Recent Data Augmentation Techniques in Natural Language Processing: A Brief Survey

Lingling Xu, Haoran Xie, *Senior Member, IEEE*, Fu Lee Wang, *Senior Member, IEEE*, and Weiming Wang

***Abstract*—Data augmentation has recently gained increasing interest in natural language processing (NLP) because of its excellent performance in low-resource settings, contrastive learning, and few-shot learning. Data augmentation is initially a strategy to increase the amount of data by employing semantically invariant transformations, such as back translation and synonym replacement, on the raw data. With the development of data augmentation, a variety of augmentation strategies are designed to produce samples with opposite labels to the original data or even samples with unseen categories. In this paper, we provide a comprehensive and thorough study of text data augmentation techniques. We first discuss various data augmentation methods and then classify them into three types: semantic-invariant augmentation, random augmentation, and generative augmentation. Subsequently, we highlight the main application scenarios and downstream tasks involving data augmentation. We also describe the challenges in developing text data augmentations and the work that can be further investigated in the future. To conclude, this paper aims to summarize data augmentation techniques in NLP and show how they work to further improve the performance of NLP tasks.**

***Index Terms*—Data Augmentation, Contrastive Learning, Low-resource Setting, Few-shot Learning, NLP, Survey.**

I. Introduction

DATA augmentation works mainly by making small changes to the data directly or by generating new data using some deep learning models. Data augmentation is extremely important in low-resource scenarios in which the number of training data is sparse, as it helps increase the number of training data while reducing the operational costs of annotating. In addition, data augmentation can create diverse data and enrich the semantic feature space of data, further enhancing the robustness of model. Data augmentation first appeared in the field of computer vision (CV), where studies [1], [2], [3] discovered that cropping, rotation, and scaling of image data greatly improved model performance. However, it is challenging to employ these continuous noises for text data augmentation due to the discrete nature of the text.

Despite this limitation, data augmentation for NLP has seen an increase in interest and demand. Inspired by the data augmentation methods of cropping and rotation in CV, [4] proposes two text data augmentation strategies, sentence cropping and sentence rotating, based on the dependency tree structure of NLP. Sentence cropping preserves some important words in a sentence and then removes the rest of the irrelevant words to generate a new sentence. Sentence rotating rotates the portable tree segment around the root of the dependency tree to form a synthetic sentence. Besides, data augmentation techniques such as random deletion and token cutoff can also be seen as a variant of cropping in CV. Inspired by the Mixup [5] image data augmentation strategy in CV, SeqMix [6] and MixText [7] are then proposed for text data augmentation. SeqMix [6] attempts to incorporate word embeddings and sentence embeddings from the convolutional neural network (CNN) [8] to form novel samples, whereas MixText [7] combines sentence embeddings from the BERT [9] to obtain new synthetic samples.

To address various NLP tasks, a large number of text data augmentation methods have been devised, resulting in many surveys on data augmentation in NLP. [10] explores text data augmentation for deep learning, which includes not only data augmentation in NLP but also in recommender systems. [11] focuses on data augmentation techniques used in text classification. [12] provides a systematic and empirical investigation of data augmentation in NLP with a small amount of labeled data. Both [13] and [14] discuss NLP data augmentation methods. [13] does not contain data augmentation used in contrastive learning, while [14] does not discuss in detail to which NLP tasks data augmentation can be applied. Therefore, we take the data augmentation approaches used in contrastive learning into account and present the NLP tasks involving data augmentation in detail.

In this paper, we aim to provide a systematic investigation of text data augmentation in NLP according to the form of data augmentation. We discover that some data augmentations are well-designed using prior knowledge to enable the semantic meaning of augmented data to remain unchanged, while certain data augmentations focus on generating label-conditioned sentences. In addition, we also give the specific application scenarios of data augmentation and downstream tasks that involve data augmentation. The remaining paper is organized as follows. Section II discusses commonly used text data augmentation techniques and classifies them as semantic invariant augmentation, random augmentation, and generative augmentation. Section III describes the application scenarios of data augmentation, including low-resource language, contrastive learning, and few-shot learning. Section IV analyzes the downstream tasks that use data augmentation. Section V presents challenges and future work in data augmentation for

The research has been supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E01/19), and Lam Woo Research Fund (LWP20019) and Direct Grant (DR23B2), Lingnan University, Hong Kong.

Lingling Xu, Weiming Wang, and Fu Lee Wang are with the School of Science and Technology, Hong Kong Metropolitan University, Hong Kong SAR.

Haoran Xie is with the Department of Computing and Decision Sciences, Lingnan University, Hong Kong.

Fu Lee Wang is the corresponding author (email: pwang@hkmu.edu.hk).

