Continual Lifelong Learning in Natural Language Processing: A Survey

Magdalena Biesialska **Katarzyna Biesialska** **Marta R. Costa-jussà**
Universitat Politècnica de Catalunya, Barcelona, Spain
{magdalena.biesialska,katarzyna.biesialska,marta.ruiz}@upc.edu

Abstract

Continual learning (CL) aims to enable information systems to learn from a continuous data stream across time. However, it is difficult for existing deep learning architectures to learn a new task without largely forgetting previously acquired knowledge. Furthermore, CL is particularly challenging for language learning, as natural language is ambiguous: it is discrete, compositional, and its meaning is context-dependent. In this work, we look at the problem of CL through the lens of various NLP tasks. Our survey discusses major challenges in CL and current methods applied in neural network models. We also provide a critical review of the existing CL evaluation methods and datasets in NLP. Finally, we present our outlook on future research directions.

1 Introduction

Human beings learn by building on their memories and applying past knowledge to understand new concepts. Unlike humans, existing neural networks (NNs) mostly learn in isolation and can be used effectively only for a limited time. Models become less accurate over time, for instance, due to the changing distribution of data – the phenomenon known as *concept drift* (Schlimmer and Granger, 1986; Widmer and Kubat, 1993). With the advent of deep learning, the problem of continual learning (CL) in Natural Language Processing (NLP) is becoming even more pressing, as current approaches are not able to effectively retain previously learned knowledge and adapt to new information at the same time.

Throughout the years, numerous methods have been proposed to address the challenge known as *catastrophic forgetting* (CF) or *catastrophic interference* (McCloskey and Cohen, 1989). Naïve approaches to mitigate the problem, such as retraining the model from scratch to adapt to a new task (or a new data distribution), are costly and time-consuming. This is reinforced by the problems of *capacity saturation* and *model expansion*. Concretely, a parametric model, while learning data samples with different distributions or progressing through a sequence of tasks, eventually reaches a point at which no more knowledge can be stored – i.e. its representational capacity approaches the limit (Sodhani et al., 2020; Aljundi et al., 2019). At this point, either model's capacity is expanded, or a selective forgetting – which likely incurs performance degradation – is applied. The latter choice may result either in a deterioration of prediction accuracy on new tasks (or data distributions) or forgetting the knowledge acquired before. This constraint is underpinned by a defining characteristic of CL, known as the *stability-plasticity dilemma*. Specifically, the phenomenon considers the model's attempt to strike a balance between its *stability* (the ability to retain prior knowledge) and its *plasticity* (the ability to adapt to new knowledge).

CL in the NLP domain, as opposed to computer vision or robotics, is still nascent (Greco et al., 2019; Sun et al., 2020). The differences are reflected in the small number of proposed methods aiming to alleviate the aforementioned issues and the evaluation benchmarks. To the best of our knowledge, apart from the work of Chen and Liu (2018), our paper is the only study summarizing the research progress related to continual, lifelong learning in NLP.

Proceedings of the 28th International Conference on Computational Linguistics, pages 6523–6541
Barcelona, Spain (Online), December 8-13, 2020

2 Learning Paradigms

In this section, we discuss principles of CL and related machine learning (ML) paradigms, as well as contemporary approaches to mitigate CF.

2.1 Continual Learning

Continual learning[1] (Ring, 1994) is a machine learning paradigm, whose objective is to adaptively learn across time by leveraging previously learned tasks to improve generalization for future tasks. Hence, CL studies the problem of sequential learning from a continuous stream of data, drawn from a potentially non-stationary distribution, and reusing gained knowledge throughout the lifetime while avoiding CF.

More formally: the goal is to sequentially learn a model $f : \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{Y}$ from a large number of tasks $\mathcal{T}$. The model is trained on examples (x_i, y_i), such that: $x_i \in \mathcal{X}_{t_i}$ is an input feature vector, $y_i \in \mathcal{Y}_{t_i}$ is a target vector (e.g. a class label), and $t_i \in \mathcal{T}$ denotes a task descriptor (in the simplest case $t_i = i$) where $i \in \mathbb{Z}$. The objective is to maximize the function f (parameterized by $\theta \in \mathbb{R}$) at the task $\mathcal{T}_i$, while minimizing CF for tasks $\mathcal{T}_1 \ldots \mathcal{T}_{i-1}$.

Although the above-mentioned definitions of CL may seem fairly general, there are certain desired properties, which are summarized in Table 1.

Property	*Definition*
Knowledge retention	The model is not prone to catastrophic forgetting.
Forward transfer	The model learns a new task while reusing knowledge acquired from previous tasks.
Backward transfer	The model achieves improved performance on previous tasks after learning a new task.
On-line learning	The model learns from a continuous data stream.
No task boundaries	The model learns without requiring neither clear task nor data boundaries.
Fixed model capacity	Memory size is constant regardless of the number of tasks and the length of a data stream.

Table 1: Desiderata of continual learning.

In practice, current CL systems often relax at least one of the requirements listed in Table 1. Most methods still follow the off-line learning paradigm – models are trained using batches of data shuffled in such a way as to satisfy the independent and identically distributed (i.i.d.) assumption. Consequently, many models are trained solely in a supervised fashion with large labeled datasets, and thus they are not exposed to more challenging situations involving few-shot, unsupervised, or self-supervised learning. Additionally, existing approaches often fail to restrict themselves to make a single pass over the data, and this entails longer learning times. Moreover, the number of tasks as well as their identity are frequently known to the system from the outset.

2.2 Related Machine Learning Paradigms

Traditionally, many ML models are designed to be trained for merely a single task. However, it has been proven that transferring knowledge learned from one task and applying it to another task is a powerful mechanism for NNs. In many respects, CL bears some resemblance to other dominant learning approaches. Therefore, in this section, we draw connections between various ML paradigms. We provide an overview of the approaches, and in particular, we shed light on the shared principles as well as on the aspects that make CL different from other ML paradigms (see Table 2).

