ProphetNet-X: Large-Scale Pre-training Models for English, Chinese, Multi-lingual, Dialog, and Code Generation

Weizhen Qi[1] [*], Yeyun Gong[2] [†], Yu Yan[3], Can Xu[3], Bolun Yao[4], Bartuer Zhou[2]
Biao Cheng[2] , Daxin Jiang[3], Jiusheng Chen[3], Ruofei Zhang[3], Houqiang Li[1], Nan Duan[2]
[1]University of Science and Technology of China, [2]Microsoft Research Asia,
[3]Microsoft, [4] Nanjing University of Science and Technology
[1]weizhen@mail.ustc.edu.com, lihq@ustc.edu.com,
[2]{yegong,bazhou,bicheng,nanduan}@microsoft.com,
[3]{yyua,caxu,djiang,jiuchen,bzhang}@microsoft.com [4]yaobl001@njust.edu.cn

Abstract

Now, the pre-training technique is ubiquitous in natural language processing field. ProphetNet is a pre-training based natural language generation method which shows powerful performance on English text summarization and question generation tasks. In this paper, we extend ProphetNet into other domains and languages, and present the ProphetNet family pre-training models, named ProphetNet-X, where X can be English, Chinese, Multi-lingual, and so on. We pre-train a cross-lingual generation model ProphetNet-Multi, a Chinese generation model ProphetNet-Zh, two open-domain dialog generation models ProphetNet-Dialog-En and ProphetNet-Dialog-Zh. And also, we provide a PLG (Programming Language Generation) model ProphetNet-Code to show the generation performance besides NLG (Natural Language Generation) tasks. In our experiments, ProphetNet-X models achieve new state-of-the-art performance on 10 benchmarks. All the models of ProphetNet-X share the same model structure, which allows users to easily switch between different models. We make the code and models publicly available[1], and we will keep updating more pre-training models and finetuning scripts.

1 Introduction

In recent years, quite a few natural language generation pre-training models are proposed (Qi et al., 2020; Lewis et al., 2019; Song et al., 2019; Brown et al., 2020). Downstream generation tasks benefit from these large scale pre-training models greatly in fluency and accuracy. Researchers also extend these general pre-training works into specific domains such as DialoGPT (Zhang et al., 2019) is extended from GPT (Brown et al., 2020) for dialog system, mBART (Liu et al., 2020b) is extended from BART (Lewis et al., 2019) for multi-lingual generation, CodeBERT (Feng et al., 2020) is extended from BERT (Devlin et al., 2018) for programming language modeling, etc.

Although there are pre-trained models for some specific domains, it is not convenient for users to find them and set them up. Besides, even some models in the same pre-training family with the same model structure and pre-training tasks, their codes and details vary a lot because of different implementation and backends selection.

ProphetNet (Qi et al., 2020) is firstly proposed as an English text pre-training model with future tokens' prediction, and successfully improves the performance on different downstream NLG tasks. In this work, we pre-train the ProphetNet on different corpus, respectively. The corpus covers different languages and domains. All the pre-trained models share the same model structure with different vocabularies. We provide six pre-trained models with downstream task finetuning scripts, including ProphetNet-En pre-trained with 160GB English raw text, ProphetNet-Zh pre-trained with 160GB Chinese raw text, ProphetNet-Multi with 101GB Wiki-100 corpus and 1.5TB Common Crawl[2] data, ProphetNet-Dialog-En with 60 million sessions Reddit open-domain dialog corpus, ProphetNet-Dialog-Zh with collected Chinese dialog corpus over 30 million sessions, and ProphetNet-Code pre-trained with 10 million codes and documents. ProphetNet-X achieves new state-of-the-art results on 10 benchmarks, including Chinese summarization (MATINF-SUMM (Xu et al., 2020a) and LCSTS (Hu et al., 2015)), Chinese question answering (MATINF-QA (Xu et al., 2020a)), cross-lingual generation (XGLUE NTG (Liang et al., 2020) and

[*] Work is done during internship at Microsoft Research Asia.

