Masked prompt learning for formal analogies beyond words

Liyan Wang[1,*], Yves Lepage[1]

[1]Waseda University, 2-7 Hibikino, Kitakyushu, 808-0135, Japan

Abstract

Prompt learning, a recent thread in few-shot learning for pre-trained language models (PLMs), has been explored for completing word analogies in the extractive way. In this paper, we reformulate the analogy task as masked analogy completion task with the use of prompting to derive a generative model for analogies beyond words. We introduce a simple prompt-based fine-tuning paradigm for language modeling on answered prompts of analogies in the sequence-to-sequence framework. To convert discrete terms of analogies into linear sequences, we present a symbolic prompt template. The sequence-to-sequence model is fine-tuned to fill in the missing span of masked prompts deduced from different masking schemes on phrase analogies extracted from a small corpus. We analyze the out-of-distribution performance on sentence analogies which are unseen cases. Our experiments demonstrate that prompt-based fine-tuning with the objective of language modeling enables models to achieve significantly better performance on in-distribution cases than PLMs. Masked prompt learning with one-term masking exhibits the best out-of-distribution generalization on sentence analogies, with a difference of only 3 characters from references.

Keywords

Prompt learning, masked analogy completion, analogies beyond words, fine-tuning

1. Introduction

Analogy, a cognition mechanism that relies on relational similarity, is growing in prominence in the field of artificial intelligence [1]. In general, it encapsulates a quadruplet relationship between terms of the same type. For example, the famous analogy between words *king* : *queen* :: *man* : *woman,* implies that *king* is to *queen* as *man* is to *woman* in terms of gender transition. Strikingly, analogical quadruples form geometric parallelograms in pre-trained embedding spaces learnt by the skip-gram model with negative sampling [2]. The parallelogram generic has drawn attention in research about the linear algebraic properties of vector analogies [3]. The simple arithmetic approach, however, was questioned as being applicable only for completing analogies with respect to certain clearly defined relations [4]. Recent efforts [5, 6] have examined the potential of machine learning techniques to learn analogies in word embedding spaces.

By formulating tasks in the manner of analogical reasoning, sentence analogy has demonstrated its versatility in various tasks in the area of natural language processing (NLP), such

IARML@IJCAI-ECAI'2022: Workshop on the Interactions between Analogical Reasoning and Machine Learning, at IJCAI-ECAI'2022, July, 2022, Vienna, Austria

*Corresponding author.

wangliyan0905@toki.waseda.jp (L. Wang); yves.lepage@waseda.jp (Y. Lepage)

as machine translation [7], text summarization [8], and question answering [9]. In contrast to word-level analogies, analogies that go beyond words may encompass manifold challenges that stem from the inherent complexity of language. A recent work [10] has demonstrated that sentence embedding models struggle to capture analogical regularities in terms of geometric parallelism. The vector offsets may not be kept within sentence embeddings. In [11], the postulate of exchange of the means has been debated for classifying positive and negative analogies between sentences.

Some work on completing sentence analogies has focused on extractive approaches, i.e., identifying the optimum solution from candidates (i.e., from a finite answer pool) [9, 8]. However, the extractive mechanism is not geared to text generation. It is relatively expensive to define candidate sets for analogies that capture complex relations between long sequences. In addition, hand-crafted candidates may leave weaknesses in linguistic creativity. Therefore, it highlights the necessity for a generative model that can automatically produce the missing term in analogy questions, which will contribute to analogy completion in NLP scenarios.

Prompt learning [12] is a fruitful learning paradigm in recent work on the adaptive performance of PLMs in the few-shot setting. The downstream tasks are reformulated as Cloze-style problems by converting inputs into natural language prompts with task-specific descriptions, which allows PLMs to predict target outputs conditioned on prompts [13]. Following [13], a filled prompt refers to a prompt where the mask slot is filled with any answer. An answered prompt refers to a prompt filled with the correct answer. Based on prompt learning, recent works have investigated the few-shot [12] and zero-shot [14] performance of PLMs in identifying word analogies from candidates by language model (LM) scoring on filled prompts.