In principle, we assume that the ability of a model to generalize can be considered one of its most important characteristics. Importantly, if tasks are related, then knowledge transfer between tasks should lead to a better generalization and faster learning (Lopez-Paz and Ranzato, 2017; Sodhani et al., 2020). Therefore, we compare the paradigms taking into account how well they are able to leverage an inductive bias. Specifically, positive backward transfer improves the performance on old tasks, while negative

[1]Continual learning in the literature is also referred to as: *lifelong learning* (Silver and Mercer, 2002; Silver et al., 2013; Chen and Liu, 2018; Chaudhry et al., 2019a; Parisi et al., 2019; Aljundi et al., 2017), *incremental learning* (Solomonoff, 1989; Chaudhry et al., 2018), *sequential learning* (McCloskey and Cohen, 1989; Shin et al., 2017; Aljundi et al., 2019), *explanation-based learning* (Thrun, 1996), and *never-ending learning* (Carlson et al., 2010).

Paradigm	*Definition*	*Properties**	*Related works*
Transfer learning	Transferring knowledge from a source task/domain to a target task/domain to improve the performance of the target task.	+ forward transfer – no backward transfer – no knowledge retention – task boundaries – off-line learning	(Pan and Yang, 2010) (Howard and Ruder, 2018) (Peters et al., 2018) (Radford et al., 2019) (Devlin et al., 2019) (Houlsby et al., 2019) (Raffel et al., 2020)
Multi-task learning	Learning multiple related tasks jointly, using parameter sharing, to improve the generalization of all the tasks.	+ positive transfer – negative transfer – task boundaries – off-line learning	(Caruana, 1997) (Zhang and Yang, 2017) (Ruder, 2017) (McCann et al., 2018) (Stickland and Murray, 2019) (Phang et al., 2020)
Meta-learning	*Learning to learn.* Learning generic knowledge, given a small set of training examples and numerous tasks, and quickly adapting to a new task.	+ forward transfer – no backward transfer – no knowledge retention – off-line learning	(Thrun and Pratt, 1998) (Finn et al., 2017) (Xu et al., 2018) (Obamuyide and Vlachos, 2019) (Hospedales et al., 2020) (Beaulieu et al., 2020)
Curriculum learning	Learning from training examples arranged in a meaningful order – task or data difficulty gradually increases.	+ forward transfer + backward transfer + knowledge retention – task boundaries – off-line learning	(Elman, 1993) (Bengio et al., 2009) (van der Wees et al., 2017) (Zhang et al., 2018) (Zhang et al., 2019) (Platanios et al., 2019) (Ruiter et al., 2020)
On-line learning	Learning over a continuous stream of training examples provided in a sequential order. Experiences *concept drift* due to non-i.i.d. data.	+ on-line learning + forward transfer – no backward transfer – no knowledge retention – single task/domain	(Bottou, 1999) (Bottou and LeCun, 2004) (Cesa-Bianchi and Lugosi, 2006) (Shalev-Shwartz, 2012) (C. de Souza et al., 2015) (Hoi et al., 2018)
On-the-job learning	Discovering new tasks, learning and adapting on-the-fly. On-the-job learning operates in an *open-world* environment, and it involves interaction with humans and the environment. It belongs to the CL family of methods.	+ on-line learning + forward transfer + backward transfer + knowledge retention + no task boundaries + open-world learning – interactive learning	(Xu et al., 2019) (Mazumder et al., 2019) (Liu, 2020)

Table 2: Comparison of related ML paradigms. * Properties aligned (+) and unaligned (–) with CL.

backward transfer deteriorates the performance on previous tasks (if high, it enables CF). Similarly, negative forward transfer impedes learning of new concepts, while positive forward transfer allows to learn a new task with just a few examples (if high, it enables zero-shot learning).

2.3 Approaches to Continual Learning

The majority of existing CL approaches tend to apply a single model structure to all tasks (Li et al., 2019) and control CF by various scheduling schemes. We distinguish three main families of methods: *rehearsal*, *regularization*, and *architectural* as well as a few hybrid categories. Importantly, the number of models originating purely from the NLP domain is quite limited.

Rehearsal methods rely on retaining some training examples from prior tasks, so that they can later be shown to a task at hand. Rebuffi et al. (2017b) proposed the most well-known method for incremental class learning, i.e. the iCaRL model. Furthermore, as training samples are kept per each task and are periodically replayed while learning the model, the computing and memory requirements of the model increase proportionally to the number of tasks. To reduce storage, it is advised to use either latent replay (Pellegrini et al., 2019) or pseudo-rehearsal (Robins, 1995) methods.

Pseudo-rehearsal methods are a sub-group of rehearsal methods. Instead of using training samples from memory, pseudo-rehearsal models generate examples by knowing the probability distributions of previous task samples. Notable approaches include a generative autoencoder (FearNet, Kemker and Kanan, 2018) and a model based on Generative Adversarial Networks (DGR, Shin et al., 2017).

Regularization methods are single-model approaches that rely on a fixed model capacity with an additional loss term that aids knowledge consolidation while learning subsequent tasks or data distributions. For instance, Elastic Weight Consolidation (EWC, Kirkpatrick et al., 2016) reduces forgetting by

regularizing the loss; in other words, it slows down the learning of parameters important for previous tasks.

Memory methods are a special case of regularization methods that can be divided into two groups: *synaptic regularization* (Zenke et al., 2017; Kirkpatrick et al., 2016; Chaudhry et al., 2018) and *episodic memory* (Li and Hoiem, 2016; Jung et al., 2016; Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2019b; d'Autume et al., 2019). The former methods are focused on reducing interference with the consolidated knowledge by adjusting learning rates in a way that prevents changes to previously learned model parameters. While the latter store training samples from previously seen data, which are later rehearsed to allow learning new classes. Importantly, Gradient Episodic Memory (GEM, Lopez-Paz and Ranzato, 2017) allows positive backward transfer and prevents the loss on past tasks from increasing. Other notable examples of this approach include A-GEM (Chaudhry et al., 2019a), MER (Riemer et al., 2018), or a method originating from NLP - MBPA++ (d'Autume et al., 2019).