[†] Corresponding Author.

[1] https://github.com/microsoft/ProphetNet

[2] https://commoncrawl.org/

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 232–239, August 1st - August 6th, 2021.

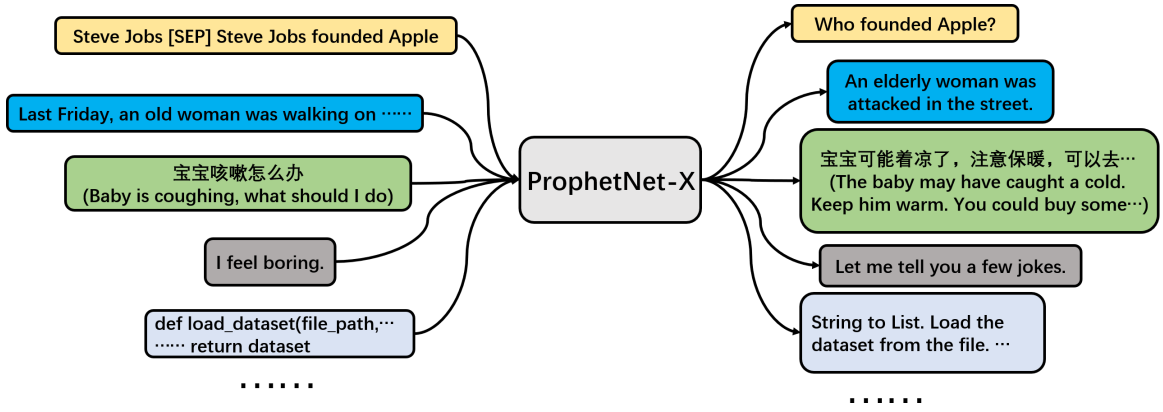

Figure 1: A diagram of ProphetNet-X framework. ProphetNet-X models share the same model structure and cover various languages and domains.

XGLUE QG (Liang et al., 2020)), English summarization (MSNews (Liu et al., 2020a)), English dialog generation (DailyDialog (Li et al., 2017), PersonaChat (Zhang et al., 2018), and DSTC7-AVSD (Alamri et al., 2019)), and code summarization (CodeXGLUE (Lu et al., 2021)). Users can simply download the ProphetNet-X repository and find corresponding pre-trained model with downstream task finetuning scripts.

The main contributions of ProphetNet-X can be described as follows:

- We provide a family of pre-trained models named ProphetNet-X, with six models including English and Chinese natural language generation in open-domain and dialog, multilingual generation, and code generation.
- All the pre-trained ProphetNet-X models share the same model structure. Users only need to simply modify one model file to use it in different language or domain tasks.
- We conduct extensive experiments, the results show that ProphetNet-X models achieve new state-of-the-art performance on 10 publicly available benchmarks.

2 ProphetNet-X

2.1 Architecture

We train different ProphetNet-X models based on ProphetNet. ProphetNet is an encoder-decoder natural language generation model with future n-gram prediction. ProphetNet leverages stacked Transformer encoder layers and stacked multi-stream self-attention Transformer decoder layers. ProphetNet aims to prevent overfitting on strong local correlations such as 2-gram combinations, and deploy future tokens' prediction to enhance auto-regressive generation ability.

Given the input sequence $x = (x_1, \ldots, x_M)$ and output sequence $y = (y_1, \ldots, y_T)$, n-gram ProphetNet-X replaces the auto-regressive predicting dependency relationship $p(y_t|y_{<t}, x)$ with $p(y_{t:t+n-1}|y_{<t}, x)$. Firstly, ProphetNet-X gets the encoded hidden states with stacked Transformer encoder layers $H_{\text{enc}} = \textbf{Encoder}(x_1, \ldots, x_M)$. Then, decoder with n-stream self-attention predicts next n tokens at each time step, as: $p(y_t|y_{<t}, x), \ldots, p(y_{t+n-1}|y_{<t}, x) = \textbf{Decoder}(y_{<t}, H_{\text{enc}})$. The optimization target of ProphetNet-X can be described as:

