Investigating Masking-based Data Generation in Language Models

Ed S. Ma*
eddsma@outlook.com

Abstract

The current era of natural language processing (NLP) has been defined by the prominence of pre-trained language models since the advent of BERT. A feature of BERT and models with similar architecture is the objective of masked language modeling, in which part of the input is intentionally masked and the model is trained to predict this piece of masked information. Data augmentation is a data-driven technique widely used in machine learning, including research areas like computer vision and natural language processing, to improve model performance by artificially augmenting the training data set by designated techniques. Masked language models (MLM), an essential training feature of BERT, have introduced a novel approach to perform effective pre-training on Transformer based models in natural language processing tasks. Recent studies have utilized masked language model to generate artificially augmented data for NLP downstream tasks. The experimental results show that Mask based data augmentation method provides a simple but efficient approach to improve the model performance. In this paper, we explore and discuss the broader utilization of these data augmentation methods based on MLM.

1 Introduction

Pre-trained language models (PLMs) have revolutionized the field of natural language processing, with BERT architectures standing out for their innovative design and impressive performance. These Transformer (Vaswani et al., 2017) based models, based largely on transformer architecture, use bi-directional representations to understand context, thus pushing the limits of previous uni-directional models. A distinctive feature of BERT and similar models is the goal of masked language modeling, in which part of the input is intentionally masked and the model is trained to predict this masked token. This strategy simulates a fuller understanding of the context and relationships between words, leading to a better understanding of language nuances and meaning. As a result, BERT-like models have found extensive application in a variety of NLP tasks, such as text classification, sentiment analysis, and question answering, significantly pushing the boundaries of what machines can understand and accomplish in the realm of human language.

NLP tasks requires a significant amount of high-quality annotated data for several critical reasons, primarily related to the inherent complexity of human language and the need for machine learning models to model it effectively understand, interpret and generate. Languages are extremely complicated and complex, with countless nuances, exceptions and rules. Consisting of morphological, syntactic, semantic, and pragmatic aspects, they require understanding not only of the words and phrases, but also of context, intent, and even cultural or social cues. In order for an NLP model to understand all of these elements and generate human-like text, it must learn from a variety of examples that show these characteristics in multiple different contexts. This is where high-quality annotated data comes into play. They provide the models with explicit labels or additional information that make it easier to understand the various features of the language.

The performance of machine learning models is highly dependent on the variety and amount of training data. Because these models learn by identifying patterns in the input data, a larger and more diverse data set allows the models to be exposed to a wider range of patterns and situations. This leads to better generalization ability and the models can process new inputs more effectively. For example, if you train an NLP model on annotated data from different domains like literature, science, law, social media, etc., it can understand and generate

* Independent project. Work in Progress. Contact: eddsma@outlook.com

texts related to each of these domains. The quality of the annotated data is important. Poorly annotated data can mislead the model during training, resulting in suboptimal performance or even completely wrong outputs. Accurate annotations are fundamental to supervised learning as they serve as the basis for the model. They help models distinguish between different elements of language, understand the relationships between words, and understand the meaning and intent behind phrases or sentences. Therefore, high-quality annotated data plays a crucial role in training robust and reliable NLP models. It provides the rich, diverse, and concise input the models need to learn the complexities of the language, ensures their applicability in different domains and scenarios, and acts as an effective guide during the training process to optimize their performance.

It is often difficult and expensive to obtain quality annotated text data in large volume. Traditional and expensive way is to hire crowd workers with target language capability to annotate data. An example is Amazon Mechanical Turk (AMT). AMT is a crowdsourcing service that enables individuals and businesses to outsource tasks to a distributed workforce who can perform these tasks virtually. Based on requirements, workers (known as 'Turkers') will go through target data, manually annotate it as per instructions. Once a worker completes a Human Intelligence Task (HIT), you can review their work, approve or reject it based on the quality of the annotation, and then pay the worker. With its vast, diverse workforce, AMT is often used to create large, annotated datasets for NLP downstream tasks. However, this annotating method is common but very expensive. Therefore, researchers have started exploring annotating methods at cheaper costs. Some examples are methods that are based on distant supervision. Unlabeled dialogue corpora in the target domain can be easily curated from previous conversation transcripts or collected via crowdsourcing (Budzianowski et al., 2018; Byrne et al., 2019) with no additional cost on human labour.

