 doi:10.20944/preprints202208.0238.v1

A Simple Survey of Pre-trained Language Models

ZHU Zhenyi
The Hong Kong University of Science and Technology
zzhubh@connect.ust.hk

Abstract

Pre-trained Language Models (PTLM) have remarkable and successful performance in solving lots of NLP tasks nowadays. And previous researchers have created many SOTA models and these models are included in many long surveys(Qiu et al., 2020). So, we would like to conduct a simple and short survey on this topic to help researchers understand the sketch of PTLM more quickly and comprehensively. In this short survey, we would provide a simple but comprehensive review of techniques, benchmarks, and methodologies in PTLM. And we would also introduce the applications evaluation of PTLM in this simple survey.
Keywords : NLP; PLTMs; benchmarks

1 Introduction

Compared with other hot topics in machine learning, such as computer vision, the use and performance of neural models are not significant. The main cause is the volume of the datasets for current supervised NLP problems is still comparable small especially for some topics such as machine translation. And previous neural networks for NLP problems could be comparably shallow and often the layers of the model could not be enough. To solve existing problem, we need to train some PTLM, which can be beneficial for downstream NLP problems. First-generation PTLM mainly focus on learning word embeddings, such as Glove and Skip-Gram. Second-generation PTLM aim in learning contextual word embeddings, such as Bert, Roberta, ELMo(Neumann et al., 2018), OpenAI GPT (Radford et al., 2018) etc. There have been many long and detailed surveys (Qiu et al., 2020) on PTLM, but we lack of some brief surveys of PTMLs. So we try to accomplish this short survey to make up for the condition. And the construction of survey is as follows. Section one would introduce some bench markings, section two would introduce some techniques, section three would introduce some methodologies, section four would introduce some current applications of PTLM, section five would introduce some evaluation of PTLM.

2 Benchmarking

2.1 First Generation PTLM

The first generation of PTLM is about pre-trained word embeddings performance and the word embeddings first comes from the neural network language model(Bengio et al., 2000). Many pretrained models are trained based on shallow neural models, such as Skip-thought vectors (Kiros et al., 2015), Context2Vec (Melamud et al., 2016) and paragraph vector (Le and Mikolov, 2014). These benchmarking are different from current successors, and they just encode the input sentences and then transfer them to fixed dimensional vector representation.

2.2 Second Generation PTLM

The second generation of PTLM is about the pre-trained contextual encoders because most NLP problems is in sentence level. For example, the Seq2Seq models was found by (Ramachandran et al., 2016a). The model can be greatly optimized and improve by using unsupervised pre training models.

2.3 Third Generation PTLM

Currently, these exist many deep PTLM performing great role in universal language representations learning task. For example, the BERT, OpenAI GPT, ULMFiT etc. nowadays, more and more fine-tuning PTLM are appearing, and in next section, we would like to introduce them in detail.

3 Techniques

The pre training process need vast amounts of training data, and we can divide the pre training tasks into several different classes: supervised learning, unsupervised learning, semi-supervised learning,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2022 doi:10.20944/preprints202208.0238.v1

self-supervised learning etc. Here are some techniques in PTLM.

3.1 Language Modeling (LM)

The probabilistic language model is the most common unsupervised model in NLP. Specifically, the LM refers to unidirectional LM or auto-regressive LM. But the drawback is also obvious, for example the representation of the tokens encoding could not be bidirectional contexts tokens.

3.2 Masked language models (MLM)

The Masked language models firstly be proposed by (Taylor, 1957). The MLM model can overcome some drawbacks caused by LM model. This language model first tries to mask out some of the tokens from the sentences imported and second it will train the model to predict the masked ones using the rest tokens.

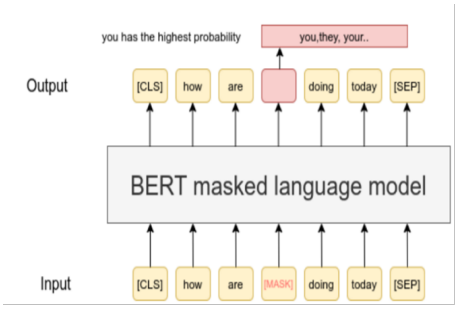

Figure 1: Bert masked language model

And in MLM model there exists a sequence-to-sequence MLM model (Seq2Seq MLM), which use encoder and decoder architecture to do the prediction. To be specific, one masked sequence will be inputted into the encoder and then the decoder will sequentially generate masked tokens in the auto regression way. And there are some improved MLM models called Enhanced Masked Language Modeling (E-MLM) such as Roberta.