In this paper, we present a preliminary study of generative completion of analogies beyond words by using LMs in conjunction with prompting. To learn analogical regularities, we introduce a novel prompt-based fine-tuning method to extract sequential features of answered prompts with the symbolic template by masked sequence-to-sequence learning. We conduct lightweight fine-tuning on phrase analogies with different masking schemes. By language modeling on answered prompts, fine-tuned models achieve over 97% accuracy on solving phrase analogies. The fine-tuned sequence-to-sequence LMs have more promising out-of-distribution generalization on sentence analogies than autoregressive LMs. In particular, one-term masking is more robust at extracting analogical regularities, which contributes to adapt effectively to unseen analogies.

2. Related Work

Based on prompting, recent works have formulated the quadruplet problem as language modeling. The GPT-3 work [12] explored the adaptation of the explosive-size LM (with 175B parameters, which is over 100 times larger than GPT-2) on Stochastic Aptitude Tests (SAT) analogy task in the few shot setting with no gradient updates. The discrete texts in analogies are mapped into natural sentences with a textual prompt template. The pre-trained GPT-3 model learns within the context consisting of few answered prompts and the query, to infer the answer with the maximum LM likelihood among the given candidates. This exhibits the potential of GPT-3 for extractive analogy completion at the word level. It gave rise to the exploration

of analogical capabilities of LMs with the help of prompts. A recent work [14] examined the adaptability of PLMs to recognize word analogies with different levels of complexity in the zero-shot setting. They conduct ablation experiments in several aspects including prompt engineering, architecture engineering and scoring engineering. With the appropriate choices, PLMs can achieve meaningful zero-shot performance on analogy identification.

Recent efforts on prompt learning found that prompt-based fine-tuning of PLMs can improve effective learning on downstream tasks. These works have the advantage of being specific to the task by tuning the LM parameters entirely [15] or partially [16] on a small number of examples. Typically, prompt-based fine-tuning is adopted on masked language models (MLMs) to predict the mask token fixedly pointing to the target label in prompts. Here, we graft prompt-based fine-tuning onto a sequence-to-sequence LM with different masking schemes and execute masked sequence-to-sequence learning on a large number of examples to extensively model prompt sequences. The analogy task is reformulated as masked analogy completion, where sequence answers are generated by a fine-tuned LM to fill in the mask token in prompts for analogy questions.

3. Method

In this section, we introduce a generative method for completing analogies beyond words by using prompts with a symbolic template (Section 3.1). To adapt to the analogy task, we propose a novel prompt-based fine-tuning paradigm (Section 3.3) for PLMs to reconstruct the missing span in prompts processed by masking schemes (Section 3.2).

3.1. Symbolic Prompt Template

Prompt design is crucial in prompt learning. For analogies beyond words, it is easy to get the tokens of analogical words mixed up with the context tokens of textual templates, as used in [12, 14]. We introduce a clear prompt in which the contents of the four terms are easily parsed out from sequences.

Symbolic prompts for analogy are formally identical to analogical equations $A : B :: C : D$, where the ratio (:) and proportion (::) characters are two symbolic tokens in the prompt template. We employ the Unicode characters U+2236 and U+2237 for ratio and proportion in sequences. Let X ($X \in \{A, B, C, D\}$) denote one of the four terms in an analogy. The length of an answered prompt can be calculated as $|A| + 1 + |B| + 1 + |C| + 1 + |D| = \sum_{X \in \{A,B,C,D\}} |X| + 3$.

Compared to textual tokens, symbolic tokens that facilitate the distinction between the four terms are straightforward. Based on ordering characteristics, they directly delimit the four terms by detecting the order of symbolic tokens in sequences. For example, : and :: delimit the term B, whereas :: and :, in that order, delimit the term C.

