Non-compositional Expression Generation and its Continual Learning

Jianing Zhou and **Suma Bhat**
University of Illinois at Urbana-Champaign
Champaign, IL USA
{zjn1746, spbhat2}@illinois.edu

Abstract

Non-compositional expressions, such as idioms, are an integral part of natural language and their figurative meanings cannot be directly derived from the meanings of their component words. Considering the scenario, where these expressions form a long-tailed process in language, either because of their occurrence in corpora and/or their gradual integration into use over time, this paper studies the ability of contemporary pre-trained language models to continually learn them and generate them. Formulating this as a mask infilling task termed as CLoNE, the study probes the combined challenges of *non-compositionality* and their *continual learning*. Using a set of three diverse idiomatic expression datasets repurposed for this task, we benchmark different large pre-trained language models and different continual learning methods on the task of non-compositional expression generation. Our experiments on the CLoNE task show that pre-trained language models are limited in their ability to generate non-compositional expressions and available continual learning methods are inadequate for our proposed CLoNE task, calling for more effective methods for continual learning of non-compositionality. Our datasets and code will be available at https://github.com/zhjjn/ContinualGeneration.git

1 Introduction

Large language model advancements have provided a new context to study non-compositional expressions (or idiomatic expressions) and the challenges they pose to classical NLP tasks involving them, such as sentiment analysis (Biddle et al., 2020), paraphrase generation (Zhou et al., 2021c), natural language inference (Zeng et al., 2023), metaphor detection (Su et al., 2020; Gong et al., 2020), and contextual disambiguation (Zeng and Bhat, 2021; Zhou et al., 2023a).Furthermore, given the significance of non-compositional expressions in enhancing the naturalness and stylistic richness

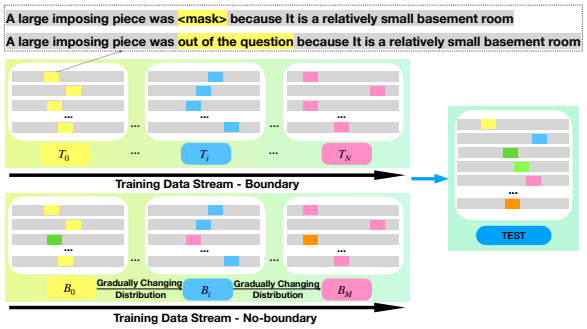


Figure 1: Training with different data streams. Different colors represent different tasks (here, idioms) in their sentences (grey). The top scenario (Boundary) refers to the setting that data for different tasks have clear boundaries and the bottom scenario (No Boundary) refers to the setting that data for different tasks do not have clear boundaries.

of everyday language (Moon et al., 1998; Baldwin and Kim, 2010), their generation by large pre-trained language models (trained without explicitly accounting for non-compositionality) remains an important, yet inadequately explored aspect in NLP (Zhou et al., 2021c).

A related, but distinct inquiry pertains to the continual addition of non-compositional expressions into language (Makkai, 2013). Specifically, it raises the question of how large pre-trained language models learn new non-compositional expressions that were not part of their training data without requiring retraining. Additionally, since the meaning of non-compositional phrases cannot be inferred from the meanings of their component words (Gläser, 1988; Makkai, 2013), and learning the meaning of one non-compositional phrase does not aid in understanding another, each non-compositional phrase must be learned individually. This calls for a continual learning (CL) setting to study large language models' ability to generate non-compositional expressions. Formulating this as a mask infilling task with idiomatic expressions, the study probes the combined challenges of *non-compositionality* and their *continual learning*.

Findings of the Association for Computational Linguistics ACL 2024, pages 2828–2839
August 11-16, 2024

Continual learning, also known as *lifelong learning*, is a machine learning paradigm to adaptively learn new knowledge from a continuous data stream over time while still being able to remember and reuse previously learned knowledge (Robins, 1995). In efforts to mitigate the issue of "catastrophic forgetting" often seen in neural models, there is a rising interest in creating CL algorithms for natural language processing (NLP) tasks (Shu et al., 2016; Xu et al., 2018; Li et al., 2019; Jin et al., 2020) that have shown limited success in handling compositional phrases. This focus of this study is the equally significant challenge of processing non-compositional phrases like idioms and metaphors.

In a recent study, Zhou et al. (2023b) investigated non-compositional expression generation by employing CL over examples organized by dynamically changing difficulty levels, potentially reintroducing previously seen idioms at later stages. In contrast, our approach explores their continual learning and generation by randomly selecting expressions from a pool without replacement. This ensures that expressions encountered earlier in training will not reappear later and better represents their continual learning, focusing on their nuanced learning rather than broader difficulty levels. In this novel setup, we study the extent to which popular CL algorithms alleviate the problem of catastrophic forgetting when learning to generate non-compositional expressions.