$$\begin{aligned}\mathcal{L} &= -\sum_{j=0}^{n-1} \alpha_j \cdot \left(\sum_{t=1}^{T-j} \log p_\theta(y_{t+j}|y_{<t}, x) \right) \\ &= \underbrace{-\alpha_0 \cdot \left(\sum_{t=1}^{T} \log p_\theta(y_t|y_{<t}, x) \right)}_{\text{language modeling loss}} \\ &\quad \underbrace{-\sum_{j=1}^{n-1} \alpha_j \cdot \left(\sum_{t=1}^{T-j} \log p_\theta(y_{t+j}|y_{<t}, x) \right)}_{\text{future n-gram loss}}\end{aligned}$$

The details of ProphetNet and multi-stream self-attention can be found in Qi et al. (2020).

2.2 Pre-training Corpus

In this section, we introduce the pre-training corpus for ProphetNet-X.

language	Fr	It	Es	De	Nl	Pt	En	Sv	Pl	Vi	Ar	Ru	Tr
size(GB)	77.25	74.01	72.97	71.48	71.19	71.05	68.34	67.48	67.44	67.43	65.18	64.09	62.96
language	Ja	Zh	Cs	El	Ko	Ro	Th	Da	Bg	Fi	Hu	No	Hi
size(GB)	61.49	58.70	56.62	55.15	45.28	44.05	35.65	32.43	28.44	27.85	27.04	25.24	17.18
language	Sk	Id	Ca	Uk	Lt	Sr	Sl	Hr	Et	Lv	Ka	Az	Ur
size(GB)	14.78	13.68	13.08	10.80	9.20	8.59	6.86	6.51	6.47	5.48	4.16	3.38	3.13
language	Kk	Ne	Gl	My	Eu	Gu	Si	Ms	Sq	Af	Cy	Sw	Bs
size(GB)	3.09	2.18	1.95	1.83	1.37	1.23	1.20	1.03	1.03	0.93	0.51	0.34	0.15

Table 1: Statistics of our multi-lingual pre-training corpus. The total pre-training corpus size is 1.54 TB. ISO codes are used to represent each language.

For ProphetNet-Zh, we collect Chinese Wikipedia, CLUE (Xu et al., 2020b) and Chinese Common Crawl data to reach 160GB. For traditional Chinese data, we firstly use OpenCC [3] to convert them to simplified Chinese. The pre-training corpus includes common webs, online forums, comments websites, Q&A websites, Chinese Wikipedia, and other encyclopedia websites. We build a simplified Chinese char vocabulary. The char vocabulary size is 9,360.

For ProphetNet-Multi, besides Wiki-100 corpus, we select 52 common languages to collect and clean multi-lingual data from Common Crawl. After cleaning and tokenizing, the Common Crawl corpus size we use is described in Table 1. The ProphetNet-Multi vocabulary is same as XLM-R (Conneau et al., 2019) 250k sentencepiece[4] model.

For ProphetNet-Dialog-En, we utilize Reddit comments dataset (Zhou et al., 2018; Galley et al., 2019). We firstly load the weights of ProphetNet-En then clean 60 million sessions for pre-training.

For ProphetNet-Dialog-Zh, we use the pre-training corpus from Wang et al. (2020) and we crawled 18.2 million dyadic dialogues (conversation between two persons) longer than or equal to 2 turns (one turn denotes one utterance from one person) from the Douban group[5] which is a popular social networking service in China. The pre-training corpus size comparison between Wang et al. (2020) and ProphetNet-Dialog-Zh is shown in Table 2. We also load the pre-trained model from ProphetNet-Zh before pre-training, which already contains external knowledge from open-domain Chinese corpus.