3.2. Masking Schemes

In light of the usual notation for analogies $A : B :: C : x$, it is natural to consider the expected solution x as the missing text that can be predicted using the left context. To learn sequential information, we explore three patterns of masking for answered prompts. We present some

examples of masked sequences that result from masking the following sentence analogy to exemplify masking schemes.

he will come tomorrow. : *he will come.* :: *i have no time tomorrow.* : *i have no time.*

Arbitrary Masking (Mask = any span) Like the document corruption method introduced in [17], we randomly mask consecutive tokens whose lengths follow a sampling distribution $\lambda = \textit{Poisson}(3)$. This masking strategy does not take into account the structure of analogies. The starting position of each masking span is selected at random. A masking span can consist of a symbolic token and parts of adjacent terms. Like the following resulted sequence, part of the first two terms are masked off along with the left ratio token.

he will [mask] *will come.* :: *i have no time tomorrow.* : *i have no time.*

One-term Masking (Mask = any term) A masking span is a whole term in analogies, selected from the quadruplet (A, B, C, D). Due to the binding meaning between the four terms in an analogy, each term can be derived from the other three. This scheme randomly masks one of the terms. It allows models to capture more comprehensively analogical regularities. The term C is masked in the following sequence.

he will come tomorrow. : *he will come.* :: [mask] : *i have no time.*

Target-oriented Masking (Mask = term *D*) In this setting, we regard the target prediction as the masking span in analogy prompts. To follow standard notations in analogy, i.e., the format $A : B :: C : x$, we specifically mask the fourth term in answered prompts as shown below.

he will come tomorrow. : *he will come.* :: *i have no time tomorrow.* : [mask]

3.3. Masked Prompt Learning

In general, the paradigm of prompt-based fine-tuning reformulates downstream tasks as masked language modeling on prompts, with the goal of optimizing the prediction of the mask token specified as target outputs [16, 15]. For text generation, we introduce a novel prompt-based fine-tuning paradigm to extract sequential information of answered prompts by masked sequence-to-sequence learning like [18]. On this basis, the analogy task is formulated as masked analogy completion, which aims to generate unknown terms through a sequence-to-sequence model trained for reconstructing the masked span in prompts.

Given a pair of sequences (X, Y) , where X is the sequence of a masked prompt (including a single mask token) obtained by applying a masking scheme to the answered prompt, Y is the target sequence of the masking span. As in regular sequence-to-sequence learning, we tune the model parameters $\Theta = (\theta_{enc}, \theta_{dec})$ to estimate the conditional probabilities of target tokens given masked prompts. The masked prompt is encoded bidirectionally. Each token in the target

sequence is predicted by the autoregressive decoder by maximizing the conditional probability given the input sequence x and its preceding sequence $Y_{<t}$:

$$P\left(Y\left|X\right.\right)=\prod_{t=1}^{|Y|}P\left(Y_t\left|Y_{<t},X;\Theta\right.\right) \tag{1}$$

The loss function for reconstructing a masked prompt is computed as the negative log-likelihood (Equation 2). Our training objective is to minimize the loss function, which is tantamount to maximizing conditional probabilities.

$$L_{\mathrm{m}}\left(X,Y\right)=-\log P\left(Y\left|X\right.\right) \tag{2}$$

The mechanism of masked prompt learning with target-oriented masking scheme (term *D*), resembles few-shot prompt-based fine-tuning for MLMs, which focuses on learning relations between prompts and target outputs of downstream tasks. We will compare the fine-tuning paradigms under different masking schemes in Section 5.3.

4. Formal Analogy between Phrases

Note that finding sentence analogies from a corpus will be difficult unless the corpus is dense. It is easier to find more analogies between small chunks from a corpus than entire sentences. Sentence constituents (i.e., phrases) are structural chunks that play grammatical roles in sentences. We focus on finding formal analogies between phrases, which can indirectly reflect the linguistic regularities contained in the given corpus.

We build analogies from the English part of a parallel corpus[1] released for the news translation task at the Workshop on Machine Translation (WMT20). The data we used is made up of 3,003 sentences with an average length of 25 words. In order to detect phrases, we use Berkeley Neural Parser[2][19] to parse sentences into constituency trees. In each sentence, word sequences between 2 and 6 in length are collected into a phrase pool from which analogies are identified. After traversing all sentences, we obtain 25,310 phrases.

