new concepts and handle the problem of concept overlap well with a pointer network.

Dataset	Pattern
CN-DBpedia	X is Y X is one of Y X is a type/a of Y X belongs to Y Y is located/founded/ in...
Probase	X is a Y that/which/who X is one of Y X refers to Y X is a member/part/form... of Y As Y, X is ...

Table 7: The Hearst patterns used in the baseline.

C.2 Experiment Settings

Our experiments are conducted on a workstation of dual GeForce GTX 1080 Ti with 32G memory and the environment of torch 1.7.1. We adopt a BERT-base with 12 layers and 12 self-attention heads as the topic classifier and concept extractor in KPCE. The training settings of our topic classifier are: d = 768, batch size = 16, learning rate = 3e-5, dropout rate = 0.1 and training epoch = 2. The training settings of our concept extractor are: d = 768, m = 30, batch size = 4, learning rate = 3e-5, dropout rate = 0.1 and training epoch = 2. The α in Eq.8 is set to 0.3 and the selection threshold of candidate spans in the concept extractor is set to 0.12 based on our parameter tuning.

C.3 Human Assessment

Some extracted concepts do not exist in the KG, which cannot be automatically assessed. Therefore, we invite some volunteers to assess whether the extracted concepts are correct for the given entities. We denote an extracted concept as an *EC* (existing concept) that has already existed in the KG for the given entity. We denote an extracted concept as an *NC* (new concept) represents a correct concept not existing in the KG for the given entity. We employ four annotators in total to ensure the quality of the assessment. All annotators are native Chinese and proficient in English. Each concept is labeled with 0, 1 or 2 by three annotators, where 0 means a wrong concept for the given entity, while 1 and 2 represent EC and NC, respectively. If the results from the three annotators are different, the fourth annotator will be hired for a final check. We protect the privacy rights of the annotators and pay the annotators above the local minimum wage.

C.4 Other Knowledge Injection Methods

As we mentioned before, the topics of the knowledge-guided prompt come from external KGs, which are better than the keyword-based topics from the text on guiding BERT to achieve accurate concept extraction.

To justify it, we compared KPCE with another variant, namely KPCE $_{LDA}$, where the topics are the keywords obtained by running Latent Dirichlet Allocation (LDA) (Blei et al., 2001) over all entities' abstracts. Specifically, the optimal number of LDA topic classes was also determined as 17 through our tuning study. For a given entity, its topic is identified as the keyword with the highest probability of its topic class. Besides, we also compared KPCE with ERNIE. ERNIE (Zhang et al., 2019) adopts entity-level masking and phrase-level masking to learn language representation. During pre-training of ERNIE, all words of the same entity mentioned or phrase are masked. In this way, ERNIE can implicitly learn the prior knowledge of phrases and entities, such as relationships between entities and types of entities, and thus have better generalization and adaptability.

The comparison results are listed in Table 4, which shows that our design of the knowledge-guided prompt in KPCE exploits external knowledge's value more thoroughly than the rest two schemes.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☑ A4. Have you used AI writing assistants when working on this paper?
Check grammar for the whole paper

B ☑ Did you use or create scientific artifacts?

Section 5.1

☑ B1. Did you cite the creators of artifacts you used?
Sections 5.1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 5.1

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 5.1

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Our data is collected from the existing KGs, and there is no offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 5.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5.1

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix C.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix C.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5.2

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We do not use existing packages

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Appendix C.3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix C.3

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix C.3

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix C.3

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. We do not have any human subjects research.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix C.3