the model is pre-trained using three types of language modeling tasks unidirec- tional, bidirectional, and sequence-to-sequence prediction . the model can be ne-tuned for both natural language understanding and generation tasks .

UNILM achieves new state-of-the-art results on ve natural language generation datasets . a shared Transformer network uses specic self-attention masks to control what context the prediction conditions on . the DSTC7 document-grounded dialog response generation NIST-4 to 2.67 human performance is 2.65 .

pre-trained LMs learn contextualized text representations by predicting words based on their context using large amounts of text data . ne-tuned models can be adapted to downstream tasks . language model LM pre-training has advanced the state of the art across a variety of natural language processing tasks 8, 29, 19, 31, 9 .

different prediction tasks and training objectives have been used for pre-training LMs of different types . a forward LM reads the text from left to right, and a backward LM encodes it from right to left .

UNIed pre-trained language model UNILM can be applied to natural language understanding tasks 44 . a bidirectional Transformer encoder fuses both the left and right context to predict the masked words . in contrast, BERT 9 employs a difcult encoder to fuse both the right and left context .

the unied LM is jointly pre-trained by multiple language modeling objectives, sharing the same parameters . we ne-tune and evaluate the pretrained LM on various datasets, including both language understanding and generation tasks .

cloze tasks differ in how the context is dened . for a left-to-right unidirectional LM, the context

consists of all the words on its left . masked word to be predicted is the context of the word on both the right and the right .

the context of the to-be-predicted word in the second target sequence consists of all the words in the rst source sequence . the pre-trained UNILM can be ne-tuned with additional task-specic layers if necessary to adapt to various downstream tasks .

the proposed UNILM has three main advantages: the unied pre-training procedure leads to a single Transformer LM that uses the shared parameters and architecture for different types of LMs . the parameter sharing makes the learned text representations more general because they are jointly optimized for different language modeling objectives .

the pre-training optimizes the shared Transformer 43 network with respect to several unsupervised language modeling objectives . the model parameters are shared across the LM objectives, i.e., bidirectional LM, and sequence-to-sequence LM .

input x is a word sequence, which is either a text segment for unidirectional LMs or a pair of segments packed together . we always add a special start-of-sequence SOS token at the beginning of input . the input representation follows that of BERT 9.

input tokens are tokenized to subword units by WordPiece 48 . UNILM is trained using multiple LM tasks . segment embeddings also play a role of LM identier in that we use different segments for different LM objectives .

2.3 Pre-training Objectives We pretrain UNILM using four cloze tasks . we randomly choose some WordPiece tokens in the input, and replace them with special token MASK . the tokens are all 0s, indicating that all tokens have access to each other.

the parameters of UNILM are learned to minimize the cross-entropy loss computed using the predicted tokens and the original tokens . the use of cloze tasks makes it possible to use the same training procedure for all LMs, unidirectional and bidirectional alike.

the representation of each token encodes only the leftward context tokens and itself . to predict the masked token of x1x2 MASK x4 only tokens x1, x2 and itself can be used .

a right-to-left LM predicts a token conditioned on its future right context . the upper triangular part of the self-attention mask is set to . and the other elements to 0, as shown in Figure 1 .

the self- attention mask M encodes contextual information from both directions . every token is allowed to attend across all positions in the input sequence . the tokens in the rst source segment can only attend to the leftward context in the target segment and itself .

t1 and t2 have access to the rst four tokens, including SOS and EOS . t4 can only attend to the six tokens . the upper right part is set to block attentions from the source segment to the target segment .

the model is learned to recover the masked tokens . the sequence-to-sequence LM pre-trains a bidirectional encoder and an unidirectional decoder . we also include the next sentence prediction task for pre-training .

gelu activation 18 is used as GPT 31 . we use a 24-layer Transformer with 1, 024 hidden size, and 16 attention heads . the weight matrix of the softmax classier is tied with token embeddings .

UNILM is initialized by BERTLARGE, and then pre-trained using English Wikipedia2 and

BookCorpus 53 . the vocabulary size is 28, 996, the maximum length of input sequence is 512 . masked positions 80 of the time we replace the token with MASK, 10 of the times with a random token, and keeping the original token for the rest .