Knowledge distillation methods bear a close resemblance to *episodic memory* methods, but unlike GEM they keep the predictions at past tasks invariant (Rebuffi et al., 2017b; Lopez-Paz and Ranzato, 2017). In particular, it is a class of methods alleviating CF by relying on knowledge transfer from a large network model (teacher) to a new, smaller network (student) (Hinton et al., 2015). The underlying idea is that the student model learns to generate predictions of the teacher model. As demonstrated in Kim and Rush (2016) and Wei et al. (2019), knowledge distillation approaches can prove especially suitable for neural machine translation models, which are mostly large, and hence reduction in size is beneficial.

Architectural methods prevent forgetting by applying modular changes to the network's architecture and introducing task-specific parameters. Typically, previous task parameters are kept fixed (Rusu et al., 2016; Mancini et al., 2018) or masked out (Serra et al., 2018; Mallya and Lazebnik, 2018). Moreover, new layers are often injected dynamically to augment a model with additional modules to accommodate new tasks. Progressive Networks (PNN, Rusu et al., 2016), and their improved versions: Dynamically Expandable Network (DEN, Yoon et al., 2018), Reinforced Continual Learning (RCL, Xu and Zhu, 2018), are prominent examples. The main drawback of such strategies is the substantially growing number of parameters. Similar to PNN, BatchEnsemble (Wen et al., 2020) is also immune to CF, in addition it supports parallel order of tasks and consumes less memory than PNN thanks to training only fast weights. In a similar vein, adapter modules aim to overcome the problem of a large number of parameters. They act as additional network layers with a small number of parameters (Rebuffi et al., 2017a; Houlsby et al., 2019; Bapna and Firat, 2019; Pfeiffer et al., 2020) that reconfigure the original network on-the-fly for a target task, while keeping the parameters of the original network untouched and shared between different tasks.

3 Evaluation

Even though CL is now experiencing a surge in the number of proposed new methods, there is no unified approach when it comes to their evaluation using benchmark datasets and metrics (Parisi et al., 2019). And as we will show in this section, this is especially true in the NLP domain. There is a scarcity of datasets and benchmark evaluation schemes available specifically for CL in NLP.

3.1 Protocols

Basically, researchers often focus on evaluating the *plasticity* (generalization) side and the *stability* (consistency) side of the model. Various protocols and methodologies for CL method evaluation have been devised throughout the years (e.g. Kemker et al., 2017; Serra et al., 2018; Sodhani et al., 2020; Pfülb and Gepperth, 2019; Chaudhry et al., 2019a); however, many of them suffer from deficiencies such as small datasets or a limited number of evaluated methods, to name a few.

Furthermore, as observed by Chaudhry et al. (2019a), the prevalent learning protocol followed in many CL research efforts stems from supervised learning, where many passes over the data of each task are performed. The authors claimed that in a CL setting this approach is flawed as with more passes over the data of a given task, the model degrades more because it forgets previously acquired knowledge.

In a similar vein, Yogatama et al. (2019) contended that NLP models are predominantly evaluated with respect to their performance on a held-out test set, which is measured after the training is done for a given task. Therefore, Chaudhry et al. (2019a) introduced a learning protocol that, according to the authors, is more suitable for CL as it satisfies the constraint of a single pass over the data, which is motivated by the need for a faster learning process. Another recent approach, proposed by d'Autume et al. (2019), relies on a sequentially presented stream of examples derived from various datasets in one pass, without revealing dataset boundary or identity to the model.

3.2 Benchmarks and Metrics

For years the NLP domain has lagged behind computer vision and other ML areas (e.g. Kirkpatrick et al., 2016; Zenke et al., 2017; Lomonaco and Maltoni, 2017; Rebuffi et al., 2017b) when it comes to the availability of standard CL-related benchmarks (Greco et al., 2019; Wang et al., 2019b). However, the situation has slightly improved recently with an introduction of a handful of multi-task benchmarks. In particular, GLUE (Wang et al., 2018; Greco et al., 2019) and SUPERGLUE (Wang et al., 2019a) benchmarks track performance on eleven and ten language understanding tasks respectively, using existing NLP datasets. Along the same line, McCann et al. (2018) presented the Natural Language Decathlon (DECANLP) benchmark for evaluating the performance of models across ten NLP tasks. The decathlon score (decaScore) is an additive combination of various metrics specific for each of the ten selected tasks (i.e. the normalized F1 metric, BLEU and ROUGE scores, among others). Similar to DECANLP, a recently proposed Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark (Hu et al., 2020) also uses a diverse set of NLP tasks and task-specific measures to evaluate the performance of cross-lingual transfer learning. XTREME consists of nine tasks derived from four different categories and uses zero-shot cross-lingual transfer with the English language as the source language for evaluation.

In principle, CL models should not only be evaluated against traditional performance metrics (such as model accuracy); it is also important to measure their ability to reuse prior knowledge. Similarly, evaluating how quickly models learn new tasks is also essential in the CL setting. Although CF is crucial to address in CL systems, there is no consensus on how to measure it (Pfülb and Gepperth, 2019). Arguably, the two most popular and general metrics to address this issue are *Average Accuracy* and *Forgetting Measure* (Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2018, 2019b). The former evaluates the average accuracy, while the latter measures forgetting after the model is trained continually on all the given task mini-batches. Concretely, we aim to measure test performance on the dataset $\mathcal{D}$ for each of the $\mathcal{T}$ tasks, letting $a_{j,i}$ be the performance of the model on the held-out test set of task ti after the model is trained on task tj. Later Chaudhry et al. (2019a) proposed the third metric, *Learning Curve Area* (*LCA*), that measures how quickly a model is able to learn. The three metrics are defined as follows:

- **Average Accuracy**: $A \in [0, 1]$ (Chaudhry et al., 2018). The average accuracy after incremental training from the first task to $\mathcal{T}$ is given as:

$$A_{\mathcal{T}} = \frac{1}{\mathcal{T}} \sum_{i=1}^{\mathcal{T}} a_{\mathcal{T},i}$$

- **Forgetting Measure**: $F \in [-1, 1]$ (Chaudhry et al., 2018). The average forgetting measure after incremental training from the first task to $\mathcal{T}$ is defined as:

$$F_{\mathcal{T}} = \frac{1}{\mathcal{T}-1} \sum_{i=1}^{\mathcal{T}-1} f_i^{\mathcal{T}}$$

  where f_i^j is the forgetting on task ti after the model is trained up to task tj and computed as:

$$f_i^j = \max_{k \in \{1,\cdots,j-1\}} a_{k,i} - a_{j,i}$$

- **Learning Curve Area**: $LCA \in [0, 1]$ (Chaudhry et al., 2019a). LCA is the area under the Z_b curve, which captures the learner's performance on all $\mathcal{T}$ tasks. Z_b is the average accuracy after observing the b-th mini-batch and is defined as:

$$Z_b = \frac{1}{\mathcal{T}} \sum_{i=1}^{\mathcal{T}} a_{i,b,i}$$

where b denotes the mini-batch number.