With the rise of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) models in NLP, there has been a significant leap forward in the field's capabilities and applications. BERT's bi-directional approach, where it considers the context from both left and right of a word, sets it apart from previous models. The experiments on several NLP benchmark shows that it allows a more accurate understanding of the meaning of a word within its context, leading to improved performance in a variety of tasks such as sentiment analysis, named entity recognition, question answering, and others. There are several data augmentation methods that utilize these pre-trained language models in unsupervised manners, with no human annotation needed.

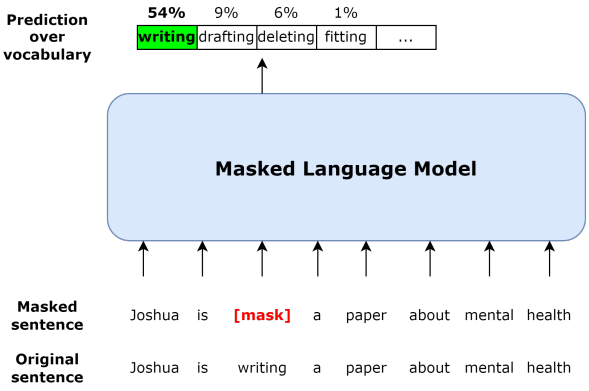

Figure 1: Visual illustration of MLM in BERT. During pre-training, some tokens (about 10 – 15 percent) are masked randomly (only one token is masked in this example) and the model is trained to predict the correct words for all masked tokens. The model will be updated based on the probability distribution over vocabulary through MLM loss functions and backprobagation. Numbers are for illustration purposes instead of real model output.

2 Related Work

In this section, we will introduce pre-trained language models and recent data augmentation methods including data augmentation methods with MLM.

2.1 Pre-trained Language Models

Pre-trained language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), BART (Lewis et al., 2020), and T5 (Raffel et al., 2020) have revolutionized the field of Natural Language Processing. These models rely on transformer-based architectures and unsupervised learning, and they are trained on large text corpora to learn useful representations of language. They can then be fine-tuned on specific tasks with smaller amounts of task-specific data. The objective is masked language modeling is essential and prominent in these models,

BERT (Devlin et al., 2019), Bidirectional Encoder Representations from Transformers, was a

pioneering pre-trained language model in the field. It is trained using a masked language modeling objective where some percentage of the input data is masked or hidden during training, and the model is trained to predict these masked words based on the context provided by the unmasked words. BERT's MLM approach enables the model to learn a deep, bidirectional representation of the sentence, rather than just predicting future words in the text like traditional language models. This improves the model's understanding of the context and the semantic relationships between words, leading to significant improvements on a wide range of NLP tasks. We show a visual illustration of MLM in BERT in Figure 1. Since BERT is a bi-directional model, the probability of predicting the correct masked word is dependent on known surrounding words in the context. RoBERTa (Liu et al., 2019), a variant of BERT developed by Meta, modifies the training process and hyperparameters of BERT, but still utilizes the MLM objective. RoBERTa extends the training time, uses larger mini-batches and learning rates, and removes the next sentence prediction objective that BERT had. These tweaks led to improved performance, demonstrating the impact of a well-thought-out training strategy.

XLNet (Yang et al., 2019) is a another transformer-based model, combining the advantages of BERT with autoregressive language models. It introduces a permutation-based training objective that allows it to learn the dependency of words on both their preceding and succeeding context, mitigating some of the limitations of BERT's MLM approach where the predicted words don't have an impact on each other.

T5 (Raffel et al., 2020) Text-to-Text Transfer Transformer takes a unique approach to its pre-training objective. Instead of employing the usual masked language modeling, T5 treats every NLP problem as a text generation task. This is achieved by reframing tasks into a unified text-to-text format where every task, be it translation, summarization, or question answering, becomes a matter of generating the target text from the source text. During pre-training, T5 uses a denoising autoencoder-style objective where random spans of text are masked out and the model learns to reconstruct the original text. T5's approach simplifies the application of the model to a wide variety of tasks, as it doesn't require task-specific model architectures. Instead, differences between tasks are captured purely in the text input and output formats. As for performance, T5 has shown state-of-the-art results on a number of benchmark datasets across different NLP tasks. Its scalability, coupled with the general text-to-text approach, makes T5 a highly flexible and powerful tool for NLP tasks, showcasing the potential of this unified framework.