For ProphetNet-Code, we conduct pre-training on both PLs (Programming Languages) and their describing NL (Natural Language). We use the pre-training corpus provided by CodeSearchNet (Husain et al., 2019). It covers 6 programming languages, including Go, Java, Javascript, PHP, Python, and Ruby. We employ the same sentencepiece tokenizer as CodeBERT (Feng et al., 2020). The tokenizer is used for both PL and NL, with a vocabulary size 50,365.

Corpus Size	Single-turn	Multi-turn
LCCC-base	3,354,382	3,466,607
LCCC-large	7,273,804	4,733,955
ProphetNet-Dialog-Zh	23,309,502	6,985,425

Table 2: Statistics of Chinese Dialog pre-training corpus

For ProphetNet-En, we directly take the model pre-trained in ProphetNet (Qi et al., 2020). It is pre-trained with 160GB English raw texts, including Wikipedia, books, stories, news, and web texts. The vocabulary of ProphetNet-En is same as BERT subwords vocabulary. The vocabulary is based on bpe subwords with a max length matching algorithm. Its vocabulary size is 30,522.

3 Experiments

3.1 Pre-training Settings

We carry out pre-training with 12-layer encoder, 12-layer decoder ProphetNet models. The hidden size is 1,024, feed forward size is 4,096, future tokens' prediction length is 2. Both the max sequence lengths of the input and output are set to 512.

For ProphetNet-En, ProphetNet-Zh, ProphetNet-Multi, ProphetNet-Dialog-En, and ProphetNet-Code, we carry out un-supervised pre-training with masked span prediction task. Spans of continuous tokens are masked out from the encoder input sentences and predicted from the decoder side. We masked continuous 9 tokens in every 64 tokens from the encoder side, and predict the 9 tokens on the decoder side. In other words, for maximum 512 encoder sequence length, totally $8(spans) \times 9(tokens\ per\ span) = 72$ tokens

[3] https://github.com/BYVoid/OpenCC
[4] https://github.com/google/sentencepiece
[5] https://www.douban.com/group

Method	MATINF-QA			MATINF-SUMM			LCSTS		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
TextRank (Mihalcea and Tarau, 2004)	-	-	-	35.53	25.78	36.84	24.38	11.97	16.76
LexRank (Erkan and Radev, 2004)	-	-	-	33.08	23.31	34.96	22.15	10.14	14.65
Seq2Seq (Sutskever et al., 2014)	16.62	4.53	10.37	23.05	11.44	19.55	-	-	-
Seq2Seq+Att (Luong et al., 2015)	19.62	5.87	13.34	43.05	28.03	38.58	33.80	23.10	32.50
WEAN (Ma et al., 2018)	-	-	-	34.63	22.56	28.92	37.80	25.60	35.20
Global Encoding (Lin et al., 2018)	-	-	-	49.28	34.14	47.64	39.40	26.90	36.50
BertAbs (Liu and Lapata, 2019)	-	-	-	57.31	44.05	**55.93**	-	-	-
MTF-S2S$_{single}$ (Xu et al., 2020a)	20.28	5.94	13.52	43.02	28.05	38.55	33.75	23.20	32.51
MTF-S2S$_{multi}$ (Xu et al., 2020a)	21.66	6.58	14.26	48.59	35.69	43.28	-	-	-
ProphetNet-Zh	**24.18**	**6.38**	**15.47**	**58.82**	**44.96**	54.26	**42.32**	**27.33**	**37.08**

Table 3: Results of ProphetNet-Zh on MATINF-QA, MATINF-SUMM, and LCSTS. "R-1", "R-2", and "R-L" represent "ROUGE-1", "ROUGE-2", and "ROUGE-L", respectively.

