MASS: Masked Sequence to Sequence Pre-training for Language Generation

Kaitao Song [*1] **Xu Tan** [*2] **Tao Qin** [2] **Jianfeng Lu** [1] **Tie-Yan Liu** [2]

Abstract

Pre-training and fine-tuning, e.g., BERT (Devlin et al., 2018), have achieved great success in language understanding by transferring knowledge from rich-resource pre-training task to the low/zero-resource downstream tasks. Inspired by the success of BERT, we propose MAsked Sequence to Sequence pre-training (MASS) for encoder-decoder based language generation. MASS adopts the encoder-decoder framework to reconstruct a sentence fragment given the remaining part of the sentence: its encoder takes a sentence with randomly masked fragment (several consecutive tokens) as input, and its decoder tries to predict this masked fragment. In this way, MASS can jointly train the encoder and decoder to develop the capability of representation extraction and language modeling. By further fine-tuning on a variety of zero/low-resource language generation tasks, including neural machine translation, text summarization and conversational response generation (3 tasks and totally 8 datasets), MASS achieves significant improvements over baselines without pre-training or with other pre-training methods. Specially, we achieve state-of-the-art accuracy (37.5 in terms of BLEU score) on the unsupervised English-French translation, even beating the early attention-based supervised model (Bahdanau et al., 2015b).

1. Introduction

Pre-training and fine-tuning are widely used when target tasks are of low or zero resource in terms of training data, while pre-training has plenty of data (Girshick et al., 2014; Szegedy et al., 2015; Ouyang et al., 2015; Dai & Le, 2015; Howard & Ruder, 2018; Radford et al., 2018; Devlin et al., 2018). For example, in computer vision, models are usually pre-trained on the large scale ImageNet dataset and then fine-tuned on downstream tasks like object detection (Szegedy et al., 2015; Ouyang et al., 2015) or image segmentation (Girshick et al., 2014). Recently, pre-training methods such as ELMo (Peters et al., 2018), OpenAI GPT (Radford et al., 2018) and BERT (Devlin et al., 2018) have attracted a lot of attention in natural language processing, and achieved state-of-the-art accuracy in multiple language understanding tasks such as sentiment classification (Socher et al., 2013), natural language inference (Bowman et al., 2015), named entity recognition (Tjong Kim Sang & De Meulder, 2003) and SQuAD question answering (Rajpurkar et al., 2016), which usually have limited supervised data. Among the pre-training methods mentioned above, BERT is the most prominent one by pre-training the bidirectional encoder representations on a large monolingual corpus through masked language modeling and next sentence prediction.

Different from language understanding, language generation aims to generate natural language sentences conditioned on some inputs, including tasks like neural machine translation (NMT) (Cho et al., 2014; Bahdanau et al., 2015a; Vaswani et al., 2017), text summarization (Ayana et al., 2016; Suzuki & Nagata, 2017; Gehring et al., 2017) and conversational response generation (Shang et al., 2015; Vinyals & Le, 2015). Language generation tasks are usually data-hungry, and many of them are low-resource or even zero-source in terms of training data. Directly applying a BERT like pre-training method on these natural language generation tasks is not feasible, since BERT is designed for language understanding, which are usually handled by just one encoder or decoder. Therefore, how to design pre-training methods for the language generation tasks (which usually adopt the encoder-decoder based sequence to sequence learning framework) is of great potential and importance.

In this paper, inspired by BERT, we propose a novel objective for pre-training: MAsked Sequence to Sequence learning (MASS) for language generation. MASS is based on the sequence to sequence learning framework: its encoder takes a sentence with a masked fragment (several consecutive tokens) as input, and its decoder predicts this masked fragment conditioned on the encoder representations. Unlike BERT or a language model that pre-trains

[*]Equal contribution [1]Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology [2]Microsoft Research. Correspondence to: Tao Qin <taoqin@microsoft.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

only the encoder or decoder, MASS is carefully designed to pre-train the encoder and decoder jointly in two steps: 1) By predicting the fragment of the sentence that is masked on the encoder side, MASS can force the encoder to understand the meaning of the unmasked tokens, in order to predict the masked tokens in the decoder side; 2) By masking the input tokens of the decoder that are unmasked in the source side, MASS can force the decoder rely more on the source representation other than the previous tokens in the target side for next token prediction, better facilitating the joint training between encoder and decoder.

MASS just needs to pre-train one model and then fine-tune on a variety of downstream tasks. We use transformer as the basic sequence to sequence model and pre-train on the WMT monolingual corpus[1], and then fine-tune on three different language generation tasks including NMT, text summarization and conversational response generation. Considering the downstream tasks cover cross-lingual task like NMT, we pre-train one model on multiple languages. We explore the low-resource setting for all the three tasks, and also consider unsupervised NMT which is a purely zero-resource setting. For NMT, the experiments are conducted on WMT14 English-French, WMT16 English-German and WMT16 English-Romanian datasets. For unsupervised NMT, we directly fine-tune the pre-trained model on monolingual data with back-translation loss (Lample et al., 2018), instead of using additional denoising auto-encoder loss as in Lample et al. (2018). For low-resource NMT, we fine-tune our model on limited bilingual data. For the other two tasks, we conduct experiments on: 1) the Gigaword corpus for abstractive text summarization; 2) the Cornell Movie Dialog corpus for conversational response generation. Our method achieves improvements on all these tasks as well as both the zero- and low-resource settings, demonstrating our method is effective and applicable to a wide range of sequence generation tasks.

