A Appendix

A.1 Limitations

Multilinguality Although constructing large-scale MRC-style training data is feasible for resource-rich languages, such as English, extending this idea to resource-poor languages might be difficult due to the relatively small amount of anchors in their corresponding Wikipedia articles. Exploring other data resources to automatically construct large-scale pre-training data can remedy this issue. For example, given a word in the monolingual dictionaries, we can regard the word itself, the definition of this word, and the example sentence of this word as the MRC answer, query, and context respectively. We believe our MRC-style pre-training is still applicable for low-resource languages with such dictionaries.

Comparison with Large Language Models In this paper, we did not compare PMR with large language models (LLM) for the following two reasons. First, existing MLMs are small in scale. Therefore, we are unable to find a suitable MLM to make a fair comparison with LLMs. Second, studies have shown that LLMs yield inferior results compared to smaller MLMs on span extraction tasks, particularly those involving structured prediction [41, 43, 61, 31]. Based on this fact, we mainly compare with existing strong generative methods of comparable model size.

Few-shot NER results of SpanBERT We ran SpanBERT [20] in our NER few-shot settings. However, its performance was below our expectations. In all our few-shot settings, SpanBERT achieved an F1 score of 0 on CoNLL and WNUT datasets. Additionally, its performance on ACE04 and ACE05 datasets was significantly lower than RoBERTa [36]. Based on these outcomes, we only compare PMR with SpanBERT in the NER full-resource setting.

A.2 Fine-tuning Tasks

For EQA, we use the MRQA benchmark [15], including SQuAD [46], TriviaQA [21], NaturalQuestion [25], NewQA [56], SearchQA [14], HotpotQA [67], BioASQ [57], DROP [13], DuoRC [49], RACE [26], RelationExtraction [28], TextbookQA [22]. EQA has always been treated as an MRC problem, where the question serves as the MRC query, and the passage containing the answers serves as the MRC context. For NER, We follow MRC-NER [32] to formulate NER into the MRC paradigm, where the entity label together with its description serves as the MRC query, and the input text serves as the MRC context. The goal is to extract the corresponding entities as answers. We use the Eq. 4 as the learning objective, where $Y^{ext}_{i,j}$ indicates that the input span $X_{i:j}$ is an answer/entity.

For sequence classification tasks, we construct the MRC query and context as followed. MCQA: The query is the concatenation of the question and one choice, and the context is the supporting document. MNLI: The query is the entailment label concatenated with the label description, and the context is the concatenation of the premise and hypothesis. SST-2: The query is the sentiment label concatenated with the label description, and the context is the input sentence. We use Eq. 3 to fine-tune the classification tasks. Note that only the correct query-context pair would get $Y^{cls} = 1$. Otherwise, the supervision is $Y^{cls} = 0$. During inference, we select the query-context pair with the highest $S_{1,1}$ among all MRC examples constructed for the sequence classification instance as the final prediction. We show concrete examples for each task in Table 7 and Table 8.

A.3 Implementations

We download the 2022-01-01 dump[4] of English Wikipedia. For each article, we extract the plain text with anchors via WikiExtractor [3] and then preprocess it with NLTK [4] for sentence segmentation and tokenization. We consider the definition articles of entities that appear as anchors in at least 10 other articles to construct the query. Then, for each anchor entity, we pair its query from the definition article with 10 relevant contexts from other mention articles that explicitly mention the corresponding anchors and construct answerable MRC examples as described in Sec. 2. Unanswerable examples are formed by pairing the query with 10 irrelevant contexts.

[4]https://dumps.wikimedia.org/enwiki/latest

Task		Example Input	Example Output
EQA (SQuAD)	Ori.	Question: Which NFL team represented the NFC at Super Bowl 50? Context: Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers to earn their third Super Bowl title.	Answer: "Carolina Panthers"
	PMR	`[CLS]` Which NFL team represented the NFC at Super Bowl 50 ? `[SEP]` `[SEP]` Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season . The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers to earn their third Super Bowl title . `[SEP]`	(53,54) - "Carolina Panthers"
NER (CoNLL)	Ori.	Two goals in the last six minutes gave holders Japan an uninspiring 2-1 Asian Cup victory over Syria on Friday.	("Japan", LOC); ("Syria", LOC); ("Asian Cup", MISC)
	PMR	`[CLS]` "ORG" . Organization entities are limited to named corporate, governmental, or other organizational entities. `[SEP]` `[SEP]` Two goals in the last six minutes gave holders Japan an uninspiring 2-1 Asian Cup victory over Syria on Friday . `[SEP]`	$\emptyset$
		`[CLS]` "PER" . Person entities are named persons or family . `[SEP]` `[SEP]` Two goals in the last six minutes gave holders Japan an uninspiring 2-1 Asian Cup victory over Syria on Friday . `[SEP]`	$\emptyset$
		`[CLS]` "LOC" . Location entities are the name of politically or geographically defined locations such as cities , countries . `[SEP]` `[SEP]` Two goals in the last six minutes gave holders Japan an uninspiring 2-1 Asian Cup victory over Syria on Friday . `[SEP]`	(32,32) - "Japan"; (40,40) - "Syria"
		`[CLS]` "MISC" . Examples of miscellaneous entities include events , nationalities , products and works of art . `[SEP]` `[SEP]` Two goals in the last six minutes gave holders Japan an uninspiring 2-1 Asian Cup victory over Syria on Friday . `[SEP]`	(34,35) - "Asian Cup"

