PRINCE: Prefix-Masked Decoding for Knowledge Enhanced Sequence-to-Sequence Pre-Training

Song Xu*, **Haoran Li***, **Peng Yuan, Youzheng Wu, Xiaodong He**
JD AI Research
{xusong28, lihaoran24}@jd.com

Abstract

Pre-trained Language Models (PLMs) have shown effectiveness in various Natural Language Processing (NLP) tasks. Denoising autoencoder is one of the most successful pre-training frameworks, learning to recompose the original text given a noise-corrupted one. The existing studies mainly focus on injecting noises into the input. This paper introduces a simple yet effective pre-training paradigm, equipped with a knowledge-enhanced decoder that predicts the next entity token with noises in the prefix, explicitly strengthening the representation learning of entities that span over multiple input tokens. Specifically, when predicting the next token within an entity, we feed masks into the prefix in place of some of the previous ground-truth tokens that constitute the entity. Our model achieves new state-of-the-art results on two knowledge-driven data-to-text generation tasks with up to 2% BLEU gains.

1 Introduction

Pre-trained language models (PLMs), such as BERT (Devlin et al., 2019), MASS (Song et al., 2019), BART (Lewis et al., 2020), and T5 (Raffel et al., 2020), have had remarkable performances in various Natural Language Processing (NLP) tasks thanks to the use of denoising autoencoder pre-training schema that is optimized to reconstruct the original text given a noise-corrupted one. Despite its common usage, to the best of our knowledge, existing work mainly focuses on injecting noises into the encoding sequence, while the feasibility of injecting noises into the decoding sequence for PLMs remains an open question. This is the primary interest of this work.

On the other hand, recent researches (Zhang et al., 2019; Peters et al., 2019; Xiong et al., 2020; Wang et al., 2019, 2020; Ke et al., 2020; Tian et al., 2020; Xu et al., 2021) demonstrate that enhancing PLMs with real-world knowledge is crucial for knowledge-driven downstream tasks, including entity typing, relation classification, sentiment analysis, and entity-related question answering. However, the pre-training objectives for the above-mentioned studies are usually designed for Natural Language Understanding (NLU) tasks. In this work, we focus on knowledge-oriented sequence-to-sequence (Seq2Seq) pre-training objectives for PLMs, so that they can be applied to Natural Language Generation (NLG) tasks, such as the data-to-text task.

Given a noise-corrupted sentence, such as "*Tom Cruise was born in [MASK] [MASK] [MASK] in the year 1962*", the standard Seq2Seq pre-training would predict the masked text fragment (i.e., "*New York City*") token by token, while our proposed pre-training would generate the correct output by learning facts regarding the concerned entity. We argue that given merely "*Tom Cruise was born in*", if a PLM can immediately predict the masked fragment to be "*New York City*", it suggests that the PLM has learned the fact ("*Tom Cruise*", "*born in*", "*New York City*"). In contrast, if the PLM could only predict "*York*" after being provided with the entity's partial ground-truth token "*New*", which is what most existing PLMs are capable of, although it indeed correctly predicts the masked tokens, it does not truly learn the fact or the entity of Tom Cruise's birthplace. Thus, we propose **PR**ef**I**x-masked decoding for k**N**owledge enhan**CE**d sequence-to-sequence pre-training (PRINCE), which decodes entity tokens with noisy prefixes rather than ground-truth tokens. For example, when predicting "*York*", a mask symbol is fed into the decoder as the prefix, in place of "*New*".

Different from the work of Chen et al. (2020) that generates knowledge-enriched text based on a knowledge subgraph from WikiData, PRINCE can directly incorporate entity knowledge into PLMs based on the raw text with extracted entities, alleviating error propagations throughout

*Equal contribution.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2675 - 2681