Task	Model	De	En	Es	Fr	It	Pt	Ru	AVG
QG	M-BERT (Devlin et al., 2018)	0.1	7.8	0.1	0.1	0.2	0.1	-	1.4
	XLM-R$_{base}$ (Conneau et al., 2019)	0.1	6.0	0.0	0.0	0.1	0.0	-	1.0
	Unicoder$_{DAE}$ (Liang et al., 2020)	3.0	14.0	12.4	4.2	15.8	8.3	-	9.6
	Unicoder$_{FNP}$ (Liang et al., 2020)	3.7	13.9	14.8	4.9	17.0	9.5	-	10.6
	ProphetNet-Multi	**4.9**	**14.9**	**17.0**	**6.0**	**19.2**	**11.3**	-	**12.2**
NTG	M-BERT (Devlin et al., 2018)	0.7	9.0	0.4	0.4	-	-	0.0	2.1
	XLM-R$_{base}$ (Conneau et al., 2019)	0.6	8.1	0.4	0.3	-	-	0.0	1.9
	Unicoder$_{DAE}$ (Liang et al., 2020)	6.8	15.6	9.0	8.7	-	-	7.7	9.6
	Unicoder$_{FNP}$ (Liang et al., 2020)	7.5	15.8	11.9	9.9	-	-	8.4	10.7
	ProphetNet-Multi	**8.7**	**16.7**	**12.7**	**11.4**	-	-	**8.5**	**11.6**

Table 4: Results of ProphetNet-Multi on XGLUE zero-shot cross-lingual generation task. Task QG and NTG represent Question Generation and News Title Generation. Numbers in this table are BLEU-4 scores.

are masked and predicted. If the last part does not reach a maximum length of 64, 15% continuous tokens are masked. ProphetNet-Dialog-En has special tokens [X_SEP] to separate turns in a session and [SEP] to separate different sessions. For ProphetNet-Dialog-Zh, we conduct supervised pre-training. Previous turns of dialogs are fed into the encoder, and the response is predicted from the decoder. It means that for a multi-turn session with n sentences, $n-1$ samples are created for pre-training. The pre-trained ProphetNet-Dialog-Zh can be used to directly generate dialogs without finetuning.

We carry out pre-training on NVIDIA Tesla V100 GPUs, and the total cost exceeds 30,000 GPU hours.

3.2 Finetuning Benchmarks

For different ProphetNet-X models, we select different benchmarks to evaluate them, respectively.

For ProphetNet-Zh, we evaluate our pre-trained model with MATINF-QA (Xu et al., 2020a) for generative question answering task, MATINF-SUMM (Xu et al., 2020a) and LCSTS (Hu et al., 2015) for summarization task.

For ProphetNet-Multi, we follow Unicoder$_{FNP}$ to evaluate on XGLUE (Liang et al., 2020) for cross-lingual zero-shot generation tasks. The pre-trained multi-lingual model is finetuned with English supervised data and inference with English and other un-seen languages data. There are NTG (News Title Generation) and QG (Question Generation) tasks.

For ProphetNet-Dialog-En, we carry out finetuning on DailyDialog (Li et al., 2017) for chit-chat generation, Persona-Chat (Zhang et al., 2018) for knowledge grounded conversation generation and DSTC7-AVSD (Alamri et al., 2019) for conversational question answering.

For ProphetNet-Dialog-Zh, we use the STC (Shang et al., 2015) single-turn open-domain dialog dataset cleaned by Wang et al. (2020), and real-world Xiaoice Chinese dialog dataset for evaluation.

For ProphetNet-Code, we evaluate the performance on code summarization task from CodeXGLUE (Lu et al., 2021).

For ProphetNet-En, we reports the results on summarization tasks CNN/DM (Hermann et al., 2015), Gigaword (Rush et al., 2015), and MSNews (Liu et al., 2020a); question generation tasks SQuAD 1.1 (Rajpurkar et al., 2016) and MSQG (Liu et al., 2020a).

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
AVSD Baseline (Alamri et al., 2019)	0.629	0485	0.383	0.309	0.215	0.487	0.746
CMU Sinbad's (Sanabria et al., 2019)	0.718	0.584	0.478	0.394	0.267	0.563	1.094
PLATO (Bao et al., 2020)	0.784	0.637	0.525	0.435	0.286	0.596	1.209
ProphetNet-Dialog-En	**0.823**	**0.688**	**0.578**	**0.482**	**0.309**	**0.631**	**1.354**

Table 5: Results of ProphetNet-Dialog-En on DSTC7-AVSD.