NLP. Section VI concludes the paper.

II. Text Data Augmentation Methods and Techniques

Numerous data augmentation strategies have been proposed to promote the performance of NLP tasks, as they can both increase the quantity of data and enrich its diversity. In this survey, we focus on studying how augmented sentences are generated from original sentences. After summarizing these approaches, we observe that data augmentation is mainly performed by some well-designed transformations, stochastic change, and generative models. We divide the existing text data augmentation methods into three categories: semantic-invariant augmentation, random augmentation, and generative augmentation.

Semantic invariant augmentation is usually carefully designed and implemented by exploiting prior knowledge or deep learning models. Random augmentation, on the other hand, emphasizes the randomness of the generation of the augmented samples, so that the semantics of the augmented samples do not always remain the same as the original sentences. Generative augmentation is usually done by using generative models like VAE and BART to generate sentences that are consistent with the label on top of the original sentence and the given label. In the following work, we will discuss these text data augmentation methods in detail. Particularly, we provide an overview of current data augmentation techniques in Fig. 1.

A. Semantic-invariant Data Augmentation

Semantic-invariant augmentation is an augmentation strategy that preserves the syntax and semantics of the sentence via making well-designed local modifications to the original sentence. Paraphrases and well-designed substitutions are two common types of semantic-invariant augmentation.

Paraphrasing is widely applied as a text data augmentation strategy in NLP tasks [15], [16], [17], as it can provide augmented text with more varied lexical choices and syntactic structures while maintaining the semantic meaning of the raw sentence. Back-translation [18], [19] is definitely the most popular paraphrasing method, which involves translating the sentence into a certain intermediate language and then translating it back into the original language. Other research aims to train an end-to-end model to produce meaningful translations [20] and augment sentences at the decoding stage by adding syntactic features [21], latent variables [22], or submodular targets [17].

Well-designed substitution is also a common data augmentation method, where certain words in a sentence are replaced with other words without changing the semantics of the sentence. An intuitive idea is to use the synonyms as replacement words for substitutions [23]. The synonyms can be words from a pre-defined corpus such as WordNet [24], words with high similarity to the replacement word [25], entities of the same type [26], [27], or words with the same morphology [4]. Additionally, work from [4] argues that we can also keep the semantics of a sentence intact by removing words that are not important. Moreover, Xie et al. [15] devise a replacement approach based on TF-IDF where the uninformative words in the sentence are replaced with other uninformative words. Hsu et al. [28] substitute the unimportant words with the predicted words generated by the auto-encoding model or the seq2seq model without altering the aspect-level polarity. Notably, these semantic-invariant augmentation techniques we discussed are unsupervised data augmentation and do not use the label information of sentences. However, Wang et al. [29] attempt to substitute representative words with their corresponding antonyms to obtain new sentences, which may be semantically irrelevant or even opposite to the original sentence.

TABLE I: The examples of word-level random augmentations.

Method	Text
Original	There is a little boy running in the playground.
Deletion	There is a boy in the playground.
Swapping	There is little a boy running in playground the.
Insertion	There is **great** a little **dog** boy running in the playground.
Substitution	There is a **beautiful cat** running in the playground.
Repetition	There there is a little boy boy running in the playground.

B. Random Data Augmentation

Semantic-invariant augmentation is crucial for tasks that require augmented samples to have the same semantic label as the original sentences. Random data augmentation, on the other hand, has also received extensive research attention due to its ease of implementation. Furthermore, random data augmentation can be roughly divided into word-level, token-level, and embedding-level augmentation, depending on the body of the noise added to the sentence.

Word-level augmentation means that noise is added to the words of sentences, either by random deletion, swapping, insertion, and substitution [30], or random repetition for some selected words [31]. These stochastic operations are easy to implement and do not always ensure that the semantic labels of the text remain unchanged. We give some examples to show this word-level random augmentation in Table I. Token-level augmentation includes token shuffling (shuffles the order of tokens randomly), token cutoff (erases some tokens randomly), feature cutoff (erases feature dimensions randomly), and span cutoff (erases token spans randomly) [32]. AEDA [33] is another easier random augmentation that generates augmented samples by inserting punctuation marks randomly in the original sentence.