3.3 Permuted Language Modeling (PLM)

Some researchers such as (Yang et al., 2019a) in 2019 found that there exist some tokens in MLM could be absent when tackling some downstream tasks, which could make a difference between pre training and fine tuning. So, they design the PLM, and in PLM, only last tokens in permuted sequences would be predicted.

3.4 Seq2seq language models

The Pre-trained Seq2seq language model are essential components of PTLM. They roughly belong to MLM. They could solve Seq2Seq-style tasks such as summarization, question answering and machine translation very well. The following is the framework of one Seq2Seq model:

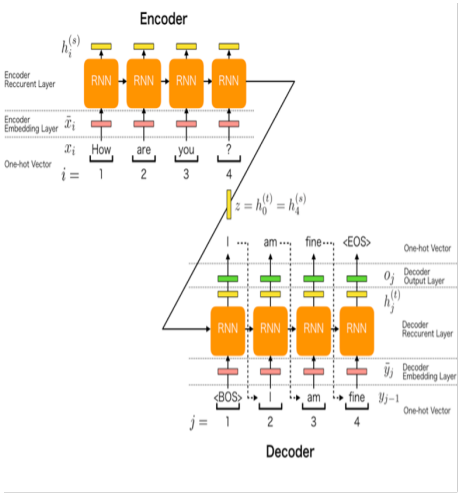

Figure 2: Framework of one Seq2Seq model

There are many typical Seq2Seq models such as Encoder-Decoder (for example in transformers, which is Generic), MASS (Song et al., 2019) T5 (Raffel et al., 2019), BART, MBART, MarianMT, RAG (Retrieval Augmented Generation - E,g, Question Answering). For example, BART use parameters both as an encoder and a decoder, and they are typically used for some enc-dec tasks, and just use the encoder as a replacement for BERT. In conclusion, BART is pre-trained on BART tasks such as take random chunks of text and then noise them, fine-tuned on the summarization dataset, and could get good results with only few summaries to fine-tune on. And for T5, the pre training process is similar to BART, but the input is different where the text here has gaps, and the output are a series of phrases filling those gaps. The following is the framework of T5:

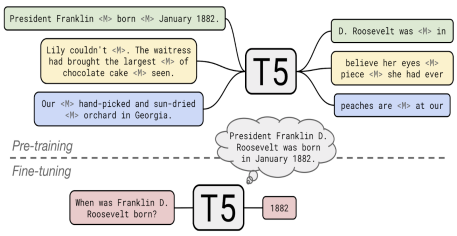

Figure 3: Framework of T5

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2022 doi:10.20944/preprints202208.0238.v1

3.5 Auto-regressive language models

Auto-regressive language models are feed forward models and extremely useful tools for modeling text's probability. And the text is constrained to forward or backward directions (Liu et al., 2021). The following is a sketch of one typical auto-regressive language model. There exist many auto-regressive

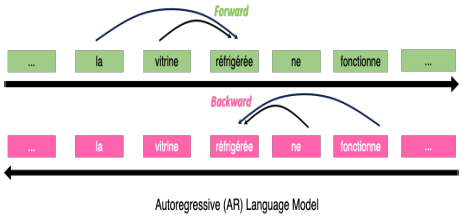

Figure 4: Sketch of one AR language model

language models, such as left-to-right autoregressive LMs, which is suitable for the prefix prompt. In 2019, XLNet was proposed by (Yang et al., 2019b), which is a generalized auto-regressive pretraining tool could overcome Bert's limitation, enable bidirectional learning, and outperforms Bert in some tasks such as QA, document ranking and sentiment analysis. GPT-like is the autoregressive language model, which is a modification of original GPT-2 framework. In 2021, a new model named Pangu-alpha (Zeng et al., 2021) was proposed, which is large-scale autoregressive language model with about 200 billion parameters, could perform various tasks extremely well under zero-shot or few-shot settings. And in 2022, an auto-regressive model named GPT-NeoX-20B is introduced by Black et al., which could gain much more in performance when evaluated five-shot than similarly sized GPT-3 and FairSeq models and serves as a particularly powerful few-shot reasoner (Black et al., 2022). So, in conclusion, the Auto-regressive language models could perform better in many tasks than some baselines.