BART (Lewis et al., 2020) Bidirectional and Auto-Regressive Transformers brings a distinctive pre-training objective to the table. BART combines the best of both worlds: the bidirectional context learning from BERT and the sequence-to-sequence nature of models like GPT. During pre-training, BART corrupts the text by applying a noising function (i.e., by masking out random spans of text, similar to T5 (Raffel et al., 2020)), and then learns how to reconstruct the original uncorrupted text. This objective is different from traditional masked language modeling as it forces the model to build an understanding of the entire sentence structure and not just predict masked tokens independently. BART has proven to be highly effective. It has achieved state-of-the-art results on a variety of benchmark datasets, including the CNN/Daily (Nallapati et al., 2016) Mail summarization task, the SQuAD question answering benchmark (Rajpurkar et al., 2016, 2018), and others. By combining the advantages of bidirectional and autoregressive transformers, BART has shown a robust ability to handle a wide range of NLP tasks, highlighting the power of sequence-to-sequence and denoising autoencoder-style pre-training.

Pre-trained language models have had a profound impact on NLP, and the Masked Language Modeling objective has been a critical part of their success. By learning to predict missing words in a sentence, models are able to gain a deeper understanding of language context and semantics. This has resulted in significant performance improvements across a wide range of NLP tasks, and it continues to drive progress in the field.

Masked language modeling plays a crucial role in these notable pre-trained language models. By learning to fill in the masked words, language models gain a deep understanding of syntax, semantics, and contextual relationships. Masked language modeling serves as a self-supervised learning task that allows models to learn from vast amounts of unlabeled text data. This technique allows pre-trained language models to capture the nuances of language and generate coherent and contextual

responses when presented with incomplete or ambiguous input. By incorporating masked language modeling, pre-trained language models have revolutionized various natural language processing tasks such as text generation, sentiment analysis, machine translation, and question answering.

2.2 Data Augmentation

Data augmentation (DA) is a technique used to generate additional training data in situations where there is a lack of sufficient data. It involves various methods, ranging from simple rule-based techniques to more advanced learnable generation-based methods. The primary objective of data augmentation is to ensure that the generated data is valid for the given task and belongs to the same distribution as the original data (Raille et al., 2020). This validity is determined by factors such as similar semantics in machine translation or the same labels in text classification as the original data. Additionally, augmented data should also exhibit diversity to enhance the generalization of models for subsequent tasks. The diversity of augmented data can be achieved through three categories of data augmentation methods: paraphrasing, noising, and sampling.

Categorization

In order to ensure validity, augmented data should also possess diversity to enhance the generalization of models in subsequent tasks. This diversity refers to the range of variations within the augmented data. (Li et al., 2022) introduces categorization of data augmentation methods based on the diversity of resulting augmented data. Paraphrasing-based methods generate augmented data that maintains a limited semantic difference from the original data by making proper and controlled changes to sentences. The goal is to produce augmented data that conveys very similar information as the original. Noising-based methods introduce discrete or continuous noise to the data while ensuring its validity. The objective of these methods is to enhance the model's robustness. Sampling-based methods, on the other hand, understand the data distributions and generate novel data samples from within these distributions. These methods output more diverse data and cater to a wider range of requirements in downstream tasks, leveraging artificial heuristics and trained models.

Paraphrases are commonly observed in natural language, initiated from early works in NLP (Barzilay and McKeown, 2001; Madnani and Dorr, 2010), serve as alternative expressions to convey identical information as the original form. Consequently, utilizing paraphrasing as a technique for data augmentation is a fitting approach. Paraphrasing encompasses multiple levels, encompassing lexical paraphrasing, phrase paraphrasing, and sentence paraphrasing.