Adam 22 with 1 0.9, 2 0.999 is used for optimization . the learning rate is 3e-5, with linear warmup over the rst 40, 000 steps and linear decay . dropout rate is 0.1 .

weight decay is 0.01. The batch size is 330 . pre-training procedure runs for about 770, 000 steps . the weight decay procedure is 0.01 and is expected to last a year . if you are a beginner, please contact us for more information .

it takes about 7 hours for 10, 000 steps using 8 Nvidia Telsa V100 32GB GPU cards with mixed precision training . we ne-tune UNILM as a bidirectional Transformer encoder, like BERT .

the ne-tuning procedure is similar to pre-training using the self-attention masks . let S1 and S2 denote source and target sequences, respectively . we take the sequence-to-sequence task as an example .

the model is ne-tuned by masking some percentage of tokens in the target sequence at random . the training objective is to maximize the likelihood of masked tokens given context . we pack them together with special tokens, to form the input SOS .

model learns when to emit EOS to terminate the generation process of the target sequence . 3.1 Abstractive Summarization Automatic text summarization produces a concise and uent summary conveying key information in the input e.g., a news article .

RG-1 RG-2 RG-L extractive summarization LEAD-3 40.42 17.62 36.67 Best Extractive 27 43.25

20.24 39.63 Abstractive Summarization PGNet 37 39.53 17.28 37.98 Bottom-Up 16 41.22 18.68 38.34 S2S-ELMo 13 41.56 18.94 38.47 UNILM 43.33 20.21 40.51 Table 4 Results of OpenNMT and Transformer are taken from 4, 39 .

the summary is not constrained to reusing the phrases or sentences in the input text . we use the non-anonymized version of the CNNDailyMail dataset 37 and Gigaword 36 for model ne-tune and evaluation .

the masking probability is 0.7 . we also use label smoothing 40 with rate of 0.1 . for CNNDailyMail, we set batch size to 32, and maximum length to 768 . the masks probability is 0.0 .

the input document is truncated to the rst 640 and 192 tokens for CNNDailyMail and Gigaword respectively . we remove duplicated trigrams in beam search, and tweak the maximum summary length on the development set 28, 13 .

bottom-Up 16 is a sequence-to-sequence model augmented with pre-trained ELMo representations . we also include in Table 3 the best reported extractive summarization result 27 on the dataset .

Transformer 43 and OpenNMT 23 implement standard attentional sequence-to-sequence models . Gigaword has different scales 10K and 3.8M . the model is pre-trained based on Transformer networks .

UNILM's model outperforms MASS by 7.08 point in ROUGE-L . 3.2 Question Answering QA The task is to answer a question given a passage 33, 34, 15 .

the rst is called extractive QA, where the answer is assumed to be a text span in the passage . the other is called generative QA . we ne-tune the pre-trained UNILM as a EM .

we conduct experiments on the Stanford Question Answering Dataset SQuAD 2.0 34 . the results are reported in Table 5, where we compare two models in Exact Match EM and F1 score . a bidirectional encoder is used to encode the task .

RMRELMo 20 is a cased model, ne-tuned on the SQuAD training data for 3 epochs, with batch size 24 and maximum length 384 . UNILM is the same model as BERTLARGE .

CoQA is a conversational question answering dataset . UNILM outperforms BERTLARGE . the answers in CoQA can be free-form texts, including a large portion is of yesno answers .

the results on CoQA are reported in Table 6 . we select a passage subspan with the highest F1 score . UNILM is ne-tuned with the same hyperparameters as BERTLARGE .

extractive methods can only predict subspans of the input passage as answers . UNILM outperforms BERTLARGE . we adapt generative question answering as a sequence-to-sequence model . the input sequence is the concatenation of conversational histories .

we ne-tune the pre-trained UNILM on the CoQA training set for 10 epochs . we set the batch size to 32, the mask probability to 0.5, and the maximum length to 512 .

the maximum length of input question and passage is 470 . we split the passage into several chunks with a sliding window approach . the seq2Seq baseline is a sequence-to-sequence model with an attention mech anism .

the PGNet model augments Seq2Seq with a copy mechanism . the generative question answering model outperforms previous generative methods by a wide margin . 3.3 Question Generation We

conduct experiments for the answer-aware question generation task 52 .

our goal is to generate a question that asks for the answer . the SQuAD 1.1 dataset 33 is used for evaluation . following 12, we split the original training set into training and BLEU-4 MTR RG-L CorefNQG 11 15.16 19.12 - SemQG 50 18.37 22.65 46.68 UNILM 22.12 25.06 51.07 MP-GSN 51 16.38 20.25 44.48 SemqG 50 20.76 24.20 48