3.6 Others

There are some other PTMLs techniques performing well, to save the space, I just introduce them briefly. Left-to-Right Language Model (LRLM), which belongs to auto-regressive language model, could predict the next words. Prefix and Encoder-Decoder (PED), the two models could complete translation and summarization tasks, encoding input texts and then generating output texts. Denoising Autoencoder (DAE): the key idea in DAE is to use partially input to get original undistorted input and it has some methods to corrupt context such as token deletion, token masking and text filling. Contrastive Learning (CTL)'s consumes less computation resources compared to LM, which is an alternative training standard for PTMs. NCE (Gutmann and Hyvärinen, 2010), Deep InfoMax (DIM), Replaced Token Detection (RTD), Next Sentence Prediction (NSP) and Sentence Order Prediction (SOP) are well-known examples of CTL.

And the following are some extensions of PTLM. First is Knowledge-Enriched PTMs (KE PTLM), much domain knowledge (semantic (Levine et al., 2019), factual (Zhang et al., 2019), commonsense (Guan et al., 2020), linguistic (Lauscher et al., 2019)) could be incorporated into these models. Multilingual and Language-Specific PTMs (MLS PTMs), which is important in most cross-lingual NLP tasks. Multi-Modal PTMs (MM PTMs) would consider many visual-based MLM, visual linguistic matching tasks. Domain-Specific and Task-Specific PTMs (DSTS PTMs) are trained on some specialty corpora. For example, SciBert is trained on scientific text, BioBert is trained on biomedical text.

4 Methodologies

This section is a complement of section one and section two. Prediction of text is the training objects of PTLM. To be specific, Standard Language Model (SLM), Corrupted Text Reconstruction (CTR) and Full Text Reconstruction (FTR). Here are some noising functions used in PTLM. Masking, Replacement, Deletion and Permutation. And another significant factor needed to be considered is the directionality of representations. There are two common methodologies: Left-to-Right and Bidirectional.

5 Applications

The PTLM now can be applied to many different tasks, and they perform in a good direction. Here are some application cases in detail.

5.1 Question Answering (QA)

Three typical problems in QA: SQuAD (Rajpurkar et al., 2016), CoQA (Guu et al., 2018), HotpotQA (Yang et al., 2018). Some PTLM such as Bert and ALBert (Zhang et al., 2020b) could transform extractive QA problems to span prediction problem. And then PTM can acted as the encoder and then predict spans. To solve CoQA tasks, Ju et al. (Ju

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2022 doi:10.20944/preprints202208.0238.v1

et al., 2019) proposed "PTM+TRTKD" model, and for HotpotQA tasks, Tu et al. (Tu et al., 2020) proposed the SAE system.

5.2 Sentiment Analysis

Bert could perform well in some widely used dataset such as aspect-based sentiment analysis (ABSA). Recently a post-training for adapting source domin of Bert was proposed by (Xu et al., 2019) And many people such as (Rietzler et al., 2019) further their work by do analysis of cross-domain post-training' s behavior. And for sentiment transfer, "Mask and Infill" was proposed by (Wu et al., 2019).

5.3 Named Entity Recognition (NER)

NER is an essential topic in information, and many models such as Bert and ELMo (Peters et al., 2018) plays an important role in downstream tasks, so there exist many pre trained models for NER. When process the word embedding, TagLM (Peters et al., 2017) use the last layer output and weighted sum of each layer's output of one pre trained LM to be the part of the word embedding. (Liu et al., 2018) speeded up ELMo's inference on NER by using dense connect and layer-wise pruning techniques.

5.4 Machine Translation (MT)

The MT is an important task in NLP, and most Neural Machine Translation models use encoder-decoder structure. The framework encodes input tokens to hidden representation first and then decodes output tokens. The encoder-decoder model was found by (Ramachandran et al., 2016b) Enlightened by Bert, (Conneau and Lample, 2019) used a pre trained Bert model to initialize encoder and decoder. And a two-stage fine-tuned Bert model for NMT was proposed by (Imamura and Sumita, 2019) Besides just pre training the encoder, (Song et al., 2019) proposed Masked Seq2Seq Pre-training (MSSS), which could use Seq2Seq MLM to jointly train encoder and decoder. And there is also a model called mBART (Liu et al., 2020) proposed by Yinhan Liu et al. in 2020, which could make a great improvement the sentence level and document level on both supervised and unsupervised machine learning translation task.

5.5 Summarization

Summarization could provide a shorter text which could preserve sufficient meaning of a longer text. (Zhong et al., 2019) used Bert to do summarization. (Zhang et al., 2020a) set Gap Sentence Generation for a pre training task. And recently BERTSUM was proposed by (Liu and Lapata, 2019), which included a novel and document-level encoder and the general framework for summarization especially the abstractive summarization and extractive summarization. And there are some summary-level framework and architecture such as MATCHSUM (Zhong et al., 2020) and Siameese-Bert which could compute the similarity between original document and the summary.