Starting from the set of phrases, we apply some functions from the Nlg tool[3] introduced in [20] to extract analogies. In this work, we explore analogical relations at the formal level. Each phrase is represented as a bag-of-word vector using *Lines2Vectors*. The dimension of vectors is the number of word types in the phrase set. We then run *Vectors2Clusters* to find analogical clusters pertaining to the differences between phrase vectors. Each cluster is made up of pairs of phrases (i.e. ratios) that have the same syntactic transformations. Thus, any two ratios in a cluster makes a syntactic analogy. Note that cluster sizes may vary greatly depending on frequencies of phrase structures contained in the corpus. To alleviate the imbalance of possible analogies, we set the maximum cluster size to 10, where the minimum size defaults to 2.

In our settings, over 1.5 million analogical clusters are extracted, where each cluster contains two phrase ratios in average. By combining every two ratios in a cluster, we are able to enumerate

[1]It can be downloaded from http://www.statmt.org/wmt20/translation-task.html
[2]https://github.com/nikitakit/self-attentive-parser
[3]http://lepage-lab.ips.waseda.ac.jp/en/projects/kakenhi-15k00317/ → Tools - Nlg Module

1,524,293 analogies that capture formal similarities between four distinct phrases. The analogy data includes 17,480 types of phrases with an average length of 3 words (20 characters). It can be roughly estimated that the average length of analogy prompts is $3 \times 4 + 3$ in words (resp. $20 \times 4 + 3$ in characters). For example, a collected phrase analogy *to say* : *want to say* :: *to go out* : *want to go out* indicates a verb phrase attachment with the modifier of the verb *want*, which consists of 15 words including three symbolic tokens in the prompt template.

5. Experiments

5.1. Datasets

We fine-tune LMs on prompts of phrase analogies. To avoid that phrase ratios in the test data appear also in the training set, we split the cluster data before making analogies. We take 1,000 clusters for building the analogies for testing and another 1,000 for validation, while the remaining ones are for training. For each cluster set, we assemble every two ratios in the same cluster into an analogy. The training, validation and test sets consist of approximately 1.5 million, 1,000 and 1,000 phrase analogies respectively.

In addition to the phrase analogy test set, we also explore the performance of Transformer models on solving sentence analogies, which are unseen analogies and have different distributions than the training data. We sample sentence analogies from the English part of bilingual analogies[4] used in an EBMT by analogy setting [21]. The set of analogies embraces formal-level analogous relationships between sentences from the Tatoeba corpus, where the length of sentences varies from 2 to 10.

Even if we swap the ordering of four phrases, each analogy only appears once in all datasets, with no duplicates. The length statistics of analogies are shown in Table 1, including lengths of terms and answered prompts. Analogies in the test sets are processed as prefix prompts, in which the mask token is the last token in sequences. Each model is tested by infilling the unknown term in masked prompts with the default format $A : B :: C :$ [mask].

5.2. Training Details

In terms of the sequence-to-sequence Transformer architecture, we experiment with a pre-trained BART [17]. For computational efficiency, we use the base-size model consisting of a 6-layer bidirectional encoder and a 6-layer autoregressive decoder. The pre-training paradigm of BART is to perform denoising learning on corrupted text, to reconstruct the entire completed sequences. To remove duplicates that to reconstruct the known tokens of masked sequences, we fine-tune BART through masked prompt learning, which reconstructs only the sequences of masking fragments.

We made some modifications to the BART fine-tuning procedure provided by the *Transformers* library [22]. To save computational memory, we conducted partial freezing on the pre-trained BART model and fine-tune only four of the twelve layers, precisely the bottom two layers of the encoder and the top two layers of the decoder. We analyze the discrepancies on fine-tuning

[4]The bilingual set of sentence analogies is the 3rd resource of experimental results at http://lepage-lab.ips.waseda.ac.jp/en/projects/kakenhi-kiban-c-18k11447/.

Table 1
Statistics of span lengths in prompting sequences including four terms and answered prompts in two analogy data: phrase analogies (PA) and sentence analogies (SA). We split over 1.5 million phrases analogies into three sets for training (1.5 million), validation (1,000) and test (1,000). In addition, we also test models on 1,000 sentence analogies which are out-of-distribution cases.