Similarly, Kemker et al. (2017) proposed three metrics for evaluating CF, i.e. the metrics evaluate the ability of a model to retain previously acquired knowledge and how well it acquires new information. In the NLP domain, Yogatama et al. (2019) introduced a new metric, based on an online (prequential) encoding (Blier and Ollivier, 2018), which measures the adoption rate of an existing model to a new task. Specifically, the metric called *online codelength* $\ell(\mathcal{D})$ is defined as follows:

$$\ell(\mathcal{D}) = \log_2 |y| - \sum_{i=2}^{N} \log_2 p\left(y_i | x_i; \theta_{\mathcal{D}_{i-1}}\right)$$

where $|y|$ denotes the number of possible labels (classes) in the dataset $\mathcal{D}$, and $\theta_{\mathcal{D}_i}$ stands for the model parameters trained on a particular subset of the dataset. Similar to *LCA* (Chaudhry et al., 2019a), *online codelength* is also related to an area under the learning curve.

While most CL methods consider settings without human-in-the-loop, some allow a human domain expert to provide the model with empirical knowledge about the task at hand. For instance, Prokopalo et al. (2020) introduced the evaluation of human assisted learning across time by leveraging user-defined model adaptation policies for NLP and speech tasks, such as machine translation and speaker diarization.

3.3 Evaluation Datasets

Most widely adopted CL benchmark datasets are image corpora such as PERMUTED MNIST (Kirkpatrick et al., 2016), CUB-200 (Welinder et al., 2010; Wah et al., 2011), or split CIFAR-10/100 (Lopez-Paz and Ranzato, 2017). Benchmark corpora have also been proposed for objects - CORE50 (Lomonaco and Maltoni, 2017) and sound - AUDIOSET (Gemmeke et al., 2017). However, none of the well-established standard datasets used in the CL field is related to NLP. Therefore, due to the scarcity of NLP-curated datasets, some of the above-mentioned datasets have also been utilized for NLP scenarios.

Name	*Details*	*Related works*
XCOPA - Cross-lingual Choice of Plausible Alternatives	• a typologically diverse multilingual dataset for causal commonsense reasoning, which is the translation and reannotation • covers 11 languages from distinct families	(Edoardo M. Ponti and Korhonen, 2020)
WEBTEXT	• a dataset of millions of webpages suitable for learning language models without supervision • 45 million links scraped from Reddit, 40 GB dataset	(Radford et al., 2019)
C4 - Colossal Clean Crawled Corpus	• a dataset constructed from Common Crawl's web crawl corpus and serves as a source of unlabeled text data • 17 GB dataset	(Raffel et al., 2020)
LIFELONG FEWREL - Lifelong Few-Shot Relation Classification Dataset	• sentence-relation pairs derived from Wikipedia distributed over 10 disjoint clusters (representing different tasks)	(Wang et al., 2019b) (Obamuyide and Vlachos, 2019)
LIFELONG SIMPLE QUESTIONS	• single-relation questions divided into 20 disjoint clusters (i.e. resulting in 20 tasks)	(Wang et al., 2019b)

Table 3: An overview of major NLP benchmark datasets to evaluate multi-task and CL methods.

Similarly, in the absence of NLP benchmark corpora, the majority of papers use adopted versions of popular NLP datasets. One such example is domain adaptation, where researchers frequently use

different, standard NLP corpora for in-domain and out-of-domain datasets. Farquhar and Gal (2018) stressed that prior research often presented incomplete evaluations, and utilized dedicated CL datasets or environments that cannot be considered general, one-size-fits-all benchmarks. As the scholars argued, such benchmarks are useful in narrow cases, limited to their respective subdomains. The number of NLP-specific CL datasets is still very limited, even though there have been lately a few notable attempts to create such corpora (summarized in Table 3).

Importantly, as Parisi et al. (2019) contended, with the increasing complexity of the evaluation dataset at hand, the overall performance of the model often decreases. The scholars attributed this to the fact that the majority of methods are tailored to work only for less complex scenarios, as they are not robust and flexible enough to alleviate CF in less controlled experimental conditions. In a similar vein, Yogatama et al. (2019) stressed that the recent tendency to construct datasets that are easy to solve without requiring generalization or abstraction is an impediment toward general linguistic intelligence. Hence, we advocate further research on establishing challenging evaluation datasets and evaluation metrics for CL in NLP that will allow to capture how well models generalize to new, unseen tasks.

4 Continual Learning in NLP Tasks

Natural language processing covers a diverse assortment of tasks. Despite the variety of NLP tasks and methods, there are some common themes. On a syntax level, sentences in any domain or task follow the same syntax rules. Furthermore, regardless of task or domain, there are words and phrases that have almost the same meaning. Therefore, sharing of syntax and semantic knowledge should be feasible across NLP tasks. In this section, we explore how CL methods are used in most popular NLP tasks.

4.1 Word and Sentence Representations

Distributed word vector representations underlie many NLP applications. Although high-quality word embeddings can considerably boost performance in downstream tasks, they cannot be considered a silver bullet as they suffer from inherent limitations. Typically, word embeddings are trained on large-size general corpora, as the size of in-domain corpora is in most cases not sufficient. This comes at a cost, since embeddings trained on general-purpose corpora are often not suitable for domain-specific downstream tasks, and in result, the overall performance suffers. In a CL setting, this also implies that vocabulary may change with respect to two dimensions: time and domain. There is an established consensus that the meaning of words changes over time due to complicated linguistic and social processes (e.g. Kutuzov et al., 2018; Shoemark et al., 2019). Hence, it is important to detect and accommodate shifts in meaning and data distribution, while preventing previously learned representations from CF.