Table 7: MRC examples of span extraction. Ori. indicates the original data format of these NLU tasks.

We use Huggingface's implementations of RoBERTa [63] as the MLM backbone. During the pre-training stage, the window size W for choosing context sentences is set to 2 on both sides. We use the first $T = 1$ sentence as the MRC query. Sometimes, the sentence segmentation would wrongly segment a few words to form a sentence, which is not meaningful enough to serve as an MRC query. Therefore, we continue to include subsequent sentences to form the query as long the query length is short than 30 words. The learning rate is set to 1e-5, and the training batch size is set to 40 and 24 for $\text{PMR}_{\texttt{base}}$ and $\text{PMR}_{\texttt{large}}$ respectively in order to maximize the usage of the GPU memory. We follow the default learning rate schedule and dropout settings used in RoBERTa. We use AdamW [37] as our optimizer. We train both $\text{PMR}_{\texttt{base}}$ and $\text{PMR}_{\texttt{large}}$ for 3 epochs on 4 A100 GPU. Since the WAE is a discriminative objective, the pre-training is extremely efficient, which tasks 36 and 89 hours to finish all training processes for two model sizes respectively. We also reserve 1,000 definition articles to build a dev set (20,000 examples) for selecting the best checkpoint. Since the queries constructed by these definition articles have never been used in training, they can be used to estimate the general language understanding ability of the model instead of hand match. The hyper-parameters of $\text{PMR}_{\texttt{large}}$ on downstream NLU tasks can be found in Table 9 and Table 11 for full-supervision and few-shot settings respectively.

Task		Example Input	Example Output
MCQA (OBQA)	Ori.	Question: A positive effect of burning biofuel is: (A) shortage of crops for the food supply. (B) an increase in air pollution (C) powering the lights in a home. (D) deforestation in the amazon to make room for crops. Context: Biofuel is used to produce electricity by burning.	Answer Choice: C
	PMR	[CLS] A positive effect of burning biofuel is shortage of crops for the food supply . [SEP] [SEP] Biofuel is used to produce electricity by burning . [SEP]	∅
		[CLS] A positive effect of burning biofuel is an increase in air pollution . [SEP] [SEP] Biofuel is used to produce electricity by burning . [SEP]	∅
		[CLS] A positive effect of burning biofuel is powering the lights in a home . [SEP] [SEP] Biofuel is used to produce electricity by burning . [SEP]	(0,0) - "[CLS]"
		[CLS] A positive effect of burning biofuel is deforestation in the amazon to make room for crops . [SEP] [SEP] Biofuel is used to produce electricity by burning . [SEP]	∅
Sentence Classification (SST-2)	Ori.	This is one of Polanski's best films.	Positive
	PMR	[CLS] Negative , feeling not good . [SEP] [SEP] This is one of Polanski 's best films . [SEP]	∅
		[CLS] Positive , having a good feeling . [SEP] [SEP] This is one of Polanski 's best films . [SEP]	(0,0) - "[CLS]"
Sen. Pair Classification (MNLI)	Ori.	Hypothesis: You and your friends are not welcome here, said Severn. Premise: Severn said the people were not welcome there.	Entailment
	PMR	[CLS] Neutral. The hypothesis is a sentence with mostly the same lexical items as the premise but a different meaning . [SEP] [SEP] Hypothesis : You and your friends are not welcome here, said Severn . Premise : Severn said the people were not welcome there . [SEP]	∅
		[CLS] Entailment . The hypothesis is a sentence with a similar meaning as the premise . [SEP] [SEP] Hypothesis : You and your friends are not welcome here, said Severn . Premise : Severn said the people were not welcome there . [SEP]	(0,0) - "[CLS]"
		[CLS] Contradiction . The hypothesis is a sentence with a contradictory meaning to the premise . [SEP] [SEP] Hypothesis : You and your friends are not welcome here, said Severn . Premise : Severn said the people were not welcome there . [SEP]	∅

Table 8: MRC examples of sequence classification.