Our study is centered around three research questions. **R1**: How difficult is it for language models trained with an MLM objective to learn to generate multi-word non-compositional expressions (which would still be generated word by word)? **R2**: To what extent do large pre-trained language models suffer from catastrophic forgetting when learning to generate non-compositional expressions? and **R3**: To what extent can popular CL algorithms reduce the problem of catastrophic forgetting when learning to generate non-compositional expressions?

Answers to these questions lead to the main contributions of our work summarized below.

1. We propose CLoNE, a new task for studying non-compositional expression generation in a continual learning setting.

2. We construct three datasets, which will be made publicly available upon acceptance, to simulate non-compositional expression acquisition and to probe a set of pretrained language models that yield comprehensive insights about their capabilities to learn non-compositional expressions, also in a CL setting.

3. Comparing CL algorithms with the constructed datasets we find that (a) current large pre-trained language models are limited in their ability to generate non-compositional expressions; (b) current large pre-trained language models are plagued by catastrophic forgetting when learning to generate non-compositional expressions; and (c) SOTA CL algorithms are not beneficial in alleviating catastrophic forgetting for the CLoNE task. Detailed ablation studies and analyses substantiate our claims providing insights about the CLoNE task.

The paper is organized as follows. In Section 3 we explore the first research question, where we fine-tune a set of pre-trained language models to generate idioms and evaluate their performance. Then we explore the second research question in Section 4, where the goal is to understand the forgetting problem for pre-trained language models when learning to generate idioms. To this end, we propose the continual learning of non-compositional expressions (CLoNE) task to measure the ability of pre-trained language models to learn to generate them in a continual manner. This is depicted in Figure 1. Finally, we explore the third research question in Section 5, where using the CLoNE dataset, we explore the extent to which popular CL algorithms alleviate the forgetting problem.

2 Related Works

Non-compositionality. Processing phrases characterized by non-compositionality has been the classical "pain in the neck" for NLP (Sag et al., 2002). Prior work has mainly focused on identifying potentially non-compositional expressions (Salehi et al., 2014; Senaldi et al., 2016; Flor and Klebanov, 2018; Amin et al., 2021; Zeng and Bhat, 2021), disambiguating between their figurative/literal use (Peng and Feldman, 2015; Köper and im Walde, 2016; Liu and Hwa, 2017, 2018; Zhou et al., 2023a), and paraphrasing between non-compositional expressions and their literal counterparts (Liu and Hwa, 2016; Agrawal et al., 2018; Shirin and Raseek, 2018; Zhou et al., 2021a,b). Yet, the task of learning and generating non-compositional expressions—a challenging task because of non-compositionality—remains unexplored and this study aims to fill the research

gap. We focus on pretrained language models' ability to generate non-compositional expressions by requiring them to reconstruct the masked non-compositional expression.

Continual Learning. Continual learning enables models to learn new knowledge and preserve knowledge acquired from a data stream by training only on the newly received data without retraining. Although a natural ability for humans, continual learning remains challenging for neural networks due to the well-known problem of *catastrophic forgetting* (French, 1993). Different methods to alleviate forgetting have been proposed, including those that are rehearsal-based (Robins, 1995; Shin et al., 2017; Aljundi et al., 2019; Pellegrini et al.), regularization-based (Kirkpatrick et al., 2017; Li and Hoiem, 2017; Chaudhry et al., 2018) and architecture-based (Xu and Zhu, 2018; Wen et al., 2019). While originally and primarily studied in computer vision settings (Rebuffi et al., 2017; Zenke et al., 2017; Aljundi et al., 2019), recently they are being studied in the context of NLP, including representation learning (Xu et al., 2018; Liu et al., 2019), language modeling (Sun et al., 2019), classification (Chen et al., 2015; Shu et al., 2016), and generation (Thompson et al., 2019). Yet, the overall number of methods purely designed for NLP problems is still quite limited.

3 Understanding Non-compositional Expression Generation in Pre-trained Language Models

3.1 Task Formulation

To explore large pre-trained language models' ability to generate non-compositional expressions, we train multiple large pre-trained language models to generate idioms, formulated as a mask infilling task. Given a sentence with a masked fragment, the task is to generate an appropriate non-compositional expression to fill the mask. In our setup, we restrict each sentence to contain only one masked fragment and the model is required to fill exactly one non-compositional expression. We create input sentences by selecting a sentence containing a non-compositional expression with a known position and replacing the entire expression with a `<mask>` token. The single `<mask>` token prevents the model from making a decision based on the number of words in the expression. For example, to test if the model is able to successfully generate "*out of the question*," we use the sentence, "It is a relatively small basement room, so a large imposing piece was *out of the question*.", converted to "It is a relatively small basement room, so a large imposing piece was `<mask>`." To account for instances where the right context of the non-compositional expressions could help with their generation, we formulate our task as a mask infill task instead of a causal generation task.