The contributions of this work are listed as follows: 1) We propose MASS, a masked sequence to sequence pre-training method for language generation; 2) We apply MASS on a variety of language generation tasks including NMT, text summarization and conversational response generation, and achieve significant improvements, demonstrating the effectiveness of our proposed method. Specially, we achieve a state-of-the art BLEU score for unsupervised NMT on two language pairs: English-French and English-German, and outperform the previous unsupervised NMT method (Lample & Conneau, 2019) by more than 4 points on English-French and 1 point on French-English in terms of BLEU score, and even beating the early attention-based supervised model (Bahdanau et al., 2015b). We will release the code on Github.

[1]The monolingual data for each language is downloaded from http://www.statmt.org/wmt16/translation-task.html.

2. Related Work

There are a lot of works on sequence to sequence learning and the pre-training for natural language processing. We briefly review several popular approaches in this section.

2.1. Sequence to Sequence Learning

Sequence to sequence learning (Cho et al., 2014; Bahdanau et al., 2015a; Wu et al., 2016; Gehring et al., 2017; Vaswani et al., 2017) is a challenging task in artificial intelligence, and covers a variety of language generation applications such as NMT (Cho et al., 2014; Bahdanau et al., 2015a; Wu et al., 2016; Gehring et al., 2017; Vaswani et al., 2017; Tan et al., 2019; Artetxe et al., 2017; Lample et al., 2017; 2018; He et al., 2018; Hassan et al., 2018), text summarization (Ayana et al., 2016; Suzuki & Nagata, 2017; Gehring et al., 2017), question answering (Yuan et al., 2017) and conversational response generation (Shang et al., 2015; Vinyals & Le, 2015).

Sequence to sequence learning has attracted much attention in recent years due to the advance of deep learning. However, many language generations tasks such as NMT lack paired data but have plenty of unpaired data. Therefore, the pre-training on unpaired data and fine-tuning with small-scale paired data will be helpful for these tasks, which is exactly the focus of this work.

2.2. Pre-training for NLP tasks

Pre-training has been widely used in NLP tasks to learn better language representation. Previous works mostly focus on natural language understanding tasks, and can be classified into feature-based approaches and fine-tuning approaches. Feature-based approaches mainly leverage pre-training to provide language representations and features to the downstream tasks, which includes word-level representations (Brown et al., 1992; Ando & Zhang, 2005; Blitzer et al., 2006; Collobert & Weston, 2008; Mikolov et al., 2013; Pennington et al., 2014) and sentence-level representations (Kiros et al., 2015; Logeswaran & Lee, 2018; Le & Mikolov, 2014), as well as context sensitive features from the NMT model (McCann et al., 2017) and ELMo (Peters et al., 2018). Fine-tuning approaches mainly pre-train a model on language modeling objective and then fine-tune the model on the downstream tasks with supervised data (Dai & Le, 2015; Howard & Ruder, 2018; Radford et al., 2018; Devlin et al., 2018). Specifically, Devlin et al. (2018) proposed BERT based on masked language modeling and next sentence prediction and achieved a state-of-the-art accuracy on multiple language understanding tasks in the GLUE benchmark (Wang et al., 2018) and SQuAD (Ra-

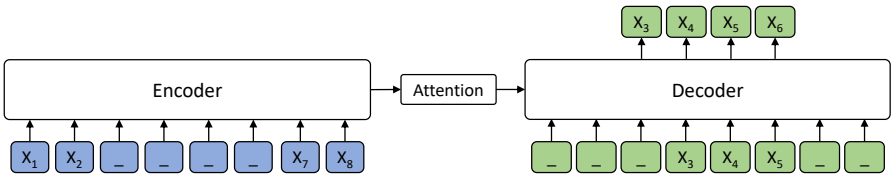


Figure 1. The encoder-decoder framework for our proposed MASS. The token "_" represents the mask symbol [$\mathbb{M}$].

jpurkar et al., 2016).

There are also some works pre-training the encoder-decoder model for language generation. Dai & Le (2015); Ramachandran et al. (2016) leverage a language model or auto-encoder to pre-train the encoder and decoder. Their improvements, although observed, are limited and not as general and significant as the pre-training methods (e.g., BERT) for language understanding. Zhang & Zong (2016) designed a sentence reordering task for pre-training, but only for the encoder part of the encoder-decoder model. Zoph et al. (2016); Firat et al. (2016) pre-train the model on similar rich-resource language pairs and fine-tuned on the target language pair, which relies on supervised data on other language pairs. Recently, XLM (Lample & Conneau, 2019) pre-trained BERT-like models both for the encoder and decoder, and achieved the previous state of the art results on unsupervised machine translation. However, the encoder and decoder in XLM are pre-trained separately and the encoder-decoder attention mechanism cannot be pre-trained, which are sub-optimal for sequence to sequence based language generation tasks.

Different from previous works, our proposed MASS is carefully designed to pre-train both the encoder and decoder jointly using only unlabeled data, and can be applied to most language generations tasks.