Basic word-/sentence-level operations

Various basic word-level and sentence-level operations can be employed to augment text data, including insertion, deletion, substitution, and swapping. Insertion, additional words or phrases are inserted into the original text. By inserting different linguistic elements, such as synonyms or context-relevant terms, the extended data can simulate different writing styles and increase the diversity of the data set. (Wei and Zou, 2019; Peng et al., 2021) This variation can allow the model to learn better and generalize to different text types. Deletion, on the other hand, is the removal of specific words or phrases from the original text. This technique helps the model focus on the most important information by eliminating non-essential elements. This operation has been employed in various works such as (Wei and Zou, 2019; Yu et al., 2019; Rastogi et al., 2020) By training on augmented data with strategically deleted content, the model can become more resilient to noise and less dependent on specific phrases or words. Substitution is another powerful data augmentation technique. Words are replaced by their synonyms or similar terms. By introducing variations in vocabulary, the model can learn to recognize and understand different word choices and expand its language skills. (Xie et al., 2017; Wang et al., 2018; Daval-Frerot and Weis, 2020; Lowell et al., 2021; Regina et al., 2021) This technique is particularly useful for improving the model's ability to deal with out-of-vocabulary words and to adapt to different linguistic expressions. Swapping is an augmentation technique that rearranges the order of words or phrases within a sentence. By rearranging the sequence of linguistic elements, the model can learn to understand different sentence structures and syntactic variations. (Wei and Zou, 2019; Zhang et al., 2020a; Longpre et al., 2020; Rastogi et al., 2020; Howard et al., 2022) This increases the model's flexibility and understanding of different sentence patterns, making it more effective in tackling sentence reordering or paraphrasing tasks.

PLM backed methods

As mentioned in Section 2.1, the success of deep pre-trained language models in recent years can be attributed to their ability to acquire extensive linguistic knowledge through pre-training. As a result, these sophisticated pre-trained models have naturally become valuable resources for data augmentation purposes.

Auto-regressive models have been explored in generating augmented data. For instance, (Peng et al., 2021) utilizes the pre-trained GPT based models to generate utterances and dialogue acts, respectively, ensuring data quality through filtering. (Queiroz Abonizio and Barbon Junior, 2020) applied DistilBERT on original sentences to generate synthetic sentences. Transformer decoder based model GPT-2 (Radford et al., 2019), in particular, has gained popularity as a model for generating augmented data. (Zhang et al., 2020a) employed GPT-2 to generate significantly diversified augmented data for extreme multi-label classification. Another GPT-2 backed data augmentation method LAMBDA is introduced in (Anaby-Tavor et al., 2020). They utilize GPT-2, which is pre-trained and fine-tuned on the training set, to generate labeled augmented sentences. These augmented sentences are subsequently filtered using a classifier to maintain data quality. (Kumar et al., 2020) applies a similar method without the use of a filtering classifier. (Quteineh et al., 2020) employed a label-conditioned GPT-2 model to generate augmented data. (Tarjan et al., 2020) utilized GPT-2 to generate augmented data and subsequently performed tokenization and splited them into subwords derived statistically, thereby avoiding vocabulary explosion in morphologically rich languages.

To obtain augmented data, some studies employ masked language models. For instance, (Ng et al., 2020) utilizes a masked language model to create both a corruption model and a reconstruction model. The model comprises a BERT-base pre-trained language model and a classification layer. The BERT parameters are additionally pre-trained on a vast Reddit dataset, ensuring a large-scale training process. The process involves generating data points that are initially distant from the original data manifold using the corruption model, and subsequently using the reconstruction model to bring those data points back to the original data manifold, resulting in the final augmented data. Such decoding sampling methods have also been utilized in (Gao et al., 2022) which first masks tokens or spans then fills in replaced tokens by pre-trained language models.

Obtaining unlabeled raw data can be relatively easy in certain scenarios, presenting an opportunity to convert such data into valid, usable data and significantly augment the overall dataset. (Thakur et al., 2021) introduce a data augmentation approach involving fine-tuning BERT on the original data, followed by utilizing the fine-tuned BERT to label unlabeled sentence pairs. These augmented data, along with the gold data, are then combined to train SBERT (Reimers and Gurevych, 2019). Data distillation has been utilized as part of the self-training process (Miao et al., 2020), where the label of unlabeled data is determined using an iteratively updated teacher model. On question answering tasks, (Yang et al., 2021) applies a similar self-training method using a cross-attention-based teacher model to determine the label for each QA pair. SentAugment (Du et al., 2021), a data augmentation method that leverages task-specific query embeddings from labeled data to retrieve sentences from a vast bank of billions of unlabeled sentences obtained from web crawling. In some cases, existing models from other tasks are directly transferred to generate pseudo-parallel corpora. (Montella et al., 2020) leverages Wikipedia to access a large volume of sentences and utilizes external information extraction package to extract triplets from these Wikipedia sentences. Recognizing BERT's proficiency in object-property (OP) relationship prediction and object-affordance (OA) relationship prediction, (Zhao et al., 2020) directly employs a fine-tuned BERT model to predict the labels of the two samples. (Wei et al., 2022) explore data augmentation on semantically-preserved continuous space using Transformers.