In general, a CL scenario for word and sentence embeddings has not received much attention so far, except for a handful of works. To tackle this problem, for example Xu et al. (2018) proposed a meta-learning method, which leverages knowledge from past multi-domain corpora to generate improved new domain embeddings. Liu et al. (2019) introduced a sentence encoder updated over time using matrix conceptors to continually learn corpus-dependent features. Importantly, Wang et al. (2019b) argued that when a NN model is trained on a new task, the embedding vector space undergoes undesired changes, and in result the embeddings are infeasible for previous tasks. To mitigate the problem of embedding space distortion, they proposed to align sentence embeddings using anchoring. Recently a research line at the intersection of word embeddings and language modeling, termed contextual embeddings, has emerged and demonstrates state-of-the-art results across numerous NLP tasks. In the next section, we will look closely at how this approach to learning embeddings is geared towards CL.

4.2 Language Modeling

Contextual representations learned via unsupervised pre-trained language models (LMs), such as: ULMFiT (Howard and Ruder, 2018), ELMo (Peters et al., 2018) or BERT (Devlin et al., 2019), allow to attain strong performance on a wide range of supervised NLP tasks. Precisely, thanks to inductive transfer, complex task-specific architectures have become less needed. In consequence, the process of training many neural-based NLP systems boils down to two steps: (1) an NN-based language model is trained

on a large unlabeled text data; (2) this pre-trained language representation model is then reused in supervised downstream tasks. In principle, a large LM trained on a sufficiently large and diverse corpus is able to perform well across many datasets and domains (Radford et al., 2019). Furthermore, Gururangan et al. (2020) studied the effects of task-adaptation as well as domain-adaptation on the transferability of adapted pre-trained LMs across domains and tasks. The authors concluded that continuous domain- and task-adaptive pre-training of LMs leads to performance gains in downstream NLP tasks.

Research interest in LM-based methods for CL in NLP has recently spiked. d'Autume et al. (2019) proposed an episodic memory-based model, MBPA++, that augments the encoder-decoder architecture. In order to continually learn, MBPA++ also performs sparse experience replay and local adaptation. The scholars claimed that MBPA++ trains faster than A-GEM, and it does not take longer to train it than an encoder-decoder model. While this is possible due to sparse experience replay, yet MBPA++ requires extra memory. In a similar vein, LAMOL (Sun et al., 2020) is based on language modeling. Unlike MBPA++, this method does not use any extra memory. LAMOL mitigates CF by means of pseudo-sample generation, as the model is trained on the mix of new task data and pseudo old samples.

4.3 Question Answering

Question answering (QA) is considered a traditional NLP task, encompassing reading comprehension as well as information and relation extraction among others. Conceptually, it is also very much related to conversational agents, such as chatbots and dialogue agents. Hence, not only in research settings but even more so in real-life scenarios (e.g. in the conversation), it is immensely important for such systems to continuously extract and accumulate new knowledge (Chen and Liu, 2018). It is believed that a good dialogue agent should be able to not only interact with users by responding and asking questions, but also to learn from both kinds of interaction (Li et al., 2017b).

Although question answering is a stand-alone NLP task, some researchers (e.g. Kumar et al., 2016; McCann et al., 2018) proposed to view NLP tasks through the lens of QA. In the context of CL, both d'Autume et al. (2019) and Sun et al. (2020) reported experimental results on a QA task. Research in dialogue agents, which are able to continually learn, is a very active area (e.g. Gasic et al., 2014; Su et al., 2016). Findings of Li et al. (2017a) indicate that a conversational model initially trained with fixed data can improve itself, when it learns from interactions with humans in an on-line fashion. Interestingly, information and relation extraction were an early subject of research interest in CL. Information extraction is considered one of the first research areas, which embraced the goal of never-ending learning. A semi-supervised NELL (Carlson et al., 2010) and an unsupervised ALICE (Banko and Etzioni, 2007) systems, which iteratively extract information and build general domain knowledge, were at the forefront of CL in NLP. In the case of relation extraction, Wang et al. (2019b) introduced an embedding alignment method to enable CL for relation extraction models. Also, Obamuyide and Vlachos (2019) proposed to extend the work of Wang et al. (2019b) by framing the lifelong relation extraction as a meta-learning problem; however, without the costly need for learning additional parameters.

4.4 Sentiment Analysis and Text Classification

Sentiment analysis (SA) is a popular choice for evaluating models on text classification. Arguably the most pressing problem of current approaches to SA is their poor performance on new domains. Therefore, various domain adaptation methods have been proposed to improve the performance of SA models in the multi-domain scenario (consult Barnes et al., 2018). This issue is of utmost importance if one thinks about CL in sentiment classification. One of the earliest approaches to CL for SA was proposed in Chen et al. (2015). According to Chen and Liu (2018), CL can enable SA models to adapt to a large number of domains, since many new domains may already be covered by other past domains. Additionally, SA systems should become more accurate not only in classification but also in the discovery of word polarities in specific domains. Research in opinion about aspects has been conducted as well. Shu et al. (2016) presented an unsupervised CL approach to classify opinion targets into entities and aspects. Furthermore, Shu et al. (2017) proposed a method based on conditional random fields to improve supervised aspect extraction across time. Experiments on text classification in the CL setting were performed in d'Autume et al. (2019) and Sun et al. (2020).