Model	DailyDialog					PersonaChat				
	B-1	B-2	D-1	D-2	AVG	B-1	B-2	D-1	D-2	AVG
Seq2Seq (Vinyals and Le, 2015)	0.336	0.238	0.03	0.128	0.183	0.448	0.353	0.004	0.016	0.205
iVAE_MI (Fang et al., 2019)	0.309	0.249	0.029	0.25	0.209	-	-	-	-	-
LIC (Golovanov et al., 2019)	-	-	-	-	-	0.405	0.320	0.019	0.113	0.214
PLATO w/o latent (Bao et al., 2020)	0.405	0.322	0.046	0.246	0.255	0.458	0.357	0.012	0.064	0.223
PLATO (Bao et al., 2020)	0.397	0.311	**0.053**	**0.291**	0.263	0.406	0.315	**0.021**	**0.121**	0.216
ProphetNet-Dialog-En	**0.461**	**0.402**	0.038	0.208	**0.277**	**0.459**	**0.382**	0.010	0.060	**0.228**

Table 6: Results of ProphetNet-Dialog-En on DailyDialog and PersonaChat. "B-1", "B-2", "D-1" and "D-2" represent "BLEU-1", "BLEU-2", "Distinct-1" and " Distinct-2", respectively.

3.3 Results

For ProphetNet-Zh, we see significant improvements in Table 3. TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2004) are extractive baselines and others are abstractive baselines. MTF-S2S$_{single}$ (Xu et al., 2020a) and MTF-S2S$_{multi}$ denote single task finetuning and multi-task finetuning on MATINF dataset. We see consistent gains on both Chinese question answering task and summarization tasks.

For ProphetNet-Multi, we show the results in Table 4, Unicoder$_{DAE}$ and Unicoder$_{FNP}$ are pre-trained on Wiki-100 with denoising auto encoder task and ProphetNet, respectively. Comparing the results between the Unicoder$_{FNP}$ and ProphetNet-Multi, we see that more pre-training corpus improves supervised English inference results and other zero-shot languages inference performance. And compared with other baseline methods, ProphetNet-Multi achieves new state-of-the-art results on both NTG and QG tasks.

For English open-domain dialog generation, we show the results in Table 5 and Table 6, compared with strong new proposed PLATO (Bao et al., 2020), we see that ProphetNet-Dialog achieves performance improvements.

Results for ProphetNet-Dialog-Zh on STC can be seen in Table 7. In addition, Table 8 shows the results on real-world Xiaoice dialog dataset with human evaluation. Results in Table 7 hint that for dialog generation, the auto-evaluation metrics (BLEU-2 and BLEU-4) may fail because open-domain dialog outputs could be very different from the given golden targets but still good responses. We observe that ProphetNet-Dialog-Zh without finetuning can generate fluent and meaningful responses but have lower BLEU scores because of the writing style difference. Thus, we conduct a human evaluation as in (Zhao et al., 2020). We randomly collect 500 single-turn and 500 multi-turn context-response pairs from the online logs of the real-word dialog system Xiaoice. Then, we recruit 3 native speakers as human annotators. The annotators have to judge which response is better, based on informativeness, consistency, and fluency of the responses. If an annotator cannot tell which response is better, he/she is required to label a "Tie". With the

Models	B-2	B-4
Seq2Seq-Attn (Luong et al., 2015)	3.93	0.9
Transformer (Vaswani et al., 2017)	6.72	3.14
GPT$_{Novel}$ (Wang et al., 2020)	5.96	2.71
CDialGPT$_{LCCC-base}$ (Wang et al., 2020)	6.48	3.08
CDialGPT2$_{LCCC-base}$ (Wang et al., 2020)	5.69	2.50
CDialGPT$_{LCCC-large}$ (Wang et al., 2020)	6.63	**3.20**
ProphetNet-Dialog-Zh w/o finetuning	2.54	0.75
ProphetNet-Dialog-Zh w/ finetuning	**6.78**	3.05

Table 7: Results of ProphetNet-Dialog-Zh on STC dataset. "B-2", and "B-4" represent "BLEU-2" and "BLEU-4", respectively.