Unsupervised methods

Unsupervised methods for text data augmentation leverage unlabeled data to enhance the quantity and quality of training data without relying on explicit or extra annotations or labels. Some unsupervised methods have been mentioned under *word-/sentence-level operations*, here we discuss methods not falling into those categories. These methods aim to harness the inherent structure, pre-training objectives and patterns present in the unlabeled data to generate additional training examples.

"Mixup" methods take a different approach to data augmentation by utilizing intermediate em-

beddings instead of generating augmented samples in natural language text. These methods involve sampling data in vector space (manifold) based on existing data, potentially resulting in samples with different labels than the original data. A prominent feature of this group of methods is no additional human annotation is required to construct new training samples, and resulting labels are automatically generated after the "mixing" process. The concept of Mixup was initially introduced in computer vision works (Zhang et al., 2018). Mixup has found wide application in numerous recent studies. Building upon this work, (Guo et al., 2019) proposed two variations of Mixup specifically for sentence classification, performing sample interpolation in the word embedding space and interpolating the hidden states of sentence encoders. Both variations generate interpolated samples as augmented data. (Wu et al., 2022) propose a data augmentation method integrating Mixup strategy with MLM that involves converting a sentence from its one-hot representation to a controllable smoothed representation. (Kong et al., 2022) introduce a framework for saliency map-informed textual data augmentation and regularization, which combines Dropout and Mixup techniques to address the issue of overfitting in text learning.

Mixup methods for text classification are explored in (Sun et al., 2020; Si et al., 2021). (Sun et al., 2020) introduces Transformer based Mixup method, a combination of Mixup and transformer-based pre-trained architecture, which is evaluated on text classification datasets. (Chen et al., 2020a) incorporates Mixup into Named Entity Recognition (NER) to improve performance in NER tasks. (Chen et al., 2020b) introduce MixText, which creates augmented training samples by interpolating text in hidden space. The method performing mixing of hidden representations of examples in different Transformer layers. The resulting labels are determined by mixing samples.

Machine translation has emerged as a popular method for data augmentation, leveraging its ability to naturally paraphrase text. With the advancements in machine translation models, this technique has found widespread application across various tasks. Researchers have explored off-the-shelf machine translation models on data augmentation with no additional human effort on annotations. Back-translation, involves translating the original text into other languages and then translating it back to the original language to generate augmented text. Unlike word-level methods, back-translation doe not merely replace individual words, but rather rewrites the entire sentence in a generated manner. (Yu et al., 2018; Xie et al., 2020; Fabbri et al., 2021) utilize English-French translation models, in both directions (known as round-trip translations), to perform back-translation on each sentence and generate paraphrases. Lowell also incorporates this method as one of the unsupervised data augmentation techniques. Additionally, Zhang employs back-translation to obtain formal expressions of the original data in the context of style transfer tasks.

Building upon the concept of vanilla back-translation, several studies have introduced additional features and techniques to enhance its effectiveness. In (Zhang et al., 2020b), a discriminator is employed to filter the sentences obtained through back-translation. By applying this filtering mechanism, the quality of the augmented data is significantly improved, as only sentences surpassing a certain threshold are retained. (Nugent et al., 2021) explore various softmax temperature settings to ensure diversity in the augmented data while preserving semantic meaning. By carefully adjusting the temperature, they achieve a balance between generating diverse examples and maintaining the original intent. Overall, these additional features and techniques bring improvements to the vanilla back-translation method, ensuring diversity and enhancing the quality of the augmented data.

The quick brown fox jumps over the lazy dog in the zoo

The [mask] brown fox jumps over the lazy dog in the zoo

The quick [mask] fox jumps over the lazy dog in the zoo

The quick brown fox [mask] over the lazy dog in the [mask]

The quick brown fox jumps over the lazy [mask] in the zoo

The quick brown [mask] jumps over the lazy [mask] in the zoo

Figure 2: Example of augmented samples with mask tokens. The top first sentence is the original sentence without any augmentation (The quick brown fox jumps over the lazy dog in the zoo). The rest five sentences are augmented sentences obtained by data augmentation with mask tokens in which an original word in the original sentence is replaced with a MASK token.