4.5 Machine Translation

The approach introduced by Luong and Manning (2015) laid the groundwork for subsequent studies in adapting neural machine translation (NMT). More specifically, the authors explored the adaptation through continued training, where an NMT model trained using large corpora in one domain can later initialize a new NMT model for another domain. Their findings suggested that fine-tuning of the NMT model trained on out-of-domain data using a small in-domain parallel corpus boosts performance. Likewise, other works (e.g. Freitag and Al-Onaizan, 2016; Chu et al., 2017) supported this claim. Khayrallah et al. (2018) pointed out that, due to over-fitting, some amount of knowledge learned from the out-of-domain corpus is being forgotten during fine-tuning. Hence, such domain adaptation techniques are prone to CF. NMT models experience difficulties when dealing with data from diverse domains, hence we argue this is not a sufficient solution. As dominant fine-tuning approaches require training and maintaining a separate model for each language or domain, Bapna and Firat (2019) proposed to add light-weight task-specific adapter modules to support parameter-efficient adaptation. We further argue that NMT – as opposed to phrase-based MT – rarely incorporates translation memory, and so it is inherently harder for NMT models to adapt using active or interactive learning. However, some attempts have been made (e.g. Peris and Casacuberta, 2018; Liu et al., 2018; Kaiser et al., 2017; Tu et al., 2018). In a similar vein, there were approaches (Sokolov et al., 2017) to incorporate bandit learners, which implicitly involve domain adaptation and on-line learning, for MT systems. We share the viewpoint of Farajian et al. (2017), that NMT models ultimately should be able to adapt on-the-fly to on-line streams of diverse data (i.e. language, domain), and thus CL for NMT is essential.

While domain adaptation methods are widely used in the context of adapting NMT models, there have also been other attempts. Multilingual NMT (Dong et al., 2015; Firat et al., 2016; Ha et al., 2016; Johnson et al., 2017; Tan et al., 2019) can be framed as a multi-task learning problem. Multilingual NMT aims to use a single model to translate between multiple languages. Such systems are beneficial not only because they can handle multiple translation directions using a single model, and thus reduce training and maintenance costs, but also due to joint training with high-resource languages they can improve performance on low- and zero-resource languages (Arivazhagan et al., 2019). To eliminate the need for retraining the entire NMT system, Escolano et al. (2020) proposed a language-specific encoder-decoder architecture, where languages are mapped into a shared space, and either encoder or decoder is frozen when training on a new language.

Another related research line is curriculum learning. Most approaches concentrate on the selection of training samples according to their relevance to the translation task at hand. Different methods have been applied, for example, van der Wees et al. (2017) and Zhang et al. (2019) adapted a model to a domain by introducing samples which are increasingly domain-relevant or domain-distant respectively. Curriculum methods based on difficulty and competence were explored in Zhang et al. (2018) and Platanios et al. (2019). Ruiter et al. (2020) proposed a self-supervised NMT model, that uses data selection to train on increasingly complex and task-related samples in combination with a denoising curriculum.

A stream of research focused on techniques more traditionally associated with CL has also been present. In the works of Miceli Barone et al. (2017); Khayrallah et al. (2018); Variš and Bojar (2019); Thompson et al. (2019) regularization approaches (e.g. EWC) were leveraged. Furthermore, Kim and Rush (2016) explored knowledge distillation, where the student model learns to match the teacher's actions at the word- and sequence-level. Wei et al. (2019) proposed an on-line knowledge distillation approach, in which the best checkpoints are utilized as the teacher model. Lately, Li et al. (2020) demonstrated that label prediction continual learning leveraging compositionality brings improvements in NMT.

5 Research Gaps and Future Directions

Although there is a growing number of task-specific approaches to CL in NLP, nevertheless, the body of research work remains rather scant (Sun et al., 2020; Greco et al., 2019). While the majority of current NLP methods is task-specific, we believe task-agnostic approaches will become much more prevalent. Contemporary methods are limited along three dimensions: data, model architectures, and hardware.

In the real world, we often deal with partial information data. Moreover, data is drawn from non-i.i.d. distributions, and is subject to agents' interventions or environmental changes. Although attempts exist, where a model learns from a stream of examples without knowing from which dataset and distribution they originate from (e.g. d'Autume et al., 2019), such approaches are rare. Furthermore, learning on a very few examples (e.g. via few-shot transfer learning) (Liu, 2020) is a major challenge for current models, even more so performing out-of-distribution generalization (Bengio, 2019). In particular, widely used in NLP sequence-to-sequence models still struggle with *systematic generalization* (Lake and Baroni, 2018; Bahdanau et al., 2019), being unable to learn general rules and reason about high-level language concepts. For instance, recent work on counterfactual language representations by Feder et al. (2020) is a promising step in that direction. The non-stationary learning problem can be alleviated by understanding and inferring causal relations from data (e.g. Osawa et al., 2019) – which is an outstanding challenge (Pearl, 2009) – and coming up with combinations that are unlikely to be present in training distributions (Bengio, 2019). Namely, language is compositional; hence, the model can dynamically manipulate the semantic concepts which can be recombined in novel situations (Lake et al., 2015) and later supported by language-based *abductive reasoning* (e.g. Bhagavatula et al., 2020).

On a model level, a combination of CL with Bayesian principles should allow to identify better the importance of each parameter of an NN and aid parameter pruning and quantization (e.g. Ebrahimi et al., 2020; Golkar et al., 2019). We believe that not only the parameter informativeness should be uncertainty-guided, but also the periodic replay of previous memories should be informed by causality. Furthermore, it is important to focus on reducing model capacity and computing requirements. Even though the over-parametrization of NNs is pervasive (Neyshabur et al., 2018), many current CL approaches promote the expansion of parameter space. We envision further research efforts focused on compression methods, such as knowledge distillation, low-rank factorization and model pruning. Importantly, while CL allows for continuous adaptation, we believe that integrating CL with meta-learning has the potential to further unlock generalization capabilities in NLP. As meta-learning is able to efficiently learn with limited samples, hence such a CL model would adapt quicker in dynamic environments (e.g. Ritter et al., 2018; Al-Shedivat et al., 2018). This would be especially beneficial for NLP systems operating in low-resource language and domain settings.

Finally, further research aiming at developing comprehensive benchmarks for CL in NLP would be an important addition to the existing studies. On the one hand, we observe a proliferation of multi-task benchmarks (e.g. McCann et al., 2018; Wang et al., 2018, 2019a). On the other hand, the CL paradigm and evaluation of CL systems call for more robust approaches than traditional performance metrics (e.g. accuracy, F1 measure) and multi-task evaluation schemes with clearly defined data and task boundaries.