A.4 Analysis of Data Construction

In addition to the defaulted way of constructing MRC examples (the first sentence in the definition article is the query, and randomly find 10 contexts for pairing 10 MRC examples), we compare with some advanced strategies to pair the query and the context, including:

- Q-C Relevance: We still use the first sentence from the definition article as the query, but we only select the top P% or top P most similar contexts to the query, where the similarity score is computed as the combination of BM25 and SimCSE [17].

Dataset	CoNLL03	WNUT	ACE04	ACE05	MRQA	RACE	DREAM	MCTest	MNLI	SST-2
Query Length	32	32	64	64	64	128	128	128	64	64
Input Length	192	160	192	192	384	512	512	512	192	192
Batch Size	32	16	64	32	16	8	2	2	16	16
Learning Rate	2e-5	1e-5	2e-5	2e-5	2e-5	2e-5	2e-5	1e-5	1e-5	2e-5
Epoch	10	5	10	5	4	4	3	8	3	2

Table 9: Hyper-parameters settings in fine-tuning downstream tasks in full-supervision settings.

ID	Strategy	Query	Context	CoNLL	SQuAD	DREAM	SST-2
0	RoBERTa$_{\text{base}}$	N.A.	N.A.	92.3	91.2	66.4	95.0
1	Random	First 1	Random 10	93.2	92.2	66.7	94.8
2	Q-C Relevance (top P%)	First 1	top 30%	93.0	91.9	65.5	95.3
3	Q-C Relevance (top P)	First 1	top 10	93.2	92.1	65.8	94.8
4	Random (Defaulted)	First 1	Random 10 + Unanswerable	93.1	92.1	70.7	94.6
5	Q-C Relevance (top P)	First 1	top 10 + Unanswerable	93.1	92.2	69.7	94.7
6	Q Diversity	Random 5	Random 10 + Unanswerable	93.2	92.2	70.6	94.8
7	C Diversity	First 1	Cluster 10 + Unanswerable	92.8	92.2	70.5	95.1

Table 10: We try various advanced strategies to pair the query and the context to form an MRC example. the **Query** and **Context** columns indicate how to select possible query and context for pairing. + Unanswerable indicates that PMR also uses Unanswerable examples and is also trained with L_{cls}. Models are base-sized.

- Q Diversity: In searching for an anchor, we hope the query should be diverse enough such that the model would not make a hard match between the fixed query and the anchor. Therefore, we randomly select one sentence from the first P sentences in the definition article to serve as the query for the anchor, while we keep the same context selection strategy.
- C Diversity: We hope the contexts should also be diverse enough such that they provide more possible usages of an anchor. Therefore, We use K-means[5] to cluster all contexts containing the anchor into P clusters and randomly select 1 context in each cluster. Similar scores in K-means are also obtained via SimCSE.

We compare those advanced strategies with our defaulted one in Table 10, where two span extraction and sequence classification tasks are selected for evaluating the effectiveness of these strategies. First, we make a fast evaluation with only L_{ext} without unanswerable examples (i.e. Strategy 1,2,3). Comparing Q-C Relevance (top P%) against Q-C Relevance (top P), we can observe that it is better to sample contexts based on absolute values. In Wikipedia, the reference frequency of anchor entities is extremely unbalanced, where some frequent anchor entities such as "the United States" are referenced more than 200,000 times, while other rare anchor entities are only mentioned once or twice in other articles. Therefore, Q-C Relevance (top P%) would waste too much focus on the well-learned frequent anchor entities and affect the learning of other less frequent anchor entities.

Then, when trained on both answerable and unanswerable examples as well well guided with both L_{cls} and L_{ext}, we only sample an absolute number of contexts. However, comparing among Strategy 4,5,6,7, no significant difference between these strategies and our random sampling is observed. We

[5]https://github.com/subhadarship/kmeans_pytorch

Dataset	EQA	NER
Query Length	64	32
Input Length	384	192
Batch Size	12	12
Learning Rate	{5e-5,1e-4}	{5e-5,1e-4}
Max Epochs/Steps	12/200	20/200

Table 11: Hyper-parameters settings in fine-tuning downstream tasks in few-shot settings.

The Broncos took an early lead in Super Bowl 50 and never trailed. Newton was limited by Denver's defense, which sacked him seven times and forced him into three turnovers, including *a fumble* which they recovered for a touchdown. Denver linebacker Von Miller was named Super Bowl MVP, recording *five solo tackles*, 2½ sacks, and two forced fumbles.