3. MASS

In this section, we first introduce the basic framework of sequence to sequence learning, and then propose MASS (MAsked Sequence to Sequence pre-training). We then discuss the differences between MASS and previous pre-training methods including the masked language modeling in BERT and standard language modeling.

3.1. Sequence to Sequence Learning

We denote $(x, y) \in (\mathcal{X}, \mathcal{Y})$ as a sentence pair, where $x = (x_1, x_2, ..., x_m)$ is the source sentence with m tokens, and $y = (y_1, y_2, ..., y_n)$ is the target sentence with n tokens, and $\mathcal{X}$ and $\mathcal{Y}$ are the source and target domains. A sequence to sequence model learns the parameter θ to estimate the conditional probability $P(y|x;\theta)$, and usually uses log likelihood as the objective function: $L(\theta;(\mathcal{X},\mathcal{Y})) = \Sigma_{(x,y)\in(\mathcal{X},\mathcal{Y})} \log P(y|x;\theta)$. The conditional probability $P(y|x;\theta)$ can be further factorized according to the chain rule: $P(y|x;\theta) = \prod_{t=1}^{n} P(y_t|y_{<t}, x;\theta)$, where $y_{<t}$ is the proceeding tokens before position t.

A major approach to sequence to sequence learning is the encoder-decoder framework: The encoder reads the source sequence and generates a set of representations; the decoder estimates the conditional probability of each target token given the source representations and its preceding tokens. Attention mechanism (Bahdanau et al., 2015a) is further introduced between the encoder and decoder to find which source representation to focus on when predicting the current token.

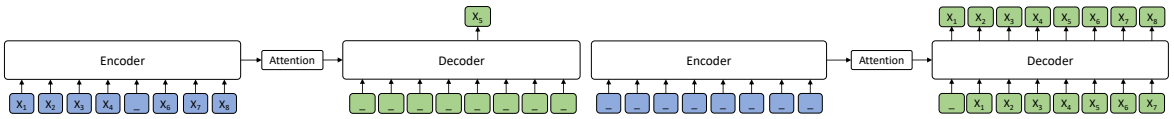
3.2. Masked Sequence to Sequence Pre-training

We introduce a novel unsupervised prediction task in this section. Given an unpaired source sentence $x \in \mathcal{X}$, we denote $x^{\backslash u:v}$ as a modified version of x where its fragment from position u to v are masked, $0 < u < v < m$ and m is the number of tokens of sentence x. We denote $k = v-u+1$ as the number of tokens being masked from position u to v. We replace each masked token by a special symbol [$\mathbb{M}$], and the length of the masked sentence is not changed. $x^{u:v}$ denotes the sentence fragment of x from u to v.

MASS pre-trains a sequence to sequence model by predicting the sentence fragment $x^{u:v}$ taking the masked sequence $x^{\backslash u:v}$ as input. We also use the log likelihood as the objective function:

$$\begin{aligned} L(\theta;\mathcal{X}) &= \frac{1}{|\mathcal{X}|}\Sigma_{x\in\mathcal{X}} \log P(x^{u:v}|x^{\backslash u:v};\theta) \\ &= \frac{1}{|\mathcal{X}|}\Sigma_{x\in\mathcal{X}} \log \prod_{t=u}^{v} P(x_t^{u:v}|x_{<t}^{u:v}, x^{\backslash u:v};\theta). \end{aligned} \tag{1}$$

We show an example in Figure 1, where the input sequence has 8 tokens with the fragment $x_3x_4x_5x_6$ being masked. Note that the model only predicts the masked fragment $x_3x_4x_5x_6$, given $x_3x_4x_5$ as the decoder input for position $4-6$, and the decoder takes the special mask symbol [$\mathbb{M}$] as inputs for the other positions (e.g., position $1-3$ and $7-8$). While our method works for any neural network based encoder-decoder frameworks, we choose Transformer in our experiments, considering that it achieves state-of-the-



(a) Masked language modeling in BERT ($k = 1$) (b) Standard language modeling ($k = m$)

Figure 2. The model structure of MASS when $k = 1$ and $k = m$. Masked language modeling in BERT can be viewed as the case $k = 1$ and standard language modeling can be viewed as the case $k = m$.

art performances in multiple sequence to sequence learning tasks.

Actually, the masked language modeling in BERT (Devlin et al., 2018) and the standard language modeling (Bengio et al., 2003; Mikolov et al., 2010) in GPT (Radford et al., 2018) can be viewed as special cases of MASS. We have an important hyperparameter k, which denotes the length of the masked fragment of the sentence. Our method with different k values can cover the special cases that are related to previous pre-training methods, as shown in Table 1.

When $k = 1$, the masked fragment in the source sentence contains only one token, and the decoder predicts this token without any tokens as input but conditioned on the unmasked source tokens, as shown in Figure 2a. It becomes the masked language modeling as used in BERT. One may argue that the model structure is a little bit different from the masked language model. However, since all the input tokens of the decoder are masked, the decoder is itself like a non-linear classifier, analogous to the softmax matrix used in BERT. In this case, the conditional probability is $P(x^u|x^{\setminus u};\theta)$ and u is the position of the masked token, which is exactly the formulation of masked language modeling used in BERT[2].