6 Conclusion

In this work, we provided a comprehensive overview of existing research on CL in NLP. We presented a classification of ML paradigms and methods for alleviating CF, as well as discussed how they are applied to various NLP tasks. Also, we summarized available benchmark datasets and evaluation approaches. Finally, we identified research gaps and outlined directions for future research endeavors. We hope this survey sparks interest in CL in NLP and inspires to view linguistic intelligence in a more holistic way.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful feedback. We also would like to thank Marc'Aurelio Ranzato for providing a detailed clarification of the *LCA* metric. We thank Carlos Escolano for a fruitful discussion on Table 2. This work is supported in part by the Catalan Agencia de Gestión de Ayudas Universitarias y de Investigación (AGAUR) through the FI PhD grant; the Spanish Ministerio de Ciencia e Innovación and by the Agencia Estatal de Investigación through the Ramón y Cajal grant and the project PCIN-2017-079; and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 947657).

References

Maruan Al-Shedivat, Trapit Bansal, Yura Burda, Ilya Sutskever, Igor Mordatch, and Pieter Abbeel. 2018. Continuous adaptation via meta-learning in nonstationary and competitive environments. In *International Conference on Learning Representations (ICLR)*.

Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. 2017. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3366–3375.

Rahaf Aljundi, Marcus Rohrbach, and Tinne Tuytelaars. 2019. Selfless sequential learning. In *International Conference on Learning Representations (ICLR)*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint*.

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. 2019. Systematic generalization: What is required and can it be learned? In *International Conference on Learning Representations (ICLR)*.

Michele Banko and Oren Etzioni. 2007. Strategies for lifelong knowledge extraction from the web. In *Proceedings of the 4th International Conference on Knowledge Capture*, K-CAP '07, page 95–102, New York, NY, USA. Association for Computing Machinery.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Projecting embeddings for domain adaption: Joint modeling of sentiment analysis in diverse domains. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 818–830, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shawn Beaulieu, Lapo Frati, Thomas Miconi, Joel Lehman, Kenneth O. Stanley, Jeff Clune, and Nick Cheney. 2020. Learning to continually learn. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 992–1001.

Yoshua Bengio. 2019. From system 1 deep learning to system 2 deep learning.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 41–48. ACM.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations (ICLR)*.

Léonard Blier and Yann Ollivier. 2018. The description length of deep learning models. In *Advances in Neural Information Processing Systems*, pages 2216–2226.

Léon Bottou. 1999. *On-Line Learning and Stochastic Approximations*, page 9–42. Cambridge University Press, USA.

Léon Bottou and Yann LeCun. 2004. Large scale online learning. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 217–224. MIT Press.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Rich Caruana. 1997. Multitask learning. *Mach. Learn.*, 28(1):41–75.

Nicolo Cesa-Bianchi and Gabor Lugosi. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European Conference on Computer Vision (ECCV)*, pages 532–547.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2019a. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations (ICLR)*.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet Kumar Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. 2019b. Continual learning with tiny episodic memories. *arXiv preprint*.

Zhiyuan Chen and Bing Liu. 2018. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3):1–207.

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, Beijing, China. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13143–13152. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Sayna Ebrahimi, Mohamed Elhoseiny, Trevor Darrell, and Marcus Rohrbach. 2020. Uncertainty-guided continual learning with bayesian neural networks. In *International Conference on Learning Representations (ICLR)*.

Olga Majewska Qianchu Liu Ivan Vulić Edoardo M. Ponti, Goran Glavaš and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2020. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. *arXiv preprint*.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on*

Machine Translation, pages 127–137, Copenhagen, Denmark. Association for Computational Linguistics.

Sebastian Farquhar and Yarin Gal. 2018. Towards robust evaluations of continual learning. *arXiv preprint*.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2020. Causalm: Causal model explanation through counterfactual language models. *arXiv preprint*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint*.

Milica Gasic, Dongho Kim, Pirros Tsiakoulis, Catherine Breslin, Matthew Henderson, Martin Szummer, Blaise Thomson, and Steve J. Young. 2014. Incremental on-line adaptation of pomdp-based dialogue managers to extended domains. In *INTERSPEECH*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Siavash Golkar, Micheal Kagan, and Kyunghyun Cho. 2019. Continual learning via neural pruning. In *NeurIPS 2019 Workshop Neuro AI*.

Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. 2019. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3601–3605, Florence, Italy. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Steven C. H. Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. 2018. Online learning: A comprehensive survey. *arXiv preprint*.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2020. Meta-learning in neural networks: A survey. *arXiv preprint*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of The 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of The 37th International Conference on Machine Learning*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. 2016. Less-forgetting learning in deep neural networks. *arXiv preprint*.

Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. 2017. Learning to remember rare events. In *International Conference on Learning Representations (ICLR)*.

Ronald Kemker and Christopher Kanan. 2018. Fearnet: Brain-inspired model for incremental learning. In *International Conference on Learning Representations (ICLR)*.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L. Hayes, and Christopher Kanan. 2017. Measuring catastrophic forgetting in neural networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Huda Khayrallah, Brian Thompson, Kevin Duh, and Philipp Koehn. 2018. Regularized training objective for continued training for domain adaptation in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 36–44.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2016. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1378–1387. PMLR.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017a. Dialogue learning with human-in-the-loop. In *International Conference on Learning Representations (ICLR)*.

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017b. Learning through dialogue interactions by asking questions. In *International Conference on Learning Representations (ICLR)*.

Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. 2019. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *Proceedings of the*

36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 3925–3934. PMLR.

Yuanpeng Li, Liang Zhao, Kenneth Church, and Mohamed Elhoseiny. 2020. Compositional language continual learning. In *International Conference on Learning Representations (ICLR)*.

Zhizhong Li and Derek Hoiem. 2016. Learning without forgetting. In *European Conference on Computer Vision (ECCV)*, pages 614–629. Springer.

Bing Liu. 2020. Learning on the job: Online lifelong and continual learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 13544–13549.

Ming Liu, Wray Buntine, and Gholamreza Haffari. 2018. Learning to actively learn neural machine translation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 334–344, Brussels, Belgium. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Vincenzo Lomonaco and Davide Maltoni. 2017. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pages 6467–6476.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Arun Mallya and Svetlana Lazebnik. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7765–7773.