(a) Phrase analogies (PA)

Span	in words	Length in chars
A	2.77±0.00	17.07±0.02
B	3.09±0.00	18.82±0.02
C	3.71±0.00	20.03±0.02
D	4.04±0.00	21.78±0.02
Prompt	16.61±0.01	80.69±0.05

(b) Sentence analogies (SA)

Span	in words	Length in chars
A	5.19±0.11	19.07±0.48
B	4.82±0.09	17.48±0.38
C	5.59±0.11	20.76±0.50
D	5.22±0.11	19.17±0.47
Prompt	23.81±0.32	79.47±1.42

strategies in Section 5.5. In order to alleviate overfitting, we stop the training if there is no improvement on the metric of the validation set after two consecutive epochs. We then save the model with the best performance among the checkpoints.

5.3. Main Results

As for the autoregressive baseline, we explore the performance of the distilled GPT-2 model consisting of 6 layers. The entire pre-trained GPT-2 is fine-tuned on answered prompts of phrase analogies for generating the last term given preceding context. To compare different methods, we employ the accuracy metric to measure the percentage of generations that exactly match the references. In addition, we also compute the Levenshtein distance (including spaces) to measure differences at the character level. Table 3 shows the performance of different fine-tuned models on completing phrase analogies and sentence analogies.

Model Architecture Prompting together with fine-tuning, benefits PLMs to accomplish effective analogy completion of in-distribution cases (phrase analogies). As far as the model architecture is concerned, the autoregressive LM excels in inferring answers to phrase analogy questions where the last term (*D*) is missing. By learning on answered prompts of phrase analogies in a feed-forward fashion, GPT-2 achieves the best accuracy 99.7% on phrase analogies. However, the sequence-to-sequence model fine-tuned for infilling the term *D* performs noticeably worse, only reaching 50.9% in accuracy. Except for the setting of target-oriented masking (term *D*), BART fine-tuned with masked prompt learning have competitive performance with GPT-2.

Masking Scheme The masking deployment of prompts has a large impact on capturing analogical regularities in masked prompt learning for the sequence-to-sequence model. The target-oriented masking scheme (term *D*), like the mechanism of few-shot prompt-based fine-tuning in MLMs, performs the worst in phrase analogies. We posit that it makes BART overly

Table 3
Comparison between the autoregressive baseline and masked prompt learning with different masking schemes over two analogy test sets.

Data	Model	Masking scheme	Acc (%)	Levenshtein distance in chars
PA	GPT-2	-	**99.7**	**0.01±0.01**
	BART	Any span	97.5	0.05±0.03
		Any term	97.0	0.05±0.03
		Term *D*	50.9	0.58±0.07
SA	GPT-2	-	4.2	12.92±0.83
	BART	Any span	11.4	4.28±0.38
		Any term	**44.4**	**2.85±0.29**
		Term *D*	12.0	3.17±0.30

imitate the features of the fourth terms of analogies rather than comprehend analogical relationships. The arbitrary masking (any span) and the one-term masking (any term) are optimal for modeling the sequential information of prompts for phrase analogies, which enables the model to perform well in generating the last phrase in analogies, with only a 0.05 character difference.

Out-of-distribution Generalization As shown in the results of SA in Table 3, the autoregressive LM exhibits poor out-of-distribution generalization capability, although it achieves excellent performance on phrase analogies. GPT-2 can only correctly answer 4.2% unseen sentence analogies where sequences are longer than the training data. This suggests that the autoregressive LM may overfit to the narrow distribution of phrase analogies, which leads to a failure in solving analogies between longer sequences. In comparison, the sequence-to-sequence models exhibit better out-of-distribution generalization. Particularly, BART with similar fine-tuning procedure, predicting the term D conditional on previous tokens, is approximately three times superior to GPT-2.

It is noticeable that masked prompt learning on prompts with one-term masking (any term) has the best generalization on sentence analogies. It is 4 times more accurate than any span masking with competitive performance on phrase analogies. It enables BART to generate sequences that very closely match the reference sentences, differing by only 3 characters on average, about 3/20 of the reference length. Table 5 presents some example errors. The fine-tuned BART model profiting from the bidirectional learning on previous and future tokens, can accomplish effective adaptation to unseen sentence analogies, with the achievement of an order of magnitude greater accuracy than GPT-2.