	F1	EM
RoBERTa	7.3	0.1
T5-v1.1	12.6	0.0
FewshotBART	0.8	0.3
PMR	17.2	10.4

1. How many solo tackles did Von Miller make at Super Bowl 50?
   Gold: five solo tackles
   RoBERTa: forced him into three turnovers, including (✗)
   T5-v1.1: context: context: context: context: context: context: context: (✗)
   FewshotBART: ∅
   PMR: five solo tackles (✓)
2. Which Newton turnover resulted in seven points for Denver?
   Gold: a fumble
   RoBERTa: trailed. Newton was limited by Denver's defense, which sacked him seven times and forced him into three turnovers, including a fumble which they recovered (✗)
   T5-v1.1: . context: Newton's first Super Bowl touchdown came in Super Bowl 50. context: (✗)
   FewshotBART: Denver linebacker Von (✗)
   PMR: two forced fumbles (✗)

Figure 5: Zero-shot performance on SQuAD and a case study. The F1/EM scores are shown in the left-top corner.

suggest that the benefits from these heuristic strategies are marginal in the presence of large-scale training data. Therefore, in consideration of the implementation simplicity, we just use the Random strategy as our final PMR implementation.

A.5 Zero-shot Learning

To reveal PMR's inherent capability from its MRC-style pretraining, we show its zero-shot performance in Figure 5, where the F1 and Exact Match (EM) scores on the entire SQuAD dev set and a case study in answering several questions are presented. Without any fine-tuning, our PMR achieves 10.4 EM, whereas T5 and RoBERTa can barely provide a meaningful answer, as shown by their near-zero EM scores. In the case study, our PMR correctly answers the first question. For the second question, although PMR gives an incorrect answer, the prediction is still a grammatical phrase. In contrast, RoBERTa and T5-v1.1 always perform random extractions and generations. Such a phenomenon verifies that PMR obtains a higher-level language digest capability from the MRC-style pretraining and can directly tackle downstream tasks to some extent.

A.6 Better Comprehending capability

To verify that PMR can better comprehend the input text, we feed the models with five different query variants during CoNLL evaluation. The five variants are:

- The default query used for fine-tuning the model:
  `"[Label]". [Label description]`
- The query template is modified (v1):
  `What is the "[Label]" entity, where [Label description]?`
- The query template is modified (v2):
  `Identify the spans (if any) related to "[Label]" entity. Details: [Label description]`
- The label description in the query is paraphrased using ChatGPT (v1):
  `"[Label]". [Paraphrased Label description v1]`
- The label description in the query is paraphrased using ChatGPT (v2):
  `"[Label]". [Paraphrased Label description v2]`

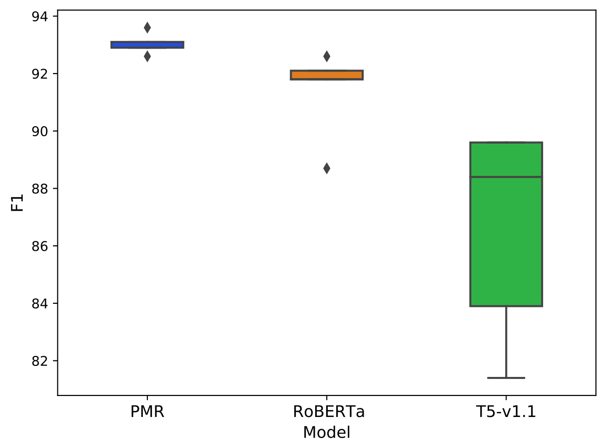

Figure 6: CoNLL performance when the models are fed with five different templates respectively during evaluation.

In Figure 6, we show the statistic results of the three models on CoNLL when five different query templates are used respectively during evaluation. Among the models, PMR demonstrated significantly higher and more stable performance than RoBERTa and T5-v1.1. Such a finding verifies our assumption that PMR can effectively comprehend the latent semantics of the input text despite being rephrased with varying lexical usage from the default query used for fine-tuning models.