DA with Mask tokens

Mask tokens are a critical component of pre-trained language models like BERT (Devlin et al., 2019). With BERT architecture, a small percentage of the input tokens are randomly masked during pre-training. These masked tokens are usually replaced with a special [MASK] token. We present a visual example of MLM in Figure 2 for ease of understanding. Using mask tokens in pre-trained language models such as BERT enables effective language representation learning and allows the models to capture the contextual relationships between words and phrases, resulting in improved performance on various NLP tasks. As a pre-training strategy, mask tokens are seen by the model during pre-training process. Mask tokens are pre-trained to preserve contextualized information (like other tokens) but not actually encountered in test-time examples. Intuitively, mask token can be used in sentences like other words without injecting undesired signals in sentences under the training scheme of BERT.

Word replacement and back-translation have been demonstrated to be effective unsupervised data augmentation methods for short written text classification in previous studies (Wei and Zou, 2019; Xie et al., 2020). However, the effectiveness of these augmentation methods is diminished when applied to pre-trained models (Shleifer, 2019). Additionally, the applicability of backt-ranslation is limited in our scenario, given that translating multi-turn dialogue is considerably more challenging compared to short text. Recent work (Yavuz et al., 2020) has examined the effectiveness of replacing a portion of original word tokens with [MASK] tokens as a data augmentation strategy in NLP. This technique allows for the generation of extended examples in which certain words are escaped, providing the model with additional training instances. (Yavuz et al., 2020) introduce MASKAUGMENT, a controllable data augmentation that augments text input by leveraging the pre-trained mask token from BERT model on the task of dialog act tagging. (Lin and Ng, 2022) extend this idea and propose a semi-supervised bootstrapping training method for dialog breakdown detection tasks. Both works utilize teacher-student learning scheme on samples with different probabilities of mask token replacement. By leveraging the pre-trained language model's ability to understand and contextualize mask tokens, this extension method introduces diversity and variation into the training data without introducing unwanted bias or noise. This DA approach has been shown to increase model robustness, improve generalization to invisible data, and increase performance on various downstream NLP tasks.

Compared to the aforementioned DA methods, mask data augmentation appears to be better suited for pre-trained language models with MLM training objectives and demonstrate effectiveness in dialog tasks. This method is intuitively simple and controllable as it only considers the position and probability of replacement. By maintaining a low probability of masking (Yavuz et al., 2020; Lin and Ng, 2022), the sentences augmented with mask tokens can retain the original meaning while undergoing natural changes.

DA with adversarial training

NLP data augmentation with adversarial training has emerged as a promising technique to improve the robustness and performance of NLP models. Adversarial training involves generating adversarial examples or perturbations that aim to deceive the model while preserving the original meaning or intent of the text. In the context of NLP, these adversarial examples can be created by applying various techniques but not limited to synonym substitution, word deletion, or sentence modification to the original text.

(Cheng et al., 2020) constructs adversarial samples based on the original samples following (Cheng et al., 2019) and then applies two Mixup strategies. (Qu et al., 2021) combines back-translation with adversarial training. (Morris et al., 2020) introduce an off-the-shelf framework for developing data augmentation with adversarial attacks covering various attack recipes from literature(Li et al., 2019; Jin et al., 2020). This integration allows them to synthesize augmented examples that are both diverse and informative, incorporating multiple transformations to enrich the data. The goal of adversarial data augmentation is to expose the model to challenging and diverse examples that mimic real-world linguistic variations and potential attacks. By incorporating these adversarial examples into the training data, NLP models can learn to handle and generalize better to such variations, making them more robust and reliable in real-world scenarios. Adversarial training not only helps models overcome the limitations of overfitting but also enhances their ability to handle noisy or adversarial inputs.