When $k = m$ where m is the number of tokens in sentence x, all the tokens on the encoder side are masked and the decoder needs to predict all tokens given previous tokens, as shown in Figure 2b. The conditional probability is $P(x^{1:m}|x^{\setminus 1:m};\theta)$, and it becomes the standard language modeling in GPT, conditioned on null information from the encoder as all the tokens in the encoder side are masked.

3.3. Discussions

MASS is a pre-training method for language generation. While its special cases are related to the previous methods including the standard language modeling in GPT and the masked language modeling in BERT, it is different from these methods in general.

- Standard language modeling has long been used for pre-training, and the most prominent ones are the recently proposed ELMo (Peters et al., 2018) and OpenAI GPT (Radford et al., 2018). BERT introduces two pre-training tasks (masked language modeling and next sentence prediction) for natural language understanding, and uses one encoder to extract the representation for a single sentence or a pair of sentences. Both standard language modeling and BERT can just pre-train the encoder or decoder separately. While achieving promising results on language understanding tasks, they are not suitable for language generation tasks which typically leverage an encoder-decoder framework for conditional sequence generation.

- MASS is designed to jointly pre-train the encoder and decoder for language generation tasks. First, by only predicting the masked tokens through a sequence to sequence framework, MASS forces the encoder to understand the meaning of the unmasked tokens, and also encourages the decoder to extract useful information from the encoder side. Second, by predicting consecutive tokens in the decoder side, the decoder can build better language modeling capability than just predicting discrete tokens. Third, by further masking the input tokens of the decoder which are not masked in the encoder side (e.g., when predicting fragment $x_3x_4x_5x_6$, only the tokens $x_3x_4x_5$ are taken as the input and other tokens are masked with $[\mathbb{M}]$), the decoder is encouraged to extract more useful information from the encoder side, rather than leveraging the abundant information from the previous tokens.

Length	Probability	Model
$k = 1$	$P(x^u\|x^{\setminus u};\theta)$	masked LM in BERT
$k = m$	$P(x^{1:m}\|x^{\setminus 1:m};\theta)$	standard LM in GPT
$k \in (1, m)$	$P(x^{u:v}\|x^{\setminus u:v};\theta)$	methods in between

Table 1. Masked language modeling in BERT and standard language modeling, as special cases covered in MASS.

4. Experiments and Results

In this section, we describe the experimental details about MASS pre-training and fine-tuning on a variety of language

[2]One may argue that the masked language modeling in BERT randomly masks multiple tokens rather than just one token at a time. However, the key idea behind masking language modeling in BERT is to leverage bidirectional context information. Masking multiple tokens at a time is mainly for training speedup.

generation tasks, including NMT, text summarization, conversational response generation.

4.1. MASS Pre-training

Model Configuration We choose Transformer (Vaswani et al., 2017) as the basic model structure, which consists of 4-layer encoder and 4-layer decoder with 512 embedding/hidden size and 2048 feed-forward filter size. Considering the language generation tasks contain monolingual tasks such as text summarization and conversational response generation, as well as cross-lingual tasks such as NMT, we pre-train a single model with multiple languages besides English. In our current setting, we consider four languages (English, German, French and Romanian) in MASS, where English is used in all downstream tasks, while German, French, and Romanian are included in the NMT task. To distinguish between different languages, we also add a language embedding to the input sentence.

Datasets We use the monolingual data from WMT News Crawl datasets[3], where we choose 50M sentences from year 2007 to 2017 for English, French, German respectively. We also include a low-resource language: Romanian, in the pre-training stage, to verify the effectiveness of MASS pre-trained with low-resource language. We use all of the available Romanian sentences from News Crawl dataset and augment it with WMT16 monolingual data, which results in 2.9M sentences. All these languages are jointly tokenized into sub-word units with Byte-Pair Encoding (Sennrich et al., 2016). We share the vocabulary of all the languages during pre-training and the vocabulary size is set to 60K.

Pre-Training Details We mask the fragment by replacing the consecutive tokens with special symbols $[\mathbb{M}]$, with random start position u. Following Devlin et al. (2018), the masked tokens in the encoder will be a $[\mathbb{M}]$ token 80% of the time, a random token 10% of the time and a unchanged token 10% of the time. We set the fragment length k as roughly 50% of the total number of tokens in the sentence and also study different k to compare their accuracy changes. To reduce the memory and computation cost, we removed the padding in the decoder (the masked tokens) but keep the positional embedding of the unmasked tokens unchanged (e.g., if the first two tokens are masked and removed, the position for the third token is still 2 but not 0). In this way, we can get similar accuracy and reduce 50% computation in the decoder. We use Adam optimizer (Kingma & Ba, 2015) with a learning rate of 10^{-4} for the pre-training. The model are trained on 4 NVIDIA P100 GPU cards and each mini-batch contains $32 * 4$ sentences.

[3]While we choose the WMT monolingual data in the current setting, pre-training on Wikipedia data is also feasible.

To verify the effectiveness of MASS, we fine-tune the pre-trained model on three language generation tasks: NMT, text summarization and conversational response generation. We explore the low-resource setting on these tasks where we just leverage few training data for fine-tuning to simulate the low-resource scenario. For NMT, we mainly investigate the zero-resource (unsupervised) setting, as unsupervised NMT has become a challenging task in recent years (Artetxe et al., 2017; Lample et al., 2017; 2018).