5.4. Fine-grained Exploration of Analogical Ability

To further explore analogical capabilities of LMs fine-tuned by masked prompt learning, we conduct a fine-grained probing on analogical questions with different formats. For each test analogy, we replace one of the four terms with the mask token to enumerate four analogy questions with different masking spans. We use an individual accuracy Acc_X (where $X \in \{A, B, C, D\}$) to indicate the performance in solving analogies in a specific format. For example,

Table 4
Average accuracy results of fine-tuned BART models with different masking schemes in solving analogy questions in different formats. We report both individual accuracies Acc_X (where $X \in \{A, B, C, D\}$) and universal accuracy Acc_{all} introduced in Subsection 5.4 .

Data	Masking Scheme	Acc (%)				
		A	B	C	D	all
PA	Any span	**99.9**	**99.9**	**99.7**	**97.5**	**97.3**
	Any term	99.8	99.7	**99.7**	97.0	96.5
	Term *D*	0.0	0.0	0.0	50.9	0.0
SA	Any span	10.8	10.6	15.7	11.4	0.6
	Any term	**24.9**	**39.3**	**44.0**	**44.4**	**12.9**
	Term *D*	1.2	0.0	0.0	12.0	0.0

Acc_A is the accuracy of answering masked prompts with the missing term A ([mask] : B :: C : D). Apart from individual accuracies, we also compute the universal score Acc_{all} as Equation 3 to compare the performance for correctly answering all of the four masked prompts for each analogy.

$$\text{Acc}_{\text{all}} = \begin{cases} 1, & \text{if } \text{Acc}_X = 1 \text{ for all } X = A, B, C, D. \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Table 4 shows the accuracy results of fine-tuned BART models on phrase analogies and sentence analogies. For the results of phrase analogies, the masking of any span has the best overall accuracy, with 97.3% of analogies being answered correctly on each term. Among the mask settings that take analogical terms into account, the masking scheme of any term achieves competitive performance in terms of superior accuracy on each individual question, while term-specific masking enables the model to answer only half of the questions where term D is masked. For performance on sentence analogies, fine-tuning for any term masking outperforms substantially other strategies in terms of both individual accuracies and universal accuracy. Concretely, the fine-tuned model performs relatively well on completing sentence analogies where one of the last triplets is missing. It is not a surprise that BART with the target-oriented masking (term *D*) fails to predict terms other than the term *D*. The reason is that the mask token does not appear in any position during the fine-tuning procedure except for the last term.

5.5. Ablation Studies

The masking scheme of any term is more adapted to solving analogies in masked prompt learning. In this subsection , we use one-term masking (any term) and ablate fine-tuning strategies of tuned layers and prompt templates.

Fine-tuned Layers As shown in Table 6, we compare various fine-tuning strategies with the off-the-shelf baseline. In the zero-shot setting, the pre-trained model is not able to generate a reliable answer for analogies beyond words. Fine-tuning BART makes a significant impact on understanding the analogical relationships between sequences. We can see that updating the

Table 5
Examples of analogy results generated by the BART fine-tuned with one-term masking. The underlined text is the difference between the output and the reference.

Input	*he is getting better.* : *tom is getting better.* :: *he doesn 't like to lose.* : [mask]
Output	*tom doesn 't like to lose*
Reference	*tom doesn 't like to lose.*
Input	*i know your name.* : *i believe you.* :: *i don 't know your name.* : [mask]
Output	*don 't believe you.*
Reference	*i don 't believe you.*
Input	*what is going on here?* : *he is intelligent.* :: *what's going on here?* : [mask]
Output	*he is intelligent.*
Reference	*he's intelligent.*
Input	*how did you do this?* : *what did you say?* :: *how do you do this?* : [mask]
Output	*what did you say?*
Reference	*what do you say?*
Input	*he will come tomorrow.* : *he will come.* :: *i have no time tomorrow.* : [mask]
Output	*he will have no time tomorrow.*
Reference	*i have no time.*

entire model helps model phrase analogies, with a significant gain of 94.2 points in predicting the last phrase in analogy questions.