3.2 Data Collection

Our data are constructed from three popular datasets for idiomatic expression usage recognition with labels whether the idiomatic expression was used figuratively or literally: SemEval (Korkontzelos et al., 2013), VNC (Cook et al., 2008) and MAGPIE (Haagsma et al., 2020). We only use the figurative sense so as to ensure the non-compositional property holds. for the mask infilling task, we first mask the non-compositional expression in each sentence. For SemEval and MAGPIE, the original SemEval and MAGPIE datasets provide the position of the non-compositional expression within the sentences, which we use to directly mask the expressions. VNC, on the other hand, provides only the identity of the expression for each sentence without its position. Given that all expressions from VNC are verb-noun compounds, we identify the position of the expressions by first lemmatizing and generating part-of-speech tags for each word in a given sentence and then matching each verb-noun phrase with the given expression[1]. Once the position is identified, we mask the expression. This method finds the position of all the idiomatic verb-noun compounds correctly.

3.3 Experiments

Set up. In these experiments, we follow the classical method of performing stochastic gradient update on a batch of data, which is randomly sampled from the entire training set. Each batch contains data sampled from all tasks use for training over multiple epochs, which we call the **offline** training.

Models. For pre-trained language models, we choose BART, T5, T0 and flan-T5 for our experiments because of their outstanding performance and scale. Besides, we also utilize GPT4 for our experiments in a zero-shot setting.

Evaluation. We resort to *ROUGE* (Lin, 2004) and *BLEU* (Papineni et al., 2002) for evaluation. Besides, *phrase-level BLEU* and *phrase-level ROUGE*

[1]The NLTK lemmatizer and POS tagger were used.

Data	Model	Acc	BLEU		ROUGE	
			Phrase	Total	Phrase	Total
S	**BART**	0.68	87.58	94.40	94.29	96.29
	T5	0.57	83.35	90.29	86.91	93.76
	T0	0.64	90.90	94.43	93.56	96.49
	flan-t5	0.63	86.85	92.81	90.14	94.79
	GPT4	0.37	33.96	63.67	51.27	75.01
V	**BART**	0.37	35.81	69.41	57.90	80.30
	T5	0.40	47.06	76.09	66.20	85.16
	T0	0.46	54.90	79.65	70.37	87.11
	flan-t5	0.49	57.79	81.25	74.07	88.12
	GPT4	0.12	9.00	47.74	23.16	61.92
M	**BART**	0.45	66.01	79.25	71.88	85.93
	T5	0.39	58.92	74.53	65.20	82.57
	T0	0.47	64.54	80.00	72.79	86.27
	flan-t5	0.44	62.30	76.95	69.39	84.07
	GPT4	0.23	33.67	54.59	38.67	67.45

Table 1: Performance of different models in different datasets. **S** refers to the SemEval dataset. **V** refers to the VNC dataset. **M** refers to the MAGPIE dataset.

scores calculated on the target non-compositional expressions and corresponding phrases in the output are used. We also evaluate with a stricter metric of phrase-level *Accuracy*, in which the generated non-compositional expression is counted as correct if and only if every word strictly matches the target expression.

3.4 Results

Table 1 presents the base model's performance in the offline setting. From Table 1, we note that even in the optimal offline setting, the performance of the pretrained large language models is limited, with the highest accuracy of 0.68 on SemEval, 0.49 on VNC and 0.47 on MAGPIE. The performance in terms of the lenient evaluation metrics, i.e., phrase-level BLEU and phrase-level ROUGE, are also limited. This suggests that the current SOTA model struggles to learn to generate non-compositional expressions despite its pre-training on the task. Only for 19 out of the 1528 idiomatic expressions in MAGPIE, the best performing pre-trained language model's accuracy is greater than 0.8. This holds for 10 out of 40 idiomatic expressions in VNC.

Furthermore, we use GPT4 in a zero shot setting by prompting to fill the mask in the given context with an idiomatic expression. As shown in Table 1, we note that GPT4 does not perform this task well.

4 Learning to Generate Non-compositional Expressions and Catastrophic Forgetting

To explore the catastrophic forgetting problem in large pre-trained language models and their abilities to learn and generate non-compositional expressions in a continual way, we propose our CLoNE task and construct the corresponding datasets for experiments.

4.1 The CLoNE Task

The CLoNE task challenges the models' ability (1) to handle *non-compositionality* and (2) to *continual learn*. To simulate a continual learning scenario, we construct two data streams, namely, the *boundary* and the *no-boundary*.

Tasks in CL: In CLoNE, the generation of each non-compositional expression type is regarded as an individual task for the purpose of CL. For example, learning to generate the non-compositional expression *out of the question* is one task while learning to generate *in the bag* is another task.

Data Stream. In the traditional setting (termed *offline training*) of the mask infill task, sentences from different tasks are all shuffled during each training pass. In CL scenarios, a model encounters data streams that are different from this traditional setting in two aspects. First, the model is not permitted to receive multiple passes through the training data. Second, data composition in CL is different from that in the offline setting. During CL, the *boundary* stream groups data by tasks with clear demarcations such that the model is presented with instances pertaining to one task as a batch during training (upper part of Figure 1), while the *no-boundary* stream allows a controlled mixture of data between tasks, thereby permitting a more natural and gradual shift in data distribution from one task to another (lower part of Figure 1).