4.2. Fine-Tuning on NMT

In this section, we first describe the experiments on the unsupervised NMT, and then introduce the experiments on low-resource NMT.

Experimental Setting For unsupervised NMT, there is no bilingual data to fine-tune the pre-trained model. Therefore, we leverage the monolingual data that is also used in the pre-training stage. Different from Artetxe et al. (2017); Lample et al. (2017; 2018), we just use back-translation to generate pseudo bilingual data for training, without using denoising auto-encoder[4]. We share the model between two translation directions, e.g., English-German and German-English, during fine-tuning, as used in Lample et al. (2018); Lample & Conneau (2019). During fine-tuning, we use Adam optimizer (Kingma & Ba, 2015) with initial learning rate 10^{-4}. During evaluation, we calculate the BLEU score with multi-bleu.pl[5] on *newstest2014* for English-French, and *newstest2016* for English-German and English-Romanian.

Results on Unsupervised NMT We first compare with Lample et al. (2018), which is the previous state-of-the-art method without pre-training. For fair comparison, we implement MASS on the codebase of Lample et al. (2018) and adopt the same data size, batch size and model size. As shown in Table 2, MASS outperforms Lample et al. (2018) on the three unsupervised translation tasks, with 6 translation directions in total. On English-French language pair, our method outperforms Lample et al. (2018) by more than 2 BLEU points on both en-fr and fr-en.

Recently, XLM is recently proposed by Lample & Conneau (2019) for cross-lingual language model pre-training, which covers several pre-training methods: masked language model (MLM) and causal language model (CLM). We also implement MASS based on the codebase of XLM[6] for comparison. We use the same data size, batch size, and model configuration with XLM, and also follow the same fine-tuning strategy of XLM, which samples 5M monolin-

[4]MASS is better than denoising auto-encoder as we will show in Table 3.

[5]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

[6]https://github.com/facebookresearch/XLM

Method	Setting	en - fr	fr - en	en - de	de - en	en - ro	ro - en
Artetxe et al. (2017)	2-layer RNN	15.13	15.56	6.89	10.16	-	-
Lample et al. (2017)	3-layer RNN	15.05	14.31	9.75	13.33	-	-
Yang et al. (2018)	4-layer Transformer	16.97	15.58	10.86	14.62	-	-
Lample et al. (2018)	4-layer Transformer	25.14	24.18	17.16	21.00	21.18	19.44
MASS	4-layer Transformer	**27.41**	**27.09**	**18.21**	**23.37**	**22.37**	**20.74**
XLM (Lample & Conneau, 2019)	6-layer Transformer	33.40	33.30	27.00	34.30	-	-
MASS*	6-layer Transformer	**37.50**	**34.60**	**28.10**	**35.00**	-	-

Table 2. The BLEU score comparisons between MASS and the previous works on unsupervised NMT. Results on en-fr and fr-en pairs are reported on *newstest2014* and the others are on *newstest2016*. MASS* adopts the same configuration with XLM, which has bigger batch sizes, data sizes and model capacity than MASS.

gual data from NewsCrawl 2007-2008. We denote this version of MASS as MASS*. We report results with beam search with a beam size of 10. As shown in the last line in Table 2, MASS* outperforms XLM by 4.1 BLEU points on en-fr and 1.3 BLEU points in fr-en translation.

Compared with Other Pre-training Methods We also compare MASS with the previous pre-training methods for language generation tasks. The first baseline is *BERT+LM*, which use masked language modeling in BERT to pre-train the encoder and the standard language modeling to pre-train the decoder. This baseline is relatively stronger than the previous works that just leverage language modeling for pre-training, since masked language modeling in BERT has shown the advantage over standard language modeling. We do not use masked language modeling to pre-train the decoder due to most language generation tasks are auto-regressive where bidirectional language modeling is not suitable. The second baseline is *DAE*, which simply uses denoising auto-encoder (Vincent et al., 2008) to pre-train the encoder and decoder. We pre-train the model with *BERT+LM* and *DAE*, and fine-tune on the unsupervised translation pairs with same fine-tuning strategy of MASS.

As shown in Table 3, *DAE* achieves higher BLEU score than *BERT+LM*, and MASS outperforms both *BERT+LM* and *DAE* on all the unsupervised translation pairs. While *DAE* usually leverages some denoising methods like randomly masking tokens or swapping adjacent tokens, the decoder can still easily learn to copy the unmasked tokens through encoder-decoder attention[7]. On the other hand, the decoder in *DAE* takes the full sentence as the input, which is enough to predict the next token like the language model, and is not forced to extract additional useful representation from the encoder.

[7]The popular encoder-decoder based model structures (Wu et al., 2016; Gehring et al., 2017; Vaswani et al., 2017) all adopt residual connection (He et al., 2016). Therefore, the token generation in the top layer of the decoder side can directly depend on the token embedding in the encoder side through residual connection and attention.