However, it is imprudent to update the entire model. The full-scale fine-tuning makes the model specialized in the training distribution, achieving only 11.7% accuracy in completing sentence analogies. Freezing the entire encoder or decoder degrades the performance of solving phrase analogies, while it increases the performance by at least 12 points over the unfrozen BART for generalizing out-of-distribution analogies. In particular, only tuning the decoder and freeze the encoder is useful to learn phrase distributions, whereas fine-tuning the encoder and freeze the decoder performs relatively better for capturing analogical regularities.

By contrast, lightweight joint fine-tuning on both encoder and decoder performs well on two analogy test sets. Fine-tuning the bottleneck layers (the top two layers of encoder and the bottom two layers of decoder), which closely updates the encoder-decoder attention, achieves accuracy scores of 95.8% and 42.0% on phrase and sentence analogies respectively. Our strategy of fine-tuning only the bottom two layers of the encoder and the top two layers of the decoder, exhibits the best performance, yielding slight gains of about 2 points over fine-tuning the bottleneck layers of BART.

Prompt Templates In Table 7, we list results of GPT-2 and BART learned on prompts with different manually written templates including symbolic prompt, textual prompt and null prompt.[5] As findings in [16], different manually-written templates lead to similar in-distribution accuracy. However, prompt templates behave differently on out-of-distribution cases. We can

[5]We also perform experiments on pre-trained GPT-2 and BART. Regardless of the template, almost none of the analogies can be answered correctly by PLMs in the zero-shot setting.

Table 6
Impact of fine-tuning strategies of fine-tuned layers in the pre-trained sequence-to-sequence model.

Fine-tuned layer		Acc (%)	
Encoder	Decoder	PA	SA
None	None	0.0	0.0
All	All	94.2	11.7
None	All	88.6	23.7
All	None	70.5	28.7
Bottom two	Top two	**97.0**	**44.4**
Top two	Bottom two	95.8	42.0

Table 7
Comparison between different prompt templates: symbolic prompt ($A : B :: C : D$), textual prompt (*A is to B as C is to D*), and null prompt (*A B C D*).

Template	Model	Acc (%)	
		PA	SA
$A : B :: C : D$	GPT-2	**99.7**	4.2
	BART	97.0	**44.4**
A is to B as C is to D	GPT-2	**99.9**	3.6
	BART	99.6	**37.9**
A B C D	GPT-2	**99.9**	0.0
	BART	99.6	0.0

observe that models fine-tuned on simple concatenation of four analogical phrases (null prompt) are not able to answer sentence analogies, although they achieve 99.9% accuracy on phrase analogies. The accuracy of non-null prompts (textual and symbolic templates) is increased by at least 37.9% on sentence analogies.

Despite a subtle drop (1.4 points) in the phrase analogy test, our prompt contributes to better adaptation on unseen analogies than the textual prompt used in [14], attaining an improvement of 4.5 points for BART (0.6 point for GPT-2). This shows the necessity of clear prompt semantics, which allow models to better learn analogical relations encapsulated in prompts.

Regardless of the template, GPT-2 models struggle in completing sentence analogies, although they excel in completing phrase analogies. In contrast, the BART model is over 10 times more accurate than GPT-2 in sentence analogies with the help of non-null prompts.

6. Conclusion

Our work demonstrated the potential of LMs to complete analogies beyond words. We introduced a simple but effective prompt-based fine-tuning paradigm for solving analogies beyond words by masked sequence-to-sequence learning on answered prompts with different masking schemes. To extract useful information about analogical regularities, we proposed three patterns

of masking on answered prompts. We found that fine-tuning with the objective of language modeling on answered prompts, is effective for adaptation of generative analogy completion on phrase analogies, except the sequence-to-sequence framework with target-oriented masking which leads to overfit to narrow features in the training data. Compared to the autoregressive framework, masked prompt learning is beneficial for out-of-distribution generalization over sentence analogies. Lightweight fine-tuning in masked prompt learning with one-term masking has the best potential for learning robust analogical capabilities. In the future, we intend to refine the fine-tuning paradigm to enhance out-of-distribution performance in the few-shot scenario. We hope to apply to other languages and build a multilingual generator for analogies beyond words.