4.2 Data Stream Construction

Our data streams are constructed from three popular datasets for idiom usage recognition: SemEval (Korkontzelos et al., 2013), VNC (Cook et al., 2008) and MAGPIE (Haagsma et al., 2020). Here we still only use the figurative idioms so as to ensure the non-compositional property. To guarantee large pre-trained language models could learn the non-compositional expressions in our datasets, which makes studying the continual learning of them reasonable, we only choose the non-

Dataset	Model	Accuracy		BLEU				ROUGE				Forget	
				Phrase		Total		Phrase		Total			
		B	N	B	N	B	N	B	N	B	N	B	N
VNC	Vanilla	0.47	0.55	34.90	52.22	66.69	75.63	48.10	78.54	67.89	87.03	0.55	0.68
-CL	Offline (1-p)	0.92		90.62		94.74		93.89		97.30		-	-
MAGPIE	Vanilla	0.58	0.63	72.01	83.45	84.39	89.29	76.75	80.29	89.51	92.66	0.65	0.55
-CL	Offline (1-p)	0.89		96.66		97.99		96.69		98.51		-	-

Table 2: Performance of different settings. **Mem** refers to memory size. **Forget** refers to the forgetting scores. **B** and **N** refer to the performance on the boundary and no-boundary data stream respectively.

compositional expressions on which large pre-trained language models could achieve an accuracy above 0.8 based on our experiments in Section 3. In this way, we make sure the model is able to learn each non-compositional expression in our constructed dataset. It should be noted that only the accuracies on one non-compositional expression in SemEval is higher than 0.8. Therefore, we exclude the SemEval dataset. We call our datasets VNC-CL and MAGPIE-CL.

Boundary Data Stream: In the boundary data stream, data for different tasks have clear boundaries. Therefore, with a set of *tasks* $\mathbb{T}$ where $T_i \in \mathbb{T}$, we randomly generate a permuted sequence of tasks: $\{T_0, T_1, ..., T_N\}$ as the order in which the data of each task arrives in the stream. Then, we group all the data by the tasks: $\mathbf{S}_{boundary} = \{\mathbf{D}^{T_i}\}_{i=0}^{N}$, where $\mathbf{D}^{T_i} = \{d_k^{T_i}\}_{k=0}^{|T_i|}$ refers to all the data of task *i*. $|T_i|$ indicates the number of training examples of task *i*. Finally, the examples of different tasks arrive sequentially with clear boundaries. Here we only consider a random order and leave more systematic orderings to a future study.

No-boundary Data Stream: For preparing the no-boundary stream, we utilize the method proposed in (Jin et al., 2020). We first generated a permuted sequence of tasks: $\{T_0, T_1, ..., T_N\}$ as the order in which the centroid of each task distribution arrives in the no-boundary stream. We assume that each task conforms to a Gaussian distribution. For example, for task *i* the parameters of its Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i)$ are $\mu_i = |T_i|/2 + \sum_{k<i} |T_k|$ and $\sigma_i = |T_i|/2$. Finally, the proposed number of instances are greedily assigned to each data batch to construct the no-boundary data stream. For all three datasets, we used their original train, development, and test splits.

4.3 Experiments

Set up. In this experiment for learning catastrophic forgetting problem, the model takes a single pass over the boundary data stream and the no-boundary data stream without applying any CL algorithm, which is called vanilla CL setting. Different from the offline setting in RQ1, during training the instances will not be shuffled to maintain the order of the tasks.

Models. Based on our experiments in Section 3, we choose the model that is consistently outperforming for each dataset to evaluate their abilities of generating non-compositional expressions in a continual way. Therefore, we choose T0.

Evaluation. Same metrics are used for evaluation. Furthermore, to evaluate models' forgetting problem, we utilize a forgetting score. The forgetting score is calculated as $f = \frac{1}{\mathcal{T}} \sum_{t \in \mathcal{T}} (A_T(D_t) - A_{e_t}(D_t))$ where $e_t = \arg\min_{c_i \in \mathcal{C}} A_{c_i}(D_t)$. $\mathcal{T}$ is the set of all tasks, $\mathcal{C}$ is the set of all checkpoints and D_t represents all test examples of task t. $A_{e_t}(D_t)$ is the averaged accuracy over all test examples of task t at the checkpoint e_t and T is the time step when the training ends. A good method that preserves the knowledge learned in the past will achieve a low forgetting score.

4.4 Results

Based on Table 2, compared to the offline setting, the performance of vanilla CL on both the boundary and no-boundary data streams is 26-45% lower in accuracy, 13.21-55.72 points lower in phrase-level BLEU and 16.4-45.79 points lower in phrase-level ROUGE. This shows that the catastrophic forgetting problem persists in the CLoNE task using both the boundary and the no-boundary streams. Note that the performance gap in sentence-level BLEU (and ROUGE) score between the vanilla CL setting and the offline setting is not as wide as that in phrase-level BLEU (and ROUGE) score. This is because copying context words around the target expression over-inflates the scores, which suggests the need for phrase-level BLEU, ROUGE and Accuracy for evaluation.