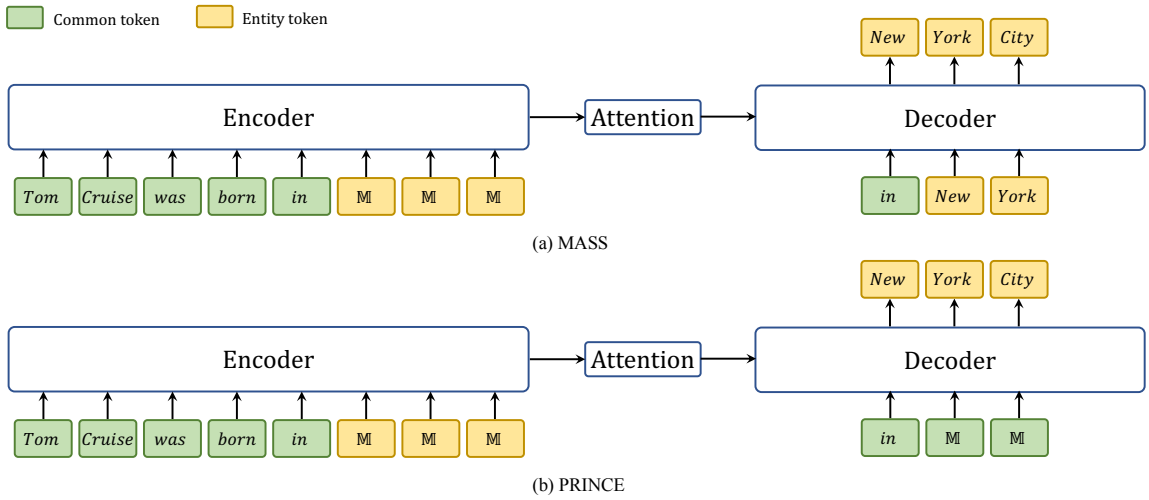

Figure 1: A pre-training framework comparison of MASS and our proposed PRINCE. The major difference is that PRINCE predicts entity tokens without feeding the previous ground-truth entity token into the decoder.

the additional data processing, such as entity alignment and knowledge triple retrieval. After pre-training PRINCE, we evaluate it on two data-to-text datasets that require entity knowledge: WebNLG (Shimorina and Gardent, 2018) and WikiBio (Liu et al., 2018). On both datasets, our model achieves new state-of-the-art results.

Our main contributions are as follows:

- We present PRINCE that predicts entity tokens with masked prefixes, aiming to improve the representation learning of entities that span over multiple tokens.
- PRINCE exhibits new state-of-the-art performances on two data-to-text tasks.

2 Background

We first introduce the existing denoising autoencoder schema for the Seq2Seq pre-training. Given a sentence with a masked text fragment, MASS (Song et al., 2019) proposes a masked Seq2seq pre-training objective that is optimized to predict the masked tokens auto-regressively. Recent work, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), has exhibited gains by applying a wide range of approaches to inject noises into the input, such as sentence permutation, document rotation, and text infilling. In this work, we adopt the framework of MASS. PRINCE is simple and flexible that we anticipate it would be able to integrate to other denoising autoencoder schemas with more sophisticated noising approaches.

Next, we will describe the basic framework of the masked Seq2seq pre-training. Given a sentence $\mathbf{x} = \{x_1, x_2, \cdots, x_s\}$, $\mathbf{x_{predict}} = \{x_p, x_{p+1}, \cdots, x_q\}$ is a text span from $\mathbf{x}$ where $1 \leq p < q \leq s$. $\mathbf{x_{mask}}$ is a sequence in which the tokens ranging from x_p to x_q are replaced by the mask symbol $\mathbb{M}$. The masked Seq2seq pre-training model maximizes the conditional probability of $\mathbf{x_{predict}}$: $P(\mathbf{x_{predict}}|\mathbf{x_{mask}})$.

The loss function $\mathcal{L}_t$ for each time t is the negative log likelihood of the token $x_t \in \mathbf{x_{predict}}$:

$$\mathcal{L}_t = -\log P(x_t|\{x_p, \cdots, x_{t-1}\}, \mathbf{x_{mask}}) \quad (1)$$

3 Our Model

PRINCE is built upon the masked Seq2seq pre-training. The difference is that PRINCE predicts entity tokens with noises, in place of the previous ground-truth entity token, as the decoding prefix. The framework of PRINCE is shown in Figure 1. Formally, for a fragment of entity $\mathbf{x_{entity}} = \{x_m, x_{m+1}, \cdots, x_n\} \in \mathbf{x_{predict}}$ where $p \leq m < n \leq q$, PRINCE predicts $x_t \in \mathbf{x_{entity}}$ $(m < t \leq n)$ by replacing the previous ground-truth entity tokens $\{x_m, \cdots, x_{t-1}\}$ with $\mathbb{M}$. The loss is:

$$\mathcal{L}_t = -\log P(x_t|\{x_p, \cdots, x_{m-1}, \mathbb{M}^*\}, \mathbf{x_{mask}}) \quad (2)$$

where we denote the replaced sequence of entity tokens as $\mathbb{M}^*$ for simplicity.