Acknowledgments

The work is supported by China Scholarship Council (CSC) under the CSC Grant No. 202008050136.

References

[1] H. Prade, G. Richard, Analogical proportions: Why they are useful in ai, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 4568–4576. doi:10.24963/ijcai.2021/621, survey Track.

[2] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, Curran Associates Inc., Red Hook, NY, USA, 2013, p. 3111–3119.

[3] O. Levy, Y. Goldberg, Linguistic regularities in sparse and explicit word representations, in: Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, Ann Arbor, Michigan, 2014, pp. 171–180. doi:10.3115/v1/W14-1618.

[4] Z. Bouraoui, S. Jameel, S. Schockaert, Relation induction in word embeddings revisited, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1627–1637. URL: https://aclanthology.org/C18-1138.

[5] S. Lim, H. Prade, G. Richard, Classifying and completing word analogies by machine learning, International Journal of Approximate Reasoning 132 (2021) 1–25. doi:https://doi.org/10.1016/j.ijar.2021.02.002.

[6] S. Alsaidi, A. Decker, P. Lay, E. Marquer, P.-A. Murena, M. Couceiro, A neural approach for detecting morphological analogies, 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA) (2021) 1–10.

[7] Y. Lepage, E. Denoual, Purest ever example-based machine translation: Detailed presentation and assessment, Machine Translation 19 (2005) 251–282. doi:10.1007/s10590-006-9010-x.

[8] B. Elayeb, A. Chouigui, M. Bounhas, O. B. Khiroun, Automatic Arabic text summarization using analogical proportions, Cognitive Computation 12 (2020) 1043–1069. doi:10.1007/s12559-020-09748-y.

[9] A. Diallo, M. Zopf, J. Fürnkranz, Learning analogy-preserving sentence embeddings for answer selection, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 910–919. doi:10.18653/v1/K19-1085.

[10] X. Zhu, G. de Melo, Sentence analogies: Linguistic regularities in sentence embeddings, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 3389–3400. doi:10.18653/v1/2020.coling-main.300.

[11] S. Afantenos, T. Kunze, S. Lim, H. Prade, G. Richard, Analogies between sentences: Theoretical aspects - preliminary experiments, in: J. Vejnarová, N. Wilson (Eds.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer International Publishing, Cham, 2021, pp. 3–18.

[12] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[13] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, arXiv preprint arXiv:2107.13586 (2021).

[14] A. Ushio, L. Espinosa Anke, S. Schockaert, J. Camacho-Collados, BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies?, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3609–3624. doi:10.18653/v1/2021.acl-long.280.

[15] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3816–3830. doi:10.18653/v1/2021.acl-long.295.

[16] R. Logan IV, I. Balazevic, E. Wallace, F. Petroni, S. Singh, S. Riedel, Cutting down on prompts and parameters: Simple few-shot learning with language models, in: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 2824–2835. doi:10.18653/v1/2022.findings-acl.222.

[17] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of

the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. doi:10.18653/v1/2020.acl-main.703.

[18] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, MASS: Masked sequence to sequence pre-training for language generation, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 5926–5936. URL: https://proceedings.mlr.press/v97/song19d.html.

[19] N. Kitaev, D. Klein, Constituency parsing with a self-attentive encoder, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2676–2686. doi:10.18653/v1/P18-1249.

[20] R. Fam, Y. Lepage, Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages, in: LREC 2018, Miyazaki, Japan, 2018. URL: https://www.aclweb.org/anthology/L18-1171.

[21] V. Taillandier, L. Wang, Y. Lepage, Réseaux de neurones pour la résolution d'analogies entre phrases en traduction automatique par l'exemple (neural networks for the resolution of analogies between sentences in EBMT), in: Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles, ATALA et AFCP, Nancy, France, 2020, pp. 108–121. URL: https://aclanthology.org/2020.jeptalnrecital-taln.9.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.