3 Discussion

3.1 Non-MLM Pre-trained Language Models

Nowadays, we are witnessing the rapid emergence of a diverse array of pre-trained language models that go beyond the traditional training objective of masked language modeling. Prominent examples include models such as T5 (Raffel et al., 2020), Flan (Chung et al., 2022) and GPT (Radford et al., 2018, 2019). T5/Flan-T5 deviate from the conventional approach of masked language modeling where individual tokens are masked. Instead, these models employ a different masking and prediction strategy that operates at the level of text spans. In this strategy, specific spans of text are masked, and the models are trained to predict the correct content within those masked spans. These models are designed with multifaceted training objectives aimed at enhancing different aspects of language understanding and generation. In addition to incorporating masked language modeling, these models can also integrate various other objectives such as sequence classification, document retrieval, question-answering, and machine translation. By encompassing this diverse range of training objectives, pre-trained language models gain the ability to capture a comprehensive understanding of linguistic properties, semantic relationships, and contextual nuances. This advancement in training objectives not only enhances the versatility and context-awareness of these language models but also facilitates their application across a wide spectrum of natural language processing tasks and domains. As ongoing research and innovation in this field continue to unfold, we can expect the emergence of even more advanced pre-trained language models, leading to significant advancements in natural language understanding and generation across diverse domains. Incorporating mask token-based data augmentation into the training pipeline of models like T5 and GPT is potentially feasible and explorable.

3.2 Emerging Large Language Models

The presence of large language models (LLM) like GPT-3 family models (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2022) and Llama (Touvron et al., 2023) have brought both challenges and significant impacts on data augmentation methods.

One challenge is the scalability and computational demands that come with data proliferation. Large language models require a large amount of training data to achieve their impressive performance. However, generating extended data at the same scale can be computationally intensive and slow, requiring additional model processing and inference. On the other hand, the influence of large language models on data augmentation is significant. These models have extensive linguistic knowledge and can generate high-quality synthetic examples using techniques such as conditional generation or controlled generation. Consequently, data augmentation techniques can benefit from integrating these models to generate diverse and contextually relevant augmented data.

Additionally, large language models can serve as powerful tools for data enrichment. By optimizing or adapting pre-trained models like GPT-3 to specific downstream tasks, they can be used to generate advanced data tailored to the target task. This approach allows models to learn from extended data and potentially improve their performance on the given task.

Moreover, large language models can also be used for data enrichment by leveraging their embeds or contextual features. These embeddings can provide valuable information about relationships and similarities between text samples and facilitate the development of new data enhancement strategies that preserve semantic and syntactic properties.

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7383–7390, New York, NY, USA. AAAI Press.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL*, pages 50–57, Toulouse, France. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In

Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of EMNLP*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. 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 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020a. Local additivity based data augmentation for semi-supervised ner. In *arXiv:2010.01677*.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020b. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of ACL*, pages 4324–4333, Florence, Italy. Association for Computational Linguistics.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. AdvAug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5961–5970, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. In *arXiv:2210.11416*.

Guillaume Daval-Frerot and Yannick Weis. 2020. WMD at SemEval-2020 tasks 7 and 11: Assessing humor and propaganda using unsupervised data augmentation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1865–1874, Barcelona (online). International Committee for Computational Linguistics.

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.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online. Association for Computational Linguistics.

Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proc. of NAACL-HLT*, pages 704–717, Online. Association for Computational Linguistics.

Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, and Ruifeng Xu. 2022. Mask-then-fill: A flexible and effective data augmentation framework for event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4537–4544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. In *arXiv:1905.08941*.

Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. 2022. NeuroCounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5056–5072, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8018–8025, New York, NY, USA. AAAI Press.

Fanshuang Kong, Richong Zhang, Xiaohui Guo, Samuel Mensah, and Yongyi Mao. 2022. DropMix: A textual data augmentation combining dropout with mixup. In *Proceedings of the 2022 Conference on Empirical*

Methods in Natural Language Processing, pages 890–899, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019*. The Internet Society.

Qian Lin and Hwee Tou Ng. 2022. A semi-supervised learning approach with two teachers to improve breakdown identification in dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11011–11019, Online. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv:1907.11692*.

Shayne Longpre, Yu Wang, and Chris DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4401–4411, Online. Association for Computational Linguistics.

David Lowell, Brian Howard, Zachary C. Lipton, and Byron Wallace. 2021. Unsupervised data augmentation with naive augmentation and without unlabeled data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4992–5001, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Lin Miao, Mark Last, and Marina Litvak. 2020. Twitter data augmentation for monitoring public opinion on COVID-19 intervention measures. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Sebastien Montella, Betty Fabre, Tanguy Urvoy, Johannes Heinecke, and Lina Rojas-Barahona. 2020. Denoising pre-training and data augmentation strategies for enhanced RDF verbalization with transformers. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 89–99, Dublin, Ireland (Virtual). Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283, Online. Association for Computational Linguistics.