For the example shown in Figure 1, the masked entity fragment is $\{New, York, City\}$. When predicting "*York*", for MASS, the previous ground-truth sequence "*New*" is fed into the decoder. While

for PRINCE, "*New*" is replaced by $\mathbb{M}$. In this way, each token in the entity fragment is generated without any indication from other entity tokens.

4 Experiments

4.1 Model Architecture

PRINCE applies the sequence-to-sequence Transformer architecture (Vaswani et al., 2017a), consisting of a 12-layer encoder and a 12-layer decoder. The size of hidden vectors is set to 1024. We adopt GELU activation (Hendrycks and Gimpel, 2016). We use Adam optimizer (Kingma and Ba, 2015) with a learning rate of 3e-5, β_1 = 0.9, β_2= 0.98, weight decay of 0.01. The dropout probability is 0.1. The maximum sequence length is set to 512. Pre-training takes 5 days with 8 Telsa V100 GPUs. We use the beam search with a beam size of 4 for the inference. Other hyper-parameters can be found in our code.

4.2 Pre-training Dataset

We use English Wikipedia as the source of our pre-training data, which is aligned to Wikidata. Entity tokens can be extracted through the alignment[1]. We set the max length of the sentence to 256 and abandon sentences containing less than three entities. To obviate data leakage during pre-training, we discard the pre-training data overlapped with the samples in downstream datasets. In the end, we obtain 14GB of pre-training data.

4.3 Pre-training Details

PRINCE prioritizes masked fragments covering entity tokens, and the number of the masked tokens is set to 30% of the length of the input sentence. When injecting noises into the decoder, following Devlin et al. (2019), the noise will be a mask symbol 80% of the time, a random token 10% of the time, and an unchanged token in the rest of the time. We adopt the architecture of BART large model.

4.4 Fine-Tuning on Data-to-Text Tasks

WebNLG (2.0) Dataset (Shimorina and Gardent, 2018) This dataset takes RDF triples as input and outputs a textual description.

WikiBio Dataset (Lebret et al., 2016) This dataset takes a Wikipedia infoboxes table as input and outputs a biography description.

[1]We use the same data as KGPT (Chen et al., 2020).

4.5 Experimental Results

4.5.1 Results on WebNLG

Models	BLEU	MTR	RG
Existing Methods			
Seq2Seq	54.0	37.0	64.0
Seq2Seq + Delex	56.0	39.0	67.0
Seq2Seq + Copy	61.0	42.0	71.0
GCN	60.8	42.76	71.13
KGPT-Graph	63.84	46.10	74.04
KGPT-Seq	64.11	46.30	74.57
MASS	60.34	43.61	69.28
ProphetNet	64.39	46.28	74.53
Our Implementations			
Transformer (No PT)	51.40	38.29	59.07
BART (14GB PT)	62.98	45.58	72.53
BART (160GB PT)	64.61	46.78	74.41
PRINCE (14GB PT)	64.35	46.31	74.18
PRINCE (BART + 14GB PT)	**66.86**	**47.48**	**75.99**

Table 1: Results (%) on the WebNLG dataset. MTR and RG are short for METEOR and ROUGE-L. PT is short for Pre-Training.

The results on the WebNLG dataset are shown in Table 1. Shimorina and Gardent (2018) apply **Seq2Seq** model with attention (Luong et al., 2015) as the baseline and address rare words by delexicalization (**Delex**) and copying (**Copy**). The **GCN** model (Marcheggiani and Perez-Beltrachini, 2018) adopts a graph convolutional networks based generator. The **KGPT** model (Chen et al., 2020) is a knowledge-grounded pre-training model with graph (**Graph**) or sequential (**Seq**) encoders. Our implementations are based on **Transformer** (Vaswani et al., 2017b). We first evaluate the **BART** (Lewis et al., 2020) pre-trained with English Wikipedia (14GB) and pre-trained with data used in BART (160GB), respectively, For our proposed pre-training method, we pre-train PRINCE from scratch and warm-start PRINCE with BART (pre-trained with 160GB data).

As we can see, pre-training PRINCE from scratch with 14GB data already achieves comparable performance with BART pre-trained with 160G data. Pre-training PRINCE based on the BART initialization leads to the best performance, which significantly improves the results over the original BART model (+ 2.25%/0.70%/1.58% for BLEU/METEOR/ROUGE-L scores, paired t-test, p-value<0.01).