MTR is short for METEOR, and RG for ROUGE . UNILM Generated Questions 84.7 87.6 Table 9 Question generation improves results on the SQuAD development set . NIST-4 BLEU-4 Div-1 Div-2 Avg len Best System in DSTC7 Shared Task 2.523 1.83 8.07 9.030 0.109 0.325 15.133 UNIL 2.669 4.39 8.27 9.195 0.120 0.391 14.807 Human Performance 2.

the question generation task is formulated as a sequence-to-sequence problem . we also conduct experiments following the data split as in 51 . a reversed dev-test split is used to generate a question .

ne-tune UNILM on the training set for 10 epochs . we set batch size to 32, masking probability to 0.7, and learning rate to 2e-5 . the other hyper-parameters are the same as pre-training .

corefNQG 11 is based on a sequence-to-sequence model with attention and a feature-rich encoder . MP-GSN 51 uses a gated self-attention encoder and semQG 50 uses two semantics-enhanced rewards . UNILM outperforms previous models and achieves a new state-of-the-art for question generation .

the model is ne-tuned on the SQuAD 2.0 data for two more epochs . the augmented data generated by UNILM improves question answering model introduced in section 3.2 . we use bidirectional masked language modeling as an auxiliary task .

the auxiliary task alleviates catastrophic forgetting 49 when ne-tuning on augmented data . 3.4 Response Generation We evaluate UNILM on the document-grounded dialog response generation task 30, 15. Given a multi-turn conversation history and a web document as the knowledge source, the system needs to 3Notice that if we directly use the tokenized references provided by Du et al. 2017, the results are 21.63 BLEU-4 25.04 METEOR 51.09 ROUGE-

ne-tune UNILM to the task as a sequence-to-sequence model . generate a natural language response that is both conversationally appropriate and reective of the contents of the web document .

the masking probability is set to 0.5 . the maximum length is 512 . we use beam search with size of 10 to decode the beams . if you are using beam search, you will be able to use the beam search tool .

UNILM outperforms the best system 41 in the DSTC7 shared task 14 . the maximum length of generated response is set to 40 . 3.5 GLUE Benchmark We evaluate the unified language understanding evaluation on the general language understanding .

GLUE is a collection of nine language understanding tasks . we use Adamax 21 as our optimizer with a learning rate of 5e-5 and a batch size of 32 . the maximum number of epochs is set to 5 .

table 11 presents the GLUE test results obtained from the benchmark evaluation server . results show that UNILM obtains comparable performance on the tasks in comparison with BERTLARGE . the model is jointly optimized for several LM objectives with shared parameters .

we will train more epochs and larger models on web- scale text corpora . we will also conduct more

experiments on end applications as well as ablation experiments to investigate the model capability and the benets of pre-training multiple language modeling tasks with the same network .

arXiv preprints arxiv1903.07785, 2019. 2 Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro and Danilo Giampiccolo . the second PASCAL recognising textual entailment challenge .

3 Luisa Bentivogli, Ido dagan, Hoa Trang Dang, Danilo Giampiccolo and Bernardo Magnini . the fth pascal recognizing textual entailment challenge .

ziqiang cao, wenjie Li, Sujian Li, and Furu Wei . rerank and rewrite Soft template based neural summarization . semeval-2017 task 1 Semantic textual similarity-multilingual and cross-lingual focused evaluation .

arXiv preprint arxiv1708.00055, 2017. 6 Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang.

the pascal recognising textual entail-ment challenge was presented in Proceedings of the first international conference on machine learning challenges . MLCW05, pages 177190, Berlin, Heidelberg, 2006; 8 Andrew M Dai and Quoc V Le.

curran Associates, Inc., 2015. 9 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova . BERT pre-training of deep bidirectional transformers for language understanding .

automatically constructing a corpus of sentential para- phrases . in Proceedings of the third international workshop on paraphrasing IWP2005, 2005 . a number of para-phrases are sententially sentential .