Based on forgetting score, we also conclude that the catastrophic forgetting problem persists in the

CLoNE task for all the large pre-trained language models based on both the boundary and the no-boundary streams. We could also see that all the models' performance on boundary stream is worse than their performance on no-boundary stream.

5 RQ3: To what extent can popular CL algorithms alleviate the problem of catastrophic forgetting

Given the conclusion from Section 4 that current large pre-trained language models still suffer from catastrophic forgetting problem when learning to generate non-compositional expressions, we next plan to explore if some traditional and popular CL algorithms could effectively alleviate this forgetting problem in the context of large pre-trained language models and non-compositionality. All the experiments on Section 4 are repeated again with the application of different CL algorithms.

5.1 Methodology

Here we introduce the CL methods used to benchmark for the CLoNE task and study RQ3.

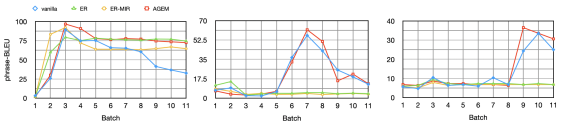
(1) Experience Replay (**ER**) (Robins, 1995) rehearses data from previous tasks by randomly sampling visited examples and then storing them in a buffer of fixed size. These stored examples are later randomly sampled and added into the training set to mitigate catastrophic forgetting.

(2) Experience Replay with Maximally Interfering Retrieval (**ER-MIR**) (Aljundi et al., 2019) provides control over the memory buffer sampling by retrieving the samples that are most negatively influenced by a subsequent update.

(3) Average Gradient Episodic Memory (**AGEM**) (Chaudhry et al., 2018) maintains a memory buffer similar to the ER-based methods. However, AGEM also controls gradient updates by projecting the gradient to a direction of decreasing average loss computed based on the examples in the memory buffer.

5.2 Experiments

Set up. All the settings are same with the settings in the experiments in Section 4 except for the application of different CL algorithms. For VNC-CL, we choose a memory size of 30, whereas for MAGPIE-CL, we use 100. Model settings and evaluation are same with those in Section 4.



(a) Begin (b) Middle (c) End

Figure 2: Comparing of continual learning algorithms, exemplified with performance on tasks in different positions in training data stream. The x-axis is the training examples visited and the y-axis is the phrase-level BLEU.

5.3 Results

To compare the different CL strategies, we see from Table 3 that ER and ER-MIR have little improvement on all the datasets, especially when evaluating with Accuracy, the strictest metric. Even in terms of the more lenient phrase-level BLEU and ROUGE scores, the improvement is still limited and unstable, ranging from 1 to 20 points.

Note that AGEM always outperforms ER and ER-MIR and has consistent improvement over the performance under vanilla continual learning setting. This can be attributed to AGEM's ability to control the gradient update. However, in Table 3, we see that AGEM's improvement is not stable: the improvement ranges from 20-30% in accuracy, sometime even worse than the ER and ER-MIR methods. The forgetting score for AGEM is still very high for both data streams (0.52 and 0.38), which indicates that AGEM is not effective on continual learning of non-compositional expressions. Even with a memory of 50 for the MAGPIE-CL dataset, the performance of AGEM is rather far from the offline performance, which emphasizes the challenges and difficulty posed by the CLoNE.

6 Analysis

6.1 Data Distribution

Boundary vs. No-boundary: The performance on the no-boundary data stream is consistently better than the performance on the boundary data stream across all the datasets, methods and memory sizes. This is expected because the lack of clear task boundaries in the no-boundary data stream helps the model alleviate the forgetting of learned knowledge by reviewing data from different non-compositional expressions in each training batch.

6.2 Influential Factors

Here we provide an analysis of different factors that influence the performance, including the length of the idioms, the number of training examples

Dataset	Model	Mem	Accuracy		BLEU				ROUGE				Forget	
					Phrase		Total		Phrase		Total			
			B	N	B	N	B	N	B	N	B	N	B	N
VNC -CL	**Vanilla**	/	0.47	0.55	34.90	52.22	66.69	75.63	48.10	78.54	67.89	87.03	0.55	0.68
	ER	10	0.57	0.67	49.95	63.95	73.88	81.50	59.44	75.86	82.82	90.00	0.44	0.63
	ER-MIR	10	0.59	0.68	50.62	65.09	75.61	83.65	61.67	78.07	84.73	91.55	**0.43**	**0.60**
	AGEM	10	**0.78**	**0.85**	**72.01**	**82.54**	**85.41**	**92.98**	**79.44**	**85.26**	**91.71**	**94.28**	0.49	0.62
MAGPIE -CL	**Vanilla**	/	0.58	0.63	72.01	83.45	84.39	89.29	76.75	80.29	89.51	92.66	0.65	0.55
	ER	50	**0.79**	0.81	93.68	86.28	95.22	93.75	91.23	90.14	96.95	95.60	0.56	0.44
	ER-MIR	50	**0.79**	0.82	**93.83**	87.19	**95.70**	94.88	**91.40**	91.31	**97.29**	96.54	0.55	0.41
	AGEM	50	0.78	0.83	89.63	**91.93**	92.38	**96.09**	87.01	**92.96**	95.37	**97.29**	**0.52**	**0.38**

Table 3: Performance of different settings.