Setting	Win	Lose	Tie	Kappa
Ours-C vs Xiaoice-C	68%	26%	6%	0.73
Ours-C vs Xiaoice-S	76%	24%	0%	0.65
Ours-S vs Xiaoice-S	81%	19%	0%	0.67

Table 8: Human evaluated results for ProphetNet-Dialog-Zh on real-world Xiaoice dataset. Here, Ours means ProphetNet-Dialog-Zh, Xiaoice means old Xiaoice retrieval based dialog system. -S(single-turn) denotes only the last turn is fed to our model or Xiaoice traditional single-turn retrieval model. -C(context) denotes feeding dialog history into our model or Xiaoice traditional multi-turn retrieval model.

Models	Ruby	Javascript	Go	Python	Java	PHP	overall
Seq2Seq (Vinyals and Le, 2015)	9.64	10.21	13.98	15.93	15.09	21.08	14.32
Transformer (Vaswani et al., 2017)	11.18	11.59	16.38	15.81	16.26	22.12	15.56
RoBERTa (Liu et al., 2019)	11.17	11.90	17.72	18.14	16.47	24.02	16.57
CodeBERT (Feng et al., 2020)	12.16	14.90	18.07	19.06	17.65	25.16	17.83
PLBART (Ahmad et al., 2021)	14.11	15.56	**18.91**	**19.30**	18.45	23.58	18.32
Prophetnet-Code	**14.37**	**16.60**	18.43	17.87	**19.39**	**24.57**	**18.54**

Table 9: Results of ProphetNet-Code on CodeXGLUE for code-to-text summarization task. Numbers in this table are smoothed BLEU-4 scores.

Method	CNN/DM			Gigaword			MSNews		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
LSTM (Bahdanau et al., 2014)	37.3	15.7	34.4	33.6	15.4	31.2	30.0	14.6	27.7
Transformer (Vaswani et al., 2017)	39.5	16.7	36.7	36.4	17.7	33.8	33.0	15.4	30.0
MASS (Song et al., 2019)	42.9	19.8	39.8	38.9	20.2	36.2	40.4	21.5	36.8
BART (Lewis et al., 2019)	44.1	**21.2**	40.9	37.5	17.6	34.3	43.8	24.0	39.2
ProphetNet-En	**44.2**	21.1	**41.3**	**39.5**	**20.4**	**36.6**	**44.1**	**24.4**	**40.2**

Table 10: Results of ProphetNet-En for text summarization. "R-1", "R-2", and "R-L" represent "ROUGE-1", "ROUGE-2", and "ROUGE-L", respectively.

experts' annotation, we see that ProphetNet-Dialog-Zh obviously outperforms Xiaoice retrieval based old system. Kappa (Fleiss and Cohen, 1973) values of all models exceed 0.6, indicating substantial agreement overall annotators.

For ProphetNet-Code, the code summarization results are shown in Table 9. We can see new state-of-the-art results are obtained with ProphetNet-Code. It shows that ProphetNet-X models not only benefit from pre-training on natural language generation tasks but also perform well in programming language tasks.

Model	SQuAD 1.1			MSQG		
	R-L	B-4	MTR	R-L	B-4	MTR
LSTM	27.2	3.8	8.9	25.3	3.5	14.1
Transformer	30.7	4.8	10.9	29.3	5.1	16.6
MASS	49.9	21.3	25.2	**38.9**	9.5	23.5
BART	50.3	22.0	**26.4**	38.8	9.2	**24.3**
ProphetNet-En	**51.5**	**22.5**	26.0	38.3	**9.6**	23.3

Table 11: Results of ProphetNet-En for question generation on SQuAD1.1 and MSQG. "R-L", "B-4", and "MTR" represent "ROUGE-L", "BLEU-4", and "METEOR", respectively.
.