A.7 Fully-Resource Results

Table 12 compares PMR with strong approaches in full-resource settings. On EQA and NER, PMR can significantly and consistently outperform previous approaches, where $\text{PMR}_{\texttt{large}}$ achieves up to 3.7 and 2.6 F1 improvements over $\text{RoBERTa}_{\texttt{large}}$ on WNUT and SearchQA, respectively. For the base-sized models, the advantage of PMR is more obvious, i.e. 1.4 F1 over $\text{RoBERTa}_{\texttt{base}}$. Apart from those, we also observe that: (1) PMR can also exceed strong generative approaches (i.e. UIE, T5-v1.1) on most tasks, demonstrating that the MRC paradigm is more suitable to tackle NLU tasks. (2) RoBERTa-Post, which leverages our Wikipedia corpus (a subset of its original pre-training data) for MLM-style continued-pretraining, performs poorly on most tasks, especially those with natural-question queries (i.e. EQA and MCQA). (3) PMR can be applied on even larger MLM such as $\text{ALBERT}_{\texttt{xxlarge}}$ [27] to gain stronger representation capability and further improve the performance of downstream tasks. Such findings suggest that with our MRC data format and WAE objective, PMR can leverage the same data to learn a high level of language understanding ability, beyond language representation.

EQA	**Unified**	**SQuAD**	**NewsQA**	**TriviaQA**	**SearchQA**	**HotpotQA**	**NQ**	**Avg.**
RBT-Post$_{\texttt{large}}$	✗	93.0	70.9	80.9	86.8	79.8	79.9	81.9
SpanBERT$_{\texttt{large}}$ [20]	✗	93.1	72.3	78.1	83.2	80.9	82.3	81.7
LUKE$_{\texttt{large}}$ [65]	✗	94.5	72.1	NA	NA	81.9	83.3	-
T5-v1.1$_{\texttt{large}}$ [45]	△	93.9	69.8	77.8	87.1	81.9	81.6	82.0
RoBERTa$_{\texttt{base}}$	✗	91.2	69.0	79.3	85.0	77.9	79.7	80.4
PMR$_{\texttt{base}}$ (OURS)	✓	92.1	71.9	81.5	86.4	80.6	81.0	82.3
RoBERTa$_{\texttt{large}}$	✗	94.2	73.8	85.1	85.7	81.6	83.3	84.0
PMR$_{\texttt{large}}$ (OURS)	✓	94.5	74.0	85.1	88.3	83.6	83.8	84.9
ALBERT$_{\texttt{xxlarge}}$	✗	94.7	75.3	86.0	89.4	83.8	83.8	85.5
PMR$_{\texttt{xxlarge}}$ (OURS)	✓	**95.0**	**75.4**	**86.7**	**89.6**	**84.5**	**84.8**	**86.0**

NER	**Unified**	**CoNLL**	**WNUT**	**ACE04**	**ACE05**	**Avg.**
Roberta$_{\texttt{large}}$+Tagging [36]	✗	92.4	55.4	-	-	-
RBT-Post$_{\texttt{large}}$	✗	92.7	53.8	86.6	86.2	79.8
SpanBERT$_{\texttt{large}}$	✗	90.3	47.2	86.4	85.4	77.3
LUKE$_{\texttt{large}}$ [65]	✗	92.4†	55.2†	-	-	-
CL-KL$_{\texttt{large}}$ [60]	✗	93.2†	59.3†	-	-	-
BARTNER$_{\texttt{large}}$ [66]	△	93.2‡	-	86.8‡	84.7‡	-
T5-v1.1$_{\texttt{large}}$ [45]	✓	90.5	46.7	83.9	82.8	76.0
UIE$_{\texttt{large}}$ [38]	✓	93.2$^{\spadesuit}$	52.5	86.9$^{\spadesuit}$	85.8$^{\spadesuit}$	79.6
RoBERTa$_{\texttt{base}}$	✗	92.3	53.9	85.8	85.2	79.3
PMR$_{\texttt{base}}$ (OURS)	✓	93.1	57.6	86.1	86.1	80.7
RoBERTa$_{\texttt{large}}$	✗	92.6	57.1	86.3	87.0	80.8
PMR$_{\texttt{large}}$ (OURS)	✓	**93.6**	**60.8**	87.5	87.4	**82.3**
ALBERT$_{\texttt{xxlarge}}$	✗	92.8	54.0	86.8	87.7	80.3
PMR$_{\texttt{xxlarge}}$ (OURS)	✓	93.2	58.3	**88.4**	**87.9**	82.0

Table 12: Performance on EQA (F1), and NER (F1). The best models are bolded. For EQA, as done in MRQA [15], we report the F1 on dev set and produce the results of SpanBERT and LUKE following the same protocol. Although we try hard to produce the results of LUKE for TriviaQA and SearchQA, its performance is unreasonably low. For CoNLL, we assume there is no additional context available and therefore we retrieve the results of CL-KL w/o context from [60]. Results labeled by †, ‡, and $^{\spadesuit}$ are cited from [60, 66, 38], respectively.