Model	Memory	Pos.	Num.	Com.	Length.
Vanilla	-	0.18	19.16	0.05	0.02
ER	50	-0.62	18.11	-0.02	-0.05
ER-MIR	50	-0.58	15.14	-0.01	0.01
AGEM	50	-0.24	37.28	-0.05	0.005

Table 4: Coefficients for different features. **Pos.** represents tasks' position. **Num.** refers to the number of training examples of tasks. **Com.** refers to the degree of compositionality of different tasks. **Length.** refers to the length of the target idiom.

of a task and the degree of compositionality of a non-compositional expression. A linear regression model is trained given the above features as input and accuracy as output for every non-compositional expression. The coefficients of different features under different settings are presented in Table 4. To provide enough samples to train such a linear regression model, we only use the MAGPIE-CL dataset because it contains the most non-compositional expressions.

For number of training examples, the coefficients under different setting are consistently positive, which is easy to understand because with more examples the performance would get better. For degree of compositionality, we use the ratio of the number of a non-compositional expression's idiomatic examples and the number of its literal examples to represent its degree of compositionality. Table 4 shows that the coefficients for degree of compositionality under different settings are consistently low, which means that the performance is not related to the degree of compositionality.

Under vanilla continual learning setting, it should be noting that the coefficient of tasks' position is positive, which means that for tasks in the later stage of training the model would have a better performance whereas the performance for tasks in the earlier stage is worse due to the forgetting problem as expected. After different continual learning methods have been utilized, it is obvious that the coefficients for taskscg' position are all negative because of the use of memory in these continual learning methods. The use of memory will force the model to review the training examples and tasks in the earlier stages, which therefore strengthens the performance on these earlier tasks. However, a successful continual learning method should not only help the model to memorize the earlier tasks but also avoid affecting the learning of later tasks negatively. Therefore, the coefficient for tasks' position should be negative and also as close to zero as possible.

6.3 Attention

Here we turn to focusing on the attentions in the large pre-trained language models to analyze their poor abilities of learning to generate idioms. The non-compositionality of idioms depends on the context – e.g. compare "in culinary school, I felt at sea" to "the sailors were at sea". Within Transformer, contextualisation of input tokens is achieved through the attention mechanisms, which is why they are expected to combine the representations of the idioms' tokens and embed the idiom in its context. This section discusses the impact of non-compostional expressions on the encoder-decoder cross-attention and decoder self-attention. We choose T0 as our model.

To analyze the attentions, we decode generation with beam size 5 and then extract the decoder's self-attention weights and the encoder-decoder cross-attention weights for those generation. Attention distribution for the weights that go from words in the generated phrase to context and the other words in the generated phrase and the attention distribution for the weights that go from the generated phrase to mask token and the context are presented. There is no significant difference between correctly generated phrase and wrongly generated phrase. As is shown in Table 6, the model always tends

Input	Generation	Truth
Everything is too loud : The amplification , my guitar , the crowd 's cheery conversation . It 's not easy . Martin takes them on unamplified , and after about an hour 's struggle , <mask> handsomely ...	gets stuck in	wins the day
... it contains only a few genuinely brilliant minds, and fewer still who are likely to <mask>.	make the grade	rock the boat
... should instead of feeling the need to get yourself noticed. Do not deliberately <mask> ...	steal the show	keep a low profile

Table 5: One example of generated output.

Generation	Decoder		EncDec	
	Context	Phrase	Context	Mask
Correct	0.82	0.18	0.98	0.02
Wrong	0.81	0.19	0.98	0.02
All	0.81	0.19	0.98	0.02

Table 6: Attention distribution for the weights of both self-attention in the decoders and cross-attention between the encoder-decoder. Correct, Wrong and All refer to the attention distribution for the weights of the correct generation, wrong generation and all the generation results respectively.

to focus more on the context instead of the words in the generated phrase on the decoder side or the mask token on the encoder side.

Therefore, we manually analyzed 100 wrongly generated examples to further study the attentions. According to our manual analysis, the poor ability of large pre-trained language models to generate non-compositional expression is due to the over-emphasis on the single token and nearby tokens and ignorance on the whole semantic meaning of the context. Among the 100 wrongly generated results, 63 of them present the attention weights that focus more on some single tokens that could trigger the generation of some related idioms. Besides, 20 wrongly generated results present the attention weights that focus more on the nearby tokens but ignore the overall semantics. The remaining 17 wrongly generated results present the attention weights that focus more on both some single tokens and nearby tokens. For example, as shown in Table 5, the model wrongly generates 'gets stuck in' due to the over-emphasis of attention on the single word 'struggle', which causes the wrong generation result.