Tim Nugent, Nicole Stelea, and Jochen L. Leidner. 2021. Detecting environmental, social and governance (esg) topics using domain-specific language models and data augmentation. In *Flexible Query Answering Systems*, pages 157–169, Cham. Springer International Publishing.

OpenAI. 2022. Introducing chatgpt. Https://openai.com/blog/chatgpt/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *arXiv:2203.02155*.

Baolin Peng, Chenguang Zhu, Michael Zeng, and Jianfeng Gao. 2021. Data Augmentation for Spoken Language Understanding via Pretrained Language Models. In *Proc. Interspeech 2021*, pages 1219–1223.

Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Weizhu Chen, and Jiawei Han. 2021. Co{da}: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. In *International Conference on Learning Representations*.

Hugo Queiroz Abonizio and Sylvio Barbon Junior. 2020. Pre-trained data augmentation for text classification. In *Intelligent Systems*, pages 551–565, Cham. Springer International Publishing.

Husam Quteineh, Spyridon Samothrakis, and Richard Sutcliffe. 2020. Textual data augmentation for efficient active learning on tiny datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7400–7410, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Guillaume Raille, Sandra Djambazovska, and Claudiu Musat. 2020. Fast cross-domain data augmentation through neural sentence editing. In *arXiv:2003.10254*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Chetanya Rastogi, Nikka Mofid, and Fang-I Hsiao. 2020. Can we achieve more with less? exploring data augmentation for toxic comment classification. In *arXiv:2007.00875*.

Mehdi Regina, Maxime Meyer, and Sebastien Goutal. 2021. Text data augmentation: Towards better detection of spear-phishing emails. In *arXiv:2007.02033*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sam Shleifer. 2019. Low resource text classification with ulmfit and backtranslation. In *arXiv:1903.09244*.

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2021. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1569–1576, Online. Association for Computational Linguistics.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip Yu, and Lifang He. 2020. Mixup-transformer: Dynamic data augmentation for NLP tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3436–3440, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Balazs Tarjan, Gyorgy Szaszak, Tibor Fegyo, and Peter Mihajlik. 2020. Deep transformer based data augmentation with subword units for morphologically rich online asr. In *arXiv:2007.06949*.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of NAACL-HLT*, pages 296–310, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Weihua Luo, and Rong Jin. 2022. Learning to generalize to more: Continuous semantic augmentation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7930–7944, Dublin, Ireland. Association for Computational Linguistics.

Xing Wu, Chaochen Gao, Meng Lin, Liangjun Zang, and Songlin Hu. 2022. Text smoothing: Enhance various data augmentation methods on text classification tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 871–875, Dublin, Ireland. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 6256—-6268, Red Hook, NY, USA. Curran Associates Inc.

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. 2017. Data noising as smoothing in neural network language models. In *Proceedings of 5th International Conference on Learning Representations, ICLR 2017*, Toulon, France. OpenReview.net.

Yinfei Yang, Ning Jin, Kuo Lin, Mandy Guo, and Daniel Cer. 2021. Neural retrieval for question answering with cross-attention supervised data augmentation. In *Proceedings of ACL-IJCNLP (Volume 2: Short Papers)*, pages 263–268, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Semih Yavuz, Kazuma Hashimoto, Wenhao Liu, Nitish Shirish Keskar, Richard Socher, and Caiming Xiong. 2020. Simple data augmentation with the mask token improves domain adaptation for dialog act tagging. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5083–5089, Online. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*.

Shujuan Yu, Jie Yang, Danlei Liu, Runqi Li, Yun Zhang, and Shengmei Zhao. 2019. Hierarchical data augmentation and the application in text classification. *IEEE Access*, 7:185476–185485.

Danqing Zhang, Tao Li, Haiyang Zhang, and Bing Yin. 2020a. On data augmentation for extreme multi-label classification. In *arXiv:2009.10778*.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Yi Zhang, Tao Ge, and Xu Sun. 2020b. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228, Online. Association for Computational Linguistics.

Zhenjie Zhao, Evangelos Papalexakis, and Xiaojuan Ma. 2020. Learning Physical Common Sense as Knowledge Graph Completion via BERT Data Augmentation and Constrained Tucker Factorization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3293–3298, Online. Association for Computational Linguistics.