For ProphetNet-En, we report the results for ProphetNet in Table 10 and Table 11. We also report the results for two new tasks MSNTG and MSQG introduced from GLGE (Liu et al., 2020a).

4 Related Work

ProphetNet (Qi et al., 2020) is the most related to our work since we carry out pre-training based on it. Other related works involve pre-training works in different domains. For English generation pre-training, MASS (Song et al., 2019) proposes an unsupervised pre-training task with span masked and recover. BART (Lewis et al., 2019) feeds corrupted sentences into the encoder and reconstructs the original sentences. GPT (Radford et al., 2019) models perform language modeling pre-training with Transformer decoder. For multilingual pre-training, mBART (Liu et al., 2020b) introduces language labels to adopt BART denoising pre-training. Based on GPT (Radford et al., 2019), DialoGPT (Zhang et al., 2019) and CDialGPT (Wang et al., 2020) adopts language model pre-training with English and Chinese dialog corpus respectively. CodeBERT (Feng et al., 2020) and GraphCodeBERT (Guo et al., 2020) are two pre-training models for programming languages. PLBART (Ahmad et al., 2021) is similar to multilingual BART with language tags to perform denoising pre-training on programming languages.

5 Conclusion

In this paper, we pre-train ProphetNet-X on various languages and domains, including open-domain (for English, Chinese, and Multi-lingual), dialog (for English and Chinese), and programming (for Ruby, Javascript, Go, Python, Java, and PHP). All the models share the same model structure and are easy to use. Extensive experiments show that ProphetNet-X achieves new state-of-the-art performance on 10 benchmarks. In the future, we will extend ProphetNet-X to support more domains such as biomedical text and protein pre-training.

References

Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Unified pre-training for program understanding and generation. *arXiv preprint arXiv:2103.06333*.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Le Fang, Chunyuan Li, Jianfeng Gao, Wen Dong, and Changyou Chen. 2019. Implicit deep latent variable models for text generation. 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 3937–3947.

Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, et al. 2020. Codebert: A pre-trained model for programming and natural languages. *arXiv preprint arXiv:2002.08155*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7. In *AAAI Dialog System Technology Challenges Workshop*.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058.

Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Jian Yin, Daxin Jiang, et al. 2020. Graphcodebert: Pre-training code representations with data flow. *arXiv preprint arXiv:2009.08366*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. *arXiv preprint arXiv:1506.05865*.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pretraining, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global encoding for abstractive summarization. *arXiv preprint arXiv:1805.03989*.

Dayiheng Liu, Yu Yan, Yeyun Gong, Weizhen Qi, Hang Zhang, Jian Jiao, Weizhu Chen, Jie Fu, Linjun Shou, Ming Gong, et al. 2020a. Glge: A new

general language generation evaluation benchmark. *arXiv preprint arXiv:2011.11928*.

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. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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*.

Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. 2021. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Shuming Ma, Xu Sun, Wei Li, Sujian Li, Wenjie Li, and Xuancheng Ren. 2018. Query and output: Generating words by querying distributed word representations for paraphrase generation. *arXiv preprint arXiv:1803.01465*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

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-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2401–2410.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*, pages 2383–2392.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389.

Ramon Sanabria, Shruti Palaskar, and Florian Metze. 2019. Cmu sinbad's submission for the dstc7 avsd challenge.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

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*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103. Springer.

Canwen Xu, Jiaxin Pei, Hongtao Wu, Yiyu Liu, and Chenliang Li. 2020a. Matinf: A jointly labeled large-scale dataset for classification, question answering and summarization. *arXiv preprint arXiv:2004.12302*.

Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai Sun, Yechen Xu, et al. 2020b. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Yufan Zhao, Can Xu, Wei Wu, and Lei Yu. 2020. Learning a simple and effective model for multi-turn response generation with auxiliary tasks. *arXiv preprint arXiv:2004.01972*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.