6.4 Insights into CLoNE

It should be noted that Jin et al. (2020) found that for CL of simple compositional phrases ER method performed better than ER-MIR and AGEM, which is in stark contrast with what we observed in our CLoNE task. This suggests that non-compositionality poses a challenge not only to traditional language learning tasks (e.g., sentiment analysis (Hwang and Hidey, 2019) and machine translation (Fadaee et al., 2018)) but also to learn them in a continual manner.

Based on our experiments and analysis on the attention patterns, it should be noted that current pre-trained language models still heavily rely on single words when processing non-compositional expressions. However, to fully understand and generate non-compositional expressions, it is necessary for them to refer to the semantic meanings of the whole sentences.

In addition to evaluating the performance of different CL methods, the CLoNE task could also be used to evaluate different models' ability to handle non-compositionality. One could hypothesize that current SOTA models' reliance on the linear combination of word representations to infer the meaning of phrases and generate them, may be insufficient for the CLoNE task due to the inherent non-compositionality. More importantly, given that non-compositional expressions cannot be learned by transferring previously learned knowledge (again due to non-compositionality), CLoNE presents definite challenges for CL methods to alleviate the catastrophic forgetting problem. This calls for more advanced models for learning non-compositionality and more effective CL methods for alleviating forgetting are both required.

7 Conclusion

In this study, we propose CLoNE, the first task focusing on generation of non-compositional expressions and their continual learning, whose challenges are non-compositionality and continual learning. Benchmarks and analysis of SOTA large pre-trained language models and continual learning methods provide exploration towards our proposed three research questions. Our experiments show that even the most SOTA large pre-trained language model struggles with non-compositional expression generation and still suffers from catastrophic forgetting problem. Furthermore, even the

best-performing CL method still struggles with continual learning of non-compositionality. The gap between the best performance in a CL setting and the upper-bound performance in an offline setting calls for more advanced models for learning non-compositionality and more effective CL algorithms.

8 Limitations

This study can be expanded in many ways. First, enlarging our CL datasets with other types of IEs beyond idioms (e.g., phrasal verbs) and studying their effect would shed more light into the process. Second, we only consider a random order for the tasks in different data streams, leaving a potentially more systematic ordering that takes the properties of non-compositional expressions into consideration, such as their frequency, for future explorations. Lastly, improved methods that address the drastic change of important weights for seen tasks in neural networks would extend this work.

It should be noted that the automatic evaluation metrics we used are only based on n-gram overlap, which is not the most appropriate evaluation metrics for non-compositional expression generation because there are cases where non-compositional expressions with different words share the same meaning. Besides, other evaluation metrics based on semantic meaning such as BERTScore are also inappropriate because they rely on the combined contextual embeddings to represent the semantic meaning whereas the meanings of non-compositional expressions are not the sum of their parts. Therefore, future works should also explore the more appropriate evaluation metrics for non-compositional expression generation.

Acknowledgements

This research was supported by the National Science Foundation under Grant No. IIS 2230817 and in part by a U.S. National Science Foundation and Institute of Education Sciences under Grant No. 2229612.

References

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, and Dipti Misra Sharma. 2018. No more beating about the bush: A step towards idiom handling for indian language nlp. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. 2019. Online continual learning with maximal interfered retrieval. Advances in Neural Information Processing Systems, 32:11849–11860.

Miriam Amin, Peter Fankhauser, Marc Kupietz, and Roman Schneider. 2021. Data-driven identification of idioms in song lyrics. MWE 2021, page 13.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. Handbook of natural language processing, 2:267–292.

Rhys Biddle, Aditya Joshi, Shaowu Liu, Cecile Paris, and Guandong Xu. 2020. Leveraging sentiment distributions to distinguish figurative from literal health reports on Twitter. In Proceedings of The Web Conference 2020, pages 1217–1227.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2018. Efficient lifelong learning with a-gem. In International Conference on Learning Representations.

Zhiyuan Chen, Nianzu Ma, and Bing Liu. 2015. Lifelong learning for sentiment classification. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 750–756.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset. In Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), pages 19–22.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Michael Flor and Beata Beigman Klebanov. 2018. Catching idiomatic expressions in efl essays. In Proceedings of the Workshop on Figurative Language Processing, pages 34–44.

Robert M French. 1993. Catastrophic interference in connectionist networks: Can it be predicted, can it be prevented? In Proceedings of the 6th International Conference on Neural Information Processing Systems, pages 1176–1177.

Rosemarie Gläser. 1988. The grading of idiomaticity as a presupposition for a taxonomy of idioms. Understanding the lexicon, pages 264–279.