4.5.2 Results on WikiBio

The results on the WikiBio dataset are shown in Table 2. **Table NLM** (Lebret et al., 2016) is a table-conditioned neural language model. **Order-Planning** (Sha et al., 2018) is an order-planning text generation model. **Field-Gating** (Liu et al., 2018) is a structure-aware seq2seq model with the field-gating encoder. **KBAtt** (Chen et al., 2019) enhances data-to-text model with external background knowledge. **Hierarchical+Auxiliary Loss** (Liu et al., 2019) is a data-to-text model trained with multiple auxiliary objectives.

Models	BLEU
Existing Methods	
Table NLM	34.70
Order-Planning	43.19
Field-Gating	44.71
KBAtt	44.59
Hierarchical + Auxiliary Loss	45.01
KGPT-Graph	45.10
KGPT-Seq	45.06
MASS	44.18
ProphetNet	47.54
Our Implementations	
Transformer (No PT)	42.46
BART (14GB PT)	45.76
BART (160GB PT)	47.80
PRINCE (w/o BART, 14GB PT)	47.38
PRINCE (BART + 14GB PT)	**49.03**

Table 2: Results (%) on the WikiBio dataset.

First, we can find that the models with pre-training outperform the models without pre-training, while the improvement is smaller than that on the WebNLG dataset, which can be ascribed to the larger size of the WikiBio dataset. Second, similar to the results on the WebNLG dataset, pre-training PRINCE with the BART initialization brings significant improvements compared to other models (+ 1.23% for BLEU score over BART, paired t-test, p-value<0.01).

4.6 Further Analysis

4.6.1 Can PRINCE Generate Entities Better?

PRINCE aims to enhance the representation learning of entities that span over multiple tokens. Looking into the WebNLG datasets, we observe that 97.61% entities are composed of multiple tokens after BPE (Sennrich et al., 2016) preprocessing. Can PRINCE generate these entities better? To answer this question, we perform manual evaluations with 100 examples from the test set of WebNLG. Three annotators are involved in deciding whether the generated entities are faithful and readable. The results are shown in Table 3, where faithfulness refers to that the generated text accurately expresses the true meaning of the input, and readability refers to that the generated text is easy to understand. The results depict that PRINCE outperforms other models .

Models	Qualified%
Transformer	87.0
BART	94.0
PRINCE (BART + 14GB PT.)	**99.0**

Table 3: Entity-oriented manual evaluation.

4.6.2 Benefit of Entity-oriented Noises

We evaluate PRINCE (BART initialized) with distinct noising strategies, including injecting noises for entities (our main model), injecting noises for common tokens[2], and no noising. The results in Table 4 demonstrate the superiority of entity-oriented noising against other strategies.

Models	WebNLG	WikiBio
BART	64.61	47.80
PRINCE (Entity noising)	**66.86**	**49.03**
PRINCE (Common token noising)	65.21	47.64
PRINCE (No noising)	65.84	47.96

Table 4: BLEU scores for different noising methods.

4.6.3 Case Study

Table 5 illustrates an example from the WebNLG dataset. PRINCE with common token noising strategy generates a text with wrong team name of "los angeles" and missing information of "draft round 2". By contrast, PRINCE with entity token noising successfully expresses exactly the same meaning as the references.

5 Discussions

5.1 Motivation of PRINCE

Denoising autoencoder pre-training can train the model to learn language representations by transferring a noise-corrupted input back to the original state. Injecting noises into the decoding sequence makes this process more challenging and strengthens the robustness of the decoder. Moreover, knowledge-oriented noises force the model to

[2]For a fair comparison, we randomly inject noises for 25% common tokens.

Input Triplets
LOS_ANGELES_RAMS FORMERTEAM AKEEM_AYERS 2 DRAFTROUND AKEEM_AYERS 39 DRAFTPICK AKEEM_AYERS
References
(1) Former Los Angeles Rams team member, Akeem Ayers, was number 39 in the draft pick, in draft round 2. (2) Akeem Ayers, whose former team was the Los Angeles Rams, was in draft round 2 and his draft pick number was 39.
Result of Transformer (without pre-training)
in draft round 2, akeem ayers was the draft pick and his draft pick is 39.
Result of PRINCE (Common token noising)
akeem ayers was number 39 in the draft pick and used to play for the los angeles.
Result of PRINCE (Entity noising)
akeem ayers, who used to play for the los angeles rams, was number 39 in the draft pick, in draft round 2.

Table 5: Case study.

predict the knowledge with noisy context. Specifically, the noises are injected when the decoder predicts the entity tokens, and the previously generated partial entity tokens are unseen for the latter. In that case, the decoder needs to predict the complete entity tokens without of any clues from the entity itself, which can motivate the model to learn better to predict the entity relying solely on the context. In this way, we argue that our model can enhance the representation learning of knowledge and the ability of knowledge reasoning.