Method	en-fr	fr-en	en-de	de-en	en-ro	ro-en
BERT+LM	26.12	25.13	17.07	21.91	21.33	19.56
DAE	26.42	25.78	17.33	22.29	21.48	19.78
MASS	**27.41**	**27.09**	**18.21**	**23.37**	**22.37**	**20.74**

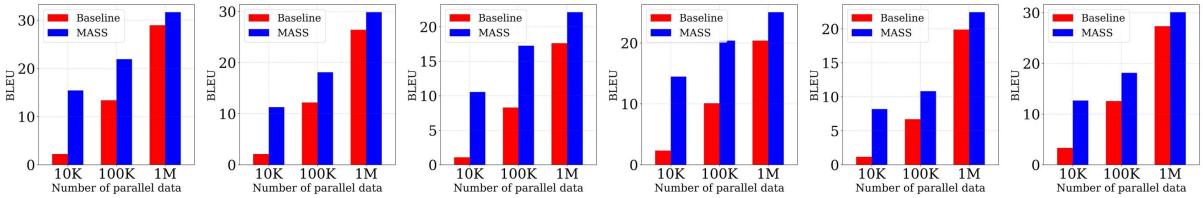
Table 3. The BLEU score comparisons between MASS and the two baseline pre-training methods.

Experiments on Low-Resource NMT In the low-resource NMT setting, we respectively sample 10K, 100K, 1M paired sentence from the bilingual training data of WMT14 English-French, WMT16 English-German and WMT16 English-Romanian, to explore the performance of our method in different low-resource scenarios. We use the same BPE codes learned in the pre-trained stage to tokenize the training sentence pairs. We fine-tune the pre-trained model on the paired data for 20,000 steps with Adam optimizer and the learning rate is set as 10^{-4}. We choose the best model according to the accuracy on development set. We report the BLEU scores on the same testsets used in the unsupervised setting. As shown in Figure 3, MASS outperforms the baseline models that are trained only on the bilingual data without any pre-training on all the six translation directions, demonstrating the effectiveness of our method in the low-resource scenarios.

4.3. Fine-Tuning on Text Summarization

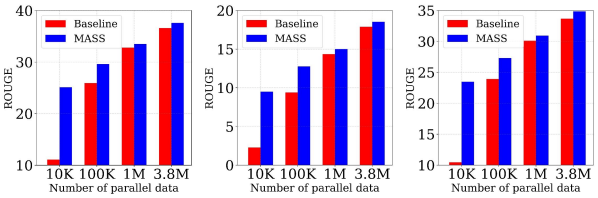
Experiment Setting Text summarization is the task of creating a short and fluent summary of a long text document, which is a typical sequence generation task. We fine-tune the pre-trained model on text summarization task with different scales (10K, 100K, 1M and 3.8M) of training data from the Gigaword corpus (Graff et al., 2003)[8], which consists of a total of 3.8M article-title pairs in English. We take the article as the encoder input and title as the decoder input for fine-tuning. We report the F1 score of ROUGE-1, ROUGE-2 and ROUGE-L on the Gigaword testset during evaluation.

[8]https://github.com/harvardnlp/sent-summary



(a) en-fr (b) fr-en (c) en-de (d) de-en (e) en-ro (f) ro-en

Figure 3. The BLEU score comparisons between MASS and the baseline on low-resource NMT with different scales of paired data.



(a) RG-1 (F) (b) RG-2 (F) (c) RG-L (F)

Figure 4. The comparisons between MASS and the baseline on text summarization task with different scales of paired data. The results are reported in ROUGE-1 (RG-1), ROUGE-2 (RG-2) and ROUGE-L (RG-L) respectively. F stands for F1-score.

We use beam search with a beam size of 5 for inference.

Results Our results are illustrated in Figure 4. We compare MASS with the model that is trained only on the paired data without any pre-training. MASS consistently outperforms the baseline on different scales of fine-tuning data (more than 10 ROUGE points gain on 10K data and 3 ROUGE points gain on 100K data), which demonstrates that MASS is effective in low-resource scenarios with different scale of training data on this task.

Method	RG-1 (F)	RG-2 (F)	RG-L (F)
BERT+LM	27.63	10.40	25.32
DAE	28.11	11.25	25.69
MASS	**29.79**	**12.75**	**27.45**

Table 4. The comparisons between MASS and two other pre-training methods in terms of ROUGE score on the text summarization task with 100K training data.

Compared with Other Pre-Training Methods We further compare MASS with the pre-training methods of *BERT+LM* and *DAE* described in Section 4.2, with 100K data on the text summarization task. As shown in Table 4, MASS consistently outperforms the two pre-training methods on the three ROUGE scores.

4.4. Fine-Tuning on Conversational Response Generation

Experimental Setting Conversational response generation generates a flexible response for the conversation (Shang et al., 2015; Vinyals & Le, 2015). We conduct experiments on the Cornell movie dialog corpus (Danescu-Niculescu-Mizil & Lee, 2011)[9] that contains 140K conversation pairs. We randomly sample 10K/20K pairs as the validation/test set and the remaining data is used for training. We adopt the same optimization hyperparameters from the pre-training stage for fine-tuning. We report the results with perplexity (PPL) following Vinyals & Le (2015).

Results We compare MASS with the baseline that is trained on the available data pairs. We conduct experiments on the 10K pairs (randomly chosen) and the whole 110K pairs, and show the results in Table 5. MASS achieves lower PPL than the baseline on both the 10K and 110K data.

Method	Data = 10K	Data = 110K
Baseline	82.39	26.38
BERT+LM	80.11	24.84
MASS	**74.32**	**23.52**

Table 5. The comparisons between MASS and other baseline methods in terms of PPL on Cornell Movie Dialog corpus.