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. 2020. Illinimet: Illinois system for metaphor

detection with contextual and linguistic information. In Proceedings of the Second Workshop on Figurative Language Processing, pages 146–153.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 279–287.

Alyssa Hwang and Christopher Hidey. 2019. Confirming the non-compositionality of idioms for sentiment analysis. In Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 125–129, Florence, Italy. Association for Computational Linguistics.

Xisen Jin, Junyi Du, Arka Sadhu, Ram Nevatia, and Xiang Ren. 2020. Visually grounded continual learning of compositional phrases. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2018–2029.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521–3526.

Maximilian Köper and Sabine Schulte im Walde. 2016. Distinguishing literal and non-literal usage of german particle verbs. In Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies, pages 353–362.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. Semeval-2013 task 5: Evaluating phrasal semantics. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 39–47.

Yuanpeng Li, Liang Zhao, Kenneth Church, and Mohamed Elhoseiny. 2019. Compositional language continual learning. In International Conference on Learning Representations.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence, 40(12):2935–2947.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81.

Changsheng Liu and Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 363–373.

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. In Thirty-First AAAI Conference on Artificial Intelligence.

Changsheng Liu and Rebecca Hwa. 2018. Heuristically informed unsupervised idiom usage recognition. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1723–1731.

Tianlin Liu, Lyle Ungar, and João Sedoc. 2019. Continual learning for sentence representations using conceptors. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3274–3279.

Adam Makkai. 2013. Idiom structure in English, volume 48. Walter de Gruyter.

Rosamund Moon et al. 1998. Fixed expressions and idioms in English: A corpus-based approach. Oxford University Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.

Lorenzo Pellegrini, Gabriele Graffieti, Vincenzo Lomonaco, and Davide Maltoni. Latent replay for real-time continual learning. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 10203–10209. IEEE.

Jing Peng and Anna Feldman. 2015. Automatic idiom recognition with word embeddings. In Information Management and Big Data, pages 17–29. Springer.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5533–5542. IEEE Computer Society.

Anthony Robins. 1995. Catastrophic forgetting, rehearsal and pseudorehearsal. Connection Science, 7(2):123–146.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In International conference on intelligent text processing and computational linguistics, pages 1–15. Springer.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Detecting non-compositional mwe components using wiktionary. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1792–1797.

Marco Silvio Giuseppe Senaldi, Gianluca E Lebani, and Alessandro Lenci. 2016. Lexical variability and compositionality: Investigating idiomaticity with distributional semantic models. In Proceedings of the 12th workshop on multiword expressions, pages 21–31.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 2994–3003.

A Fathima Shirin and C Raseek. 2018. Replacing idioms based on their figurative usage. In 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR), pages 1–6. IEEE.

Lei Shu, Bing Liu, Hu Xu, and Annice Kim. 2016. Lifelong-rl: Lifelong relaxation labeling for separating entities and aspects in opinion targets. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 225–235.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In Proceedings of the second workshop on figurative language processing, pages 30–39.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2019. Lamol: Language modeling for lifelong language learning. In International Conference on Learning Representations.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2062–2068.

Yeming Wen, Dustin Tran, and Jimmy Ba. 2019. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In International Conference on Learning Representations.

Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2018. Lifelong domain word embedding via meta-learning. In IJCAI.

Ju Xu and Zhanxing Zhu. 2018. Reinforced continual learning. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pages 907–916.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. Transactions of the Association for Computational Linguistics, 9:1546–1562.

Ziheng Zeng, Kellen Cheng, Srihari Nanniyur, Jianing Zhou, and Suma Bhat. 2023. Iekg: A commonsense knowledge graph for idiomatic expressions. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 14243–14264.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In International Conference on Machine Learning, pages 3987–3995. PMLR.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021a. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021), pages 33–48, Online. Association for Computational Linguistics.

Jianing Zhou, Hongyu Gong, Srihari Nanniyur, and Suma Bhat. 2021b. From solving a problem boldly to cutting the gordian knot: Idiomatic text generation. arXiv preprint arXiv:2104.06541.

Jianing Zhou, Ziheng Zeng, and Suma Bhat. 2023a. Clcl: Non-compositional expression detection with contrastive learning and curriculum learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 730–743.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2021c. Idiomatic expression paraphrasing without strong supervision. arXiv preprint arXiv:2112.08592.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2023b. Non-compositional expression generation based on curriculum learning and continual learning. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 4320–4335.

Dataset	Train Size	Dev Size	Test Size	# Tasks
VNC-CL	285	0	60	10
MAGPIE-CL	1230	157	131	19

Table 7: Statistics of our constructed datasets.

A Dataset

The statistis of our constructed dataset is presented in Table 7.

B Prompt

We provide the masked sentence to GPT4 and ask it to fill in the mask with an idiom. The complete prompt is as follows:

Fill in the mask in the given sentences with an idiom. Please return the whole given sentence with <mask> replaced by an idiom. Sentence: [sentence]

We also provide the format of the input to other models:

Fill in the mask in the given sentence with an idiom: [sentence]