5.6 Adversarial Attacks and Defenses (AAD)

Because of discrete feature of languages, adversarial attacks and defenses could be very challenging for text. So, we could use some PTLM to generate adversarial samples. (Li et al., 2020) created a Bert based attacker BERT-Attack. By turning Bert against another fine-tuned Bert model on some downstream tasks, and then successfully misguiding the prediction of target model and outperforming SOTA attack models in both perturb percentage and success rate. And in the future AAD will become more promising, they could make PTLM more robust and generalized for NLP tasks(Li et al., 2020) (Zhu et al., 2019).

There are also some other applications of PTLM, such as general evaluation benchmark, which would be introduced in section six. Then I just list some other common applications and omit the details: Knowledge Probing (like Factual Probing and Linguistic Probing), Text Classification, Natural Language Inference, Information Extraction, "Reasoning" in NLP (like Commonsense Reasoning, Mathematical Reasoning), Text Generation, Automatic Evaluation of Text Generation, Multimodal Learning, Meta-Applications (like Domain Adaptation and Dibiasing).

6 Evaluation

To find a comparable metric to evaluate the performance of PTLM is an important task. And General Language Understanding Evaluation (GLUE) is a bench consisting of nine language understanding tasks, but because some progress such as a new benchmark names SuperGLUE's tasks are more diverse. Some of evaluation for some SOTA PTLM is as follows.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2022 doi:10.20944/preprints202208.0238.v1

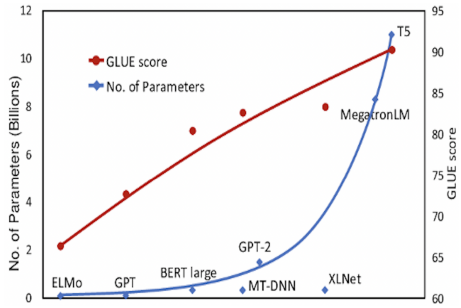

Figure 5: Evaluation of some SOTA PTLM

7 Future Directions

Prompt-based learning attempts to utilize the knowledge got by pre-trained language models, which could solve more downstream tasks. So, in the future the PTMLs could combine with prompt engineering, answer engineering or even multi-prompt learning to generate better models. And I think there are some other tasks PTLM need to solve in the future: Reliability and Interpretability of PTMs, Knowledge Transfer beyond Fine-tuning, Architecture and framework of PTMLs, upper bounds of PTMLs and Model Compression in PTMLs (Qiu et al., 2020).

8 Conclusion

This simple survey gives a rough description of benchmarking, techniques, methodologies, applications, and evaluation in PTLM.

9 Acknowledgements

I thank Professor Yangqiu Song, TA Tianqing Fang and Xin Liu for their guidance.

References

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

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

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

Kenji Imamura and Eiichiro Sumita. 2019. Recycling a pre-trained bert encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31.

Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on conversational question answering. *arXiv preprint arXiv:1909.10772*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, and Richard S Zemel. 2015. Antonio torralba, raquel urtasun, sanja fidler. skip-thought vectors. *Proceedings of Advances in Neural Information Processing Systems*.

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*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. Sensebert: Driving some sense into bert. *arXiv preprint arXiv:1908.05646*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*.

Liyuan Liu, Xiang Ren, Jingbo Shang, Jian Peng, and Jiawei Han. 2018. Efficient contextualized representation: Language model pruning for sequence labeling. *arXiv preprint arXiv:1804.07827*.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 51–61.

ME Peters M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, and Kenton Lee. 2018. Luke 440 zettlemoyer. deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 441.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

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

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Prajit Ramachandran, Peter J Liu, and Quoc V Le. 2016a. Unsupervised pretraining for sequence to sequence learning. *arXiv preprint arXiv:1611.02683*.

Prajit Ramachandran, Peter J Liu, and Quoc V Le. 2016b. Unsupervised pretraining for sequence to sequence learning. *arXiv preprint arXiv:1611.02683*.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. *arXiv preprint arXiv:1908.11860*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Wilson L Taylor. 1957. " cloze" readability scores as indices of individual differences in comprehension and aptitude. *Journal of Applied Psychology*, 41(1):19.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9073–9080.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. " mask and infill": Applying masked language model to sentiment transfer. *arXiv preprint arXiv:1908.08039*.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019a. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. 2021. Pangu-: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020b. Retrospective reader for machine reading comprehension. *arXiv preprint arXiv:2001.09694*, 1:1–9.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.

 doi:10.20944/preprints202208.0238.v1

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. *arXiv preprint arXiv:1907.03491*.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.