Compared with Other Pre-Training Methods We also compare MASS with the pre-training methods of *BERT+LM* and *DAE* on conversational response generation. As shown in Table 5, MASS consistently outperforms the two pre-training methods with lower PPL on 10K and 110K training data respectively.

4.5. Analysis of MASS

Study of Different k The length of the masked fragment k is an important hyperparameter of MASS and we have varied k in Section 3.2 to cover the special cases of masked language modeling in BERT and standard language modeling. In this section, we study the performance of MASS with different k, where we choose k from 10% to 90% percentage of the sentence length m with a step size of 10%, plus with $k = 1$ and $k = m$.

We observe both the performance of MASS after pre-training, as well as the performance after fine-tuning on several language generation tasks, including unsupervised

[9] https://github.com/suriyadeepan/datasets/tree/master/seq2seq/cornell_movie_corpus

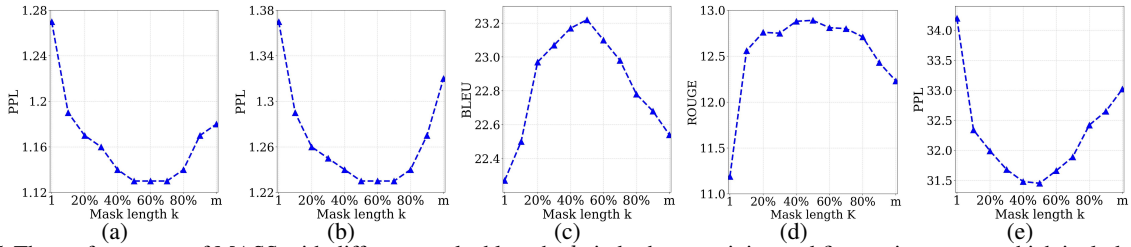


Figure 5. The performances of MASS with different masked lengths k, in both pre-training and fine-tuning stages, which include: the PPL of the pre-trained model on English (Figure a) and French (Figure b) sentences from WMT newstest2013 on English-French translation; the BLEU score of unsupervised English-French translation on WMT newstest2013 (Figure c); the ROUGE score (F1 score in RG-2) on the validation set of text summarization (Figure d); the PPL on the validation set of conversational response generation (Figure e).

Method	BLEU	Method	BLEU	Method	BLEU
Discrete	26.76	*Feed*	25.56	MASS	27.41

Table 6. The comparison between MASS and the ablation methods in terms of BLEU score on the unsupervised en-fr translation.

English-French translation, text summarization and conversational response generation. We first show the perplexity (PPL) of the pre-training model on the English and French languages with different k. We choose the English and French sentences from newstest2013 of WMT En-Fr as the validation set, and plot the PPL in Figure 5a (English) and 5b (French). It can be seen that the pre-trained model achieves the best validation PPL when k is between 50% and 70% of the sentence length m. We then observe the performance on fine-tuning tasks. We show the curve of the validation BLEU scores on unsupervised En-Fr translation in Figure 5c, the validation ROUGE scores on text summarization in Figure 5d, and the validation PPL on conversational response generation in Figure 5e. It can be seen that MASS achieves best performance on these downstream tasks when k is nearly 50% of the sentence length m. Therefore, we set $k = 50\%$ of m for MASS in our experiments.

Actually, $k = 50\%$ of m is a good balance between the encoder and decoder. Too few valid tokens in the encoder side or in the decoder side will bias the model to concentrate more on the other side, which is not suitable for language generation task that typically leverages the encoder-decoder framework to extract the sentence representation in the encoder, as well as to model and generate the sentence in the decoder. The extreme cases are $k = 1$ (masked language modeling in BERT) and $k = m$ (standard language modeling), as illustrated in Figure 2. Neither $k = 1$ nor $k = m$ can achieve good performance on the downstream language generation tasks, as shown in Figure 5.

Ablation Study of MASS In our masked sequence to sequence pre-training, we have two careful designs: (1) We mask consecutive tokens in the encoder side, and thus predict consecutive tokens in the decoder side, which can build better language modeling capability than just predicting discrete tokens. (2) We mask the input tokens of the decoder which are not masked in the encoder side (e.g., when predicting fragment $x_3x_4x_5x_6$ in Figure 1, only the tokens $x_3x_4x_5$ are taken as the input and other tokens are masked with $[\mathbb{M}]$), to encourage the decoder to extract more useful information from the encoder side, rather than leveraging the abundant information from the previous tokens. In this section, we conduct two ablation studies to verify the effectiveness of the two designs in MASS. The first study is to randomly mask discrete tokens instead of consecutive tokens in MASS, denoted as *Discrete*. The second study is to feed all the tokens to the decoder instead of masking the input tokens of the decoder that are not masked in the encoder side, denoted as *Feed*. We compare MASS with the two ablation methods on the unsupervised English-French translation, as shown in Table 6. It can be seen that both *Discrete* and *Feed* perform worse than MASS, demonstrating the effectiveness of the two designs in MASS.