Xinya Du, Junru Shao, and Claire Cardie learn to ask Neural question generation for reading comprehension . the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1 Long Papers, pages 13421352, 2017.

pre-trained language model representations for language generation . 14 Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao and Bill Dolan . in AAAI Dialog System Technology Challenges Workshop, 2019.

founda-tions and Trends in information retrieval, 132-3127298, 2019. 16 Sebastian Gehrmann, Yuntian Deng, and Alexander Rush . the conference on Empirical Methods in natural language processing, pages 40984109, Brussels, Belgium, October-November 2018.

the third PASCAL recognizing textual entailment challenge . 17 Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan . the aCL-PASCAL Workshop on Textual Entailment and Paraphrasing, pages 19, Prague, June 2007.

arXiv preprints arxiv1606.08415, 2016. 19 Jeremy Howard and Sebastian Ruder. ne-tuning for text classi- cation . 20 Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Ming Zhou .

diederik Kingma and Jimmy Ba. Adam A method for stochastic optimization . read verify Machine reading comprehension with unanswerable questions . arXiv preprint arXIv1412.6980, 2014 .

openNMT Open-source toolkit for neural machine translation . the toolkit was developed by the u.s. university of california . it is a toolkit that can be used to interpret neural neural machines .

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao . the winograd schema challenge is a package for automatic evaluation of summaries .

multi-task deep neural networks for natural language understanding . a deep reinforced model for abstractive summarization . 28 Romain Paulus, Caiming Xiong, and Richard Socher . the model is based on a deeper reinforced model .

the 2018 conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies . 30 Lianhui Qin, Michel Galley, Chris Brockett, xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao .

the Association for Computational Linguistics, 57th Annual Meeting, Florence, Italy, July 2019 . generative pre-training aims to improve language understanding by generating neural conversation with on-demand machine reading . the association for computational linguistics is presenting its findings in a new book .

language models are unsupervised multitask learners . 33 Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang . SQuAD 100,000 ques- tions for machine comprehension of text .

the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2 Short Papers, pages 784789, 2018. 35 Siva Reddy, Danqi Chen, and Christopher D. Manning.

a neural attention model for abstractive sentence summarization . in Proceedings of the 2015 conference on Empirical Methods in Natural Language Processing, pages 379389, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

the 55th Annual Meeting of the Association for Computational Linguistics, pages 10731083,

Vancouver, Canada, July 2017 . 38 Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng.

in Proceedings of the 2013 conference on empirical methods in natural language processing, pages 16311642, 2013. 39 Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu.

arXiv preprint arxiv1905.02450, 2019. 40 Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna .

Y Tam, Jiachen Ding, Cheng Niu, and Jie Zhou . cluster-based beam search for pointer-generator chatbot grounded by knowledge . y-taylor, jiachen ding, niu et al.

AAAI Dialog System Technology Challenges Workshop, 2019. 42 Wilson L Taylor . Cloze procedure A new tool for measuring readability. journalism Bulletin, 304415433, 1953. 43 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin .

GLUE A multi-task benchmark and analysis platform for natural language understanding . 45 Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman . arXiv preprint arxiv1805.12471, 2018.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao . Xiaobing Liu, Lukasz Kaiser, Stephan Gouws .

arXiv preprint arxiv1901.11373, 2019. 50 Shiyue Zhang and Mohit Bansal . addressing semantic drift in question generation for semi- supervised question answering for question answering. .

paragraph-level neural question generation with maxout pointer and gated self-attention networks . 52 Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao and Ming Zhou .

Xuanjing Huang, Jing Jiang, Dongyan Zhao, Yansong Feng, and Yu Hong, editors, Natural Language Processing and Chinese Computing, pages 662671 . a preliminary study was conducted in a chinese language .

the model can generate text samples using the left-to-right setting . we picked three text samples sampled from left to right using our model, as shown in Table 12 . the model could be used to generate a text sample .

model output is like a short paragraph in a fantasy novel for the other two with input from Wikipedia . the model makes up facts with plausible names, time stamps and events written in the Wikipedia style . we use the top-40 truncating sampling strategy 32 and forbid duplicate 4-grams during generation .

input Output Winston sat back. a sense of complete helplessness had descended upon him . he was now helpless and unmoving . we summarize the data size and the evaluation metrics used for the GLUE benchmark .

the demon known as Azrael would never know exactly what that was, or even how it would end . he was trapped within her embrace, a trapped, torn, battered, crippling presence he only knew it would mean was eternity .

he is a yellow muppet character on the long running childrens television show, sesame Street . he would typically appear as a sidekick or villain in the show . his voice was provided by Michael Combs .

rst work, le Grand Cours d Auvergne, was composed in 1909 by Maurice Ravel, a student of Jules Massenet . it was re-published in 1912 by the publisher J.S.D.M. de lEtablissement Musicale de la Musique Francaise .