Massimiliano Mancini, Elisa Ricci, Barbara Caputo, and Samuel Rota Bulò. 2018. Adding new tasks to a single network with weight transformations using binary masks. In *European Conference on Computer Vision (ECCV)*, pages 0–0.

Sahisnu Mazumder, Bing Liu, Shuai Wang, and Nianzu Ma. 2019. Lifelong and interactive learning of factual knowledge in dialogues. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 21–31, Stockholm, Sweden. Association for Computational Linguistics.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1489–1494, Copenhagen, Denmark. Association for Computational Linguistics.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. 2018. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint*.

Abiola Obamuyide and Andreas Vlachos. 2019. Meta-learning improves lifelong relation extraction. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 224–229, Florence, Italy. Association for Computational Linguistics.

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. 2019. Practical deep learning with bayesian principles. In *Advances in neural information processing systems*, pages 4287–4299.

S. J. Pan and Q. Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Lorenzo Pellegrini, Gabrile Graffieti, Vincenzo Lomonaco, and Davide Maltoni. 2019. Latent replay for real-time continual learning. *arXiv preprint*.

Álvaro Peris and Francisco Casacuberta. 2018. Active learning for interactive neural machine translation of data streams. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 151–160, Brussels, Belgium. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. *arXiv preprint*.

B. Pfülb and A. Gepperth. 2019. A comprehensive, application-oriented study of catastrophic forgetting in DNNs. In *International Conference on Learning Representations (ICLR)*.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. *arXiv preprint*.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for 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 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Yevhenii Prokopalo, Sylvain Meignier, Olivier Galibert, Loic Barrault, and Anthony Larcher. 2020. Evaluation of lifelong learning systems. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1833–1841, Marseille, France. European Language Resources Association.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017a. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, pages 506–516.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017b. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. 2018. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations (ICLR)*.

Mark B. Ring. 1994. Continual learning in reinforcement environments. In *GMD-Bericht*.

Samuel Ritter, Jane Wang, Zeb Kurth-Nelson, Siddhant Jayakumar, Charles Blundell, Razvan Pascanu, and Matthew Botvinick. 2018. Been there, done that: Meta-learning with episodic recall. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4354–4363. PMLR.

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

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint*.

Dana Ruiter, Josef van Genabith, and Cristina Espa na Bonet. 2020. Self-Induced Curriculum Learning in Self-Supervised Neural Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint*.

Jeffrey C Schlimmer and Richard H Granger. 1986. Incremental learning from noisy data. *Machine learning*, 1(3):317–354.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4548–4557.

Shai Shalev-Shwartz. 2012. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, pages 2990–2999.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

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, Austin, Texas. Association for Computational Linguistics.

Lei Shu, Hu Xu, and Bing Liu. 2017. Lifelong learning CRF for supervised aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 148–154, Vancouver, Canada. Association for Computational Linguistics.

Daniel L Silver and Robert E Mercer. 2002. The task rehearsal method of life-long learning: Overcoming impoverished data. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 90–101. Springer.

Daniel L Silver, Qiang Yang, and Lianghao Li. 2013. Lifelong machine learning systems: Beyond learning algorithms. In *Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium Series*.

Shagun Sodhani, Sarath Chandar, and Yoshua Bengio. 2020. Toward training recurrent neural networks for lifelong learning. *Neural Computation*, 32(1):1–35.

Artem Sokolov, Julia Kreutzer, Kellen Sunderland, Pavel Danchenko, Witold Szymaniak, Hagen Fürstenau, and Stefan Riezler. 2017. A shared task on bandit learning for machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 514–524, Copenhagen, Denmark. Association for Computational Linguistics.

Ray J Solomonoff. 1989. A system for incremental learning based on algorithmic probability. In *Proceedings of the Sixth Israeli Conference on Artificial Intelligence, Computer Vision and Pattern Recognition*, pages 515–527.

José G. C. de Souza, Matteo Negri, Elisa Ricci, and Marco Turchi. 2015. Online multitask learning for machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 219–228, Beijing, China. Association for Computational Linguistics.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR.

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Continuously learning neural dialogue management. *arXiv preprint*.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2020. LAMOL: LAnguage MOdeling for Lifelong Language Learning. In *International Conference on Learning Representations (ICLR)*.

Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations (ICLR)*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Sebastian Thrun. 1996. Is learning the n-th thing any easier than learning the first? In *Advances in neural information processing systems*, pages 640–646.

Sebastian Thrun and Lorien Pratt. 1998. *Learning to Learn: Introduction and Overview*, page 3–17. Kluwer Academic Publishers, USA.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Dušan Variš and Ondřej Bojar. 2019. Unsupervised pretraining for neural machine translation using elastic weight consolidation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 130–135, Florence, Italy. Association for Computational Linguistics.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3266–3280.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019b. Sentence embedding alignment for lifelong relation extraction. 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 796–806, Minneapolis, Minnesota. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark. Association for Computational Linguistics.

Hao-Ran Wei, Shujian Huang, Ran Wang, Xin-yu Dai, and Jiajun Chen. 2019. Online distilling from checkpoints for 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 1932–1941, Minneapolis, Minnesota. Association for Computational Linguistics.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.

Yeming Wen, Dustin Tran, and Jimmy Ba. 2020. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations (ICLR)*.

Gerhard Widmer and Miroslav Kubat. 1993. Effective learning in dynamic environments by explicit context tracking. In *European Conference on Machine Learning (ECML)*, pages 227–243. Springer.

Hu Xu, Bing Liu, Lei Shu, and P. Yu. 2019. Open-world learning and application to product classification. In *The World Wide Web Conference*, WWW '19, page 3413–3419, New York, NY, USA. Association for Computing Machinery.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Lifelong domain word embedding via meta-learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4510–4516. International Joint Conferences on Artificial Intelligence Organization.

Ju Xu and Zhanxing Zhu. 2018. Reinforced continual learning. In *Advances in Neural Information Processing Systems*, pages 899–908.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint*.

JaeHong Yoon, Jeongtae Lee, Eunho Yang, and Sung Ju Hwang. 2018. Lifelong learning with dynamically expandable network. In *International Conference on Learning Representations (ICLR)*.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995. PMLR.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. *arXiv preprint*.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in 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 1903–1915, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *arXiv preprint*.