5. Conclusion

In this work, we have proposed MASS: masked sequence to sequence pre-training for language generation tasks, which reconstructs a sentence fragment given the remaining part of the sentence in the encoder-decoder framework. MASS just needs to pre-train one model and then fine-tune on multiple language generation tasks such as neural machine translation, text summarization and conversational response generation. Through experiments on the three above tasks and total eight datasets, MASS achieved significant improvements over the baseline without pre-training or with other pre-training methods. More specifically, MASS achieved the state-of-the-art BLEU scores for unsupervised NMT on three language pairs, outperforming the previous state-of-the-art by more than 4 BLEU points on English-French.

For future work, we will apply MASS to more language generation tasks such as sentence paraphrasing, text style transfer and post editing. We will also investigate more of the theoretical and empirical analysis on our masked sequence to sequence pre-training method.

Acknowledgements

This work was partially supported by the National Key Research and Development Program of China under Grant 2018YFB1004904. We also thank Yichong Leng and Weicong Chen for the further development on the work of MASS.

References

Ando, R. K. and Zhang, T. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. Unsupervised neural machine translation. *CoRR*, 2017.

Ayana, Shen, S., Liu, Z., and Sun, M. Neural headline generation with minimum risk training. *ArXiv*, 2016.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *ICLR 2015*, 2015a.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. 2015b.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Blitzer, J., McDonald, R., and Pereira, F. Domain adaptation with structural correspondence learning. In *EMNLP*, pp. 120–128. Association for Computational Linguistics, 2006.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

Collobert, R. and Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, pp. 160–167. ACM, 2008.

Dai, A. M. and Le, Q. V. Semi-supervised sequence learning. In *NIPS*, pp. 3079–3087, 2015.

Danescu-Niculescu-Mizil, C. and Lee, L. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *ACL Workshop*, 2011.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, 2018.

Firat, O., Sankaran, B., Al-Onaizan, Y., Vural, F. T. Y., and Cho, K. Zero-resource translation with multi-lingual neural machine translation. In *EMNLP*, pp. 268–277, 2016.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. Convolutional sequence to sequence learning. In *ICML*, volume 70, pp. 1243–1252, 2017.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pp. 580–587, 2014.

Graff, David, Kong, Junbo, Chen, Ke, Maeda, and Kazuaki. English gigaword. In *Linguistic Data Consortium*, 2003.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., et al. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

He, T., Tan, X., Xia, Y., He, D., Qin, T., Chen, Z., and Liu, T.-Y. Layer-wise coordination between encoder and decoder for neural machine translation. In *Advances in Neural Information Processing Systems*, pp. 7944–7954, 2018.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. In *ACL*, volume 1, pp. 328–339, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. Skip-thought vectors. In *NIPS*, pp. 3294–3302, 2015.

Lample, G. and Conneau, A. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291, 2019.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. Unsupervised machine translation using monolingual corpora only. *CoRR*, 2017.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. Phrase-based & neural unsupervised machine translation. In *EMNLP*, pp. 5039–5049, 2018.

Le, Q. and Mikolov, T. Distributed representations of sentences and documents. In *ICML*, pp. 1188–1196, 2014.

Logeswaran, L. and Lee, H. An efficient framework for learning sentence representations. *CORR*, 2018.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. Learned in translation: Contextualized word vectors. In *NIPS*, pp. 6294–6305, 2017.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *NIPS*, pp. 3111–3119, 2013.

Ouyang, W., Li, H., Zeng, X., and Wang, X. Learning deep representation with large-scale attributes. In *CVPR*, pp. 1895–1903, 2015.

Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. In *EMNLP*, pp. 1532–1543, 2014.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *NAACL*, volume 1, pp. 2227–2237, 2018.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. 2018.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. *CoRR*, 2016.

Ramachandran, P., Liu, P. J., and Le, Q. V. Unsupervised pretraining for sequence to sequence learning. *CoRR*, abs/1611.02683, 2016.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *ACL*, volume 1, pp. 1715–1725, 2016.

Shang, L., Lu, Z., and Li, H. Neural responding machine for short-text conversation. In *ACL*, volume 1, pp. 1577–1586, 2015.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pp. 1631–1642, 2013.

Suzuki, J. and Nagata, M. Cutting-off redundant repeating generations for neural abstractive summarization. In *ACL*, pp. 291–297, 2017.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *CVPR*, pp. 1–9, 2015.

Tan, X., Ren, Y., He, D., Qin, T., and Liu, T.-Y. Multilingual neural machine translation with knowledge distillation. In *ICLR*, 2019.

Tjong Kim Sang, E. F. and De Meulder, F. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *NAACL*, pp. 142–147. Association for Computational Linguistics, 2003.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NIPS*, pp. 6000–6010, 2017.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *ICML*, pp. 1096–1103. ACM, 2008.

Vinyals, O. and Le, Q. V. A neural conversational model. *CoRR*, abs/1506.05869, 2015.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461, 2018.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

Yang, Z., Chen, W., Wang, F., and Xu, B. Unsupervised neural machine translation with weight sharing. In *ACL*, pp. 46–55, 2018.

Yuan, X., Wang, T., Gulcehre, C., Sordoni, A., Bachman, P., Zhang, S., Subramanian, S., and Trischler, A. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 15–25, 2017.

Zhang, J. and Zong, C. Exploiting source-side monolingual data in neural machine translation. In *EMNLP*, pp. 1535–1545, 2016.

Zoph, B., Yuret, D., May, J., and Knight, K. Transfer learning for low-resource neural machine translation. In *EMNLP*, pp. 1568–1575, 2016.