5.2 Type of Knowledge to be Injected Noises

While in this work, we regard entities stored in Wikidata as the knowledge to be masked, PRINCE is actually not designed for any specific type of knowledge. Other knowledge, such as lexical relation (Lauscher et al., 2019; Wang et al., 2020), sentiment words (Ke et al., 2020; Tian et al., 2020), keywords (Li et al., 2020b; Xu et al., 2020), entailment (Eichler et al., 2017; Li et al., 2018), and domain attribute schema (Li et al., 2020a; Zhu et al., 2020), would be compatible with our model as well.

5.3 Comparison with ProphetNet

The motivation of ProphetNet (Qi et al., 2020) partially resembles that of PRINCE. ProphetNet predicts future n-gram based on an n-stream self-attention mechanism that contains a main stream and n predicting stream. Only the main stream is maintained for fine-tuning on the downstream tasks. In contrast, the PRINCE decoder is consistent throughout, bridging the gap between pre-training and fine-tuning. Besides, ProphetNet predicts a fixed n-step ahead, while PRINCE is more flexible designed for knowledge fragments with variable lengths.

6 Conclusion and Future Work

We propose PRINCE, a Seq2Seq pre-training model equipped with a knowledge-enhanced decoder that predicts entity tokens with masked prefixes. PRINCE achieves new state-of-the-art results on two data-to-text datasets. In the future, we will adopt PRINCE to other pre-training schemas and more knowledge-driven tasks. Our code is publicly available[3].

Limitations

A limitation of our work is that it is designed for entity-oriented text generation tasks (i.e., data-to-text tasks), where the text is generated from structural data, such as RDF triples and infoboxes table. Based on the observation of our work, we can conclude that the performance of pre-training models can be improved for data-to-text tasks, while the improvements for other general text-to-text tasks are sometimes not significant. Thus, it asks for further explorations on whether universal pre-training with denoising common token noises is helpful for general text-to-text tasks.

Second, we consider entities as the knowledge in our work, and the pre-training aims to learn better representations for the entities. We argue that a broader definition of the knowledge may lead to a much wider set of application scenarios.

In addition, various types of noises injected into the input text have been proved effective (Lewis et al., 2020; Raffel et al., 2020), while we only test with noises in a simple form. We believe that more complex noises are potentially extended to the decoder.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600. We thank the anonymous reviewers for their helpful comments and suggestions.

[3] https://github.com/xu-song/prince

References

Shuang Chen, Jinpeng Wang, Xiaocheng Feng, Feng Jiang, Bing Qin, and Chin-Yew Lin. 2019. Enhancing neural data-to-text generation models with external background knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3022–3032.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648.

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 (NAACL)*, pages 4171–4186.

Kathrin Eichler, Feiyu Xu, Hans Uszkoreit, and Sebastian Krause. 2017. Generating pattern-based entailment graphs for relation extraction. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 220–229.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2020. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, (ICLR)*.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2019. Specializing unsupervised pretraining models for word-level semantic similarity. *arXiv preprint arXiv:1909.02339*.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213.

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

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8188–8195.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He. 2020b. Keywords-guided abstractive sentence summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8196–8203.

Tianyu Liu, Fuli Luo, Qiaolin Xia, Shuming Ma, Baobao Chang, and Zhifang Sui. 2019. Hierarchical encoder with auxiliary supervision for neural table-to-text generation: Learning better representation for tables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6786–6793.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Diego Marcheggiani and Laura Perez-Beltrachini. 2018. Deep graph convolutional encoders for structured data to text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 1–9.

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

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits

of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang, and Zhifang Sui. 2018. Order-planning neural text generation from structured data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Anastasia Shimorina and Claire Gardent. 2018. Handling rare items in data-to-text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 360–370.

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

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

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

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

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

Song Xu, Haoran Li, Peng Yuan, Yujia Wang, Youzheng Wu, Xiaodong He, Ying Liu, and Bowen Zhou. 2021. K-PLUG: Knowledge-injected pre-trained language model for natural language understanding and generation in E-commerce. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1–17.

Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Self-attention guided copy mechanism for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1355–1362, Online. Association for Computational Linguistics.

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 (ACL)*, pages 1441–1451.

Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Multimodal joint attribute prediction and value extraction for e-commerce product. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2129–2139.