The embedding-level random augmentation can be mainly performed by Mixup [6], [7], [34] and adversarial training. Inspired by Mixup [5], a data augmentation method that linearly interpolates two input images to obtain a target sample. Guo et al. [6] apply this method to the domain of text and proposed SeqMix, which creates augmented sentences by interpolating word embeddings and sentence embeddings linearly with CNN [8] and LSTM [35] as sentence encoders. Similarly, Chen et al. [7] use BERT as an encoder to generate sentence embeddings for sentence Mixup, and Sun et al. [34] employ a pretrained transformer as an encoder to obtain sentence embeddings for linear interpolation, further demonstrating the effectiveness

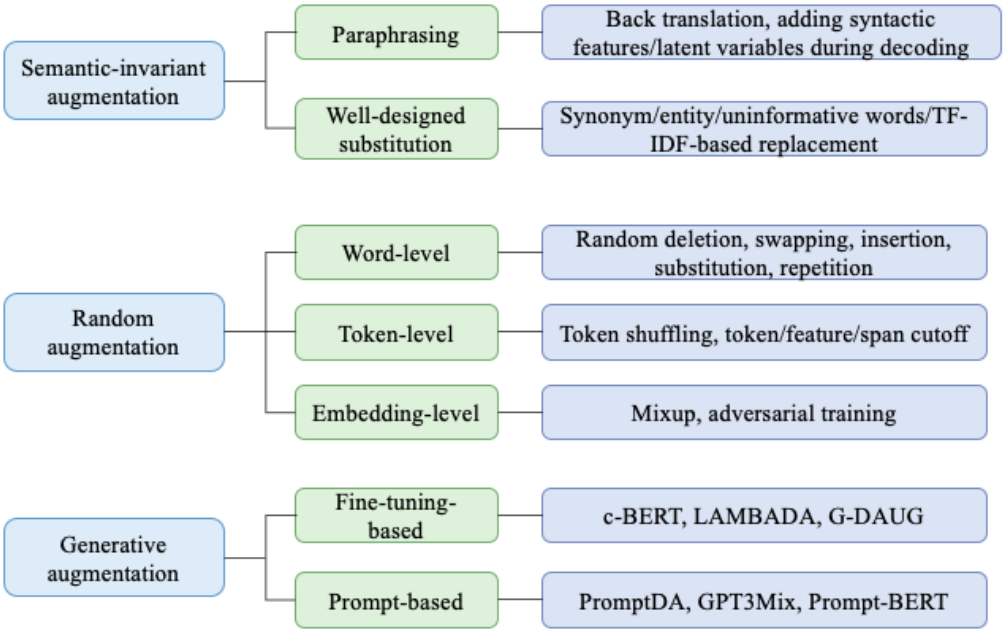

Fig. 1: An overview of recent data augmentation methods in NLP.

and generality of the Mixup augmentation in the text domain. Significantly, Mixup data augmentation requires label information to be known and is thus a supervised data augmentation technique. Assuming that the sentence embeddings of the two input sentences are e_i and e_j, and the labels are y_i and y_j, the augmented sentences and labels can be expressed as Equation 1:

$$\begin{aligned} e &= \lambda e_i + (1-\lambda)e_j, \\ y &= \lambda y_i + (1-\lambda)y_j, \end{aligned} \tag{1}$$

in which λ is sampled from the Beta distribution. Since the generated embeddings are a linear interpolation combination of two sentence embeddings, Mixup data augmentation can create semantically rich sentences. Additionally, the generated sentence labels vary because they are also an interpolation of two labels.

Adversarial training methods are commonly used to improve the robustness of models in text data [36], [37], [38]. It can also be used as a data augmentation technique to create adversarial examples using gradient-based noise. Specifically, for the input sentence embedding e_i with the label y_i, then the augmented sentence embeddings can be written as Equation 2:

$$e_i^* = e_i + \epsilon \frac{g}{\|g\|^2}, g = \nabla_{e_i} \mathcal{L}(f(e_i, y_i)), \tag{2}$$

where ϵ is random noise. Significantly, Mixup and adversarial training all require the participation of the label, thus they are supervised data augmentation approaches.

Moreover, dropout is another random augmentation method that is widely applied to contrastive learning [39], [40], which utilizes the dropout in the embedding layer and attention layer of BERT [9] to produce augmented samples. Concretely, a sentence is passed to the BERT encoder twice to obtain two different sentence representations.

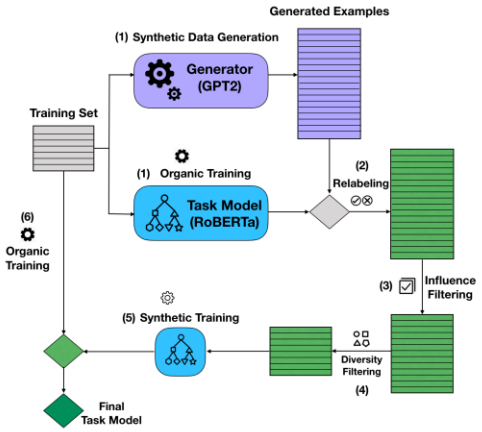

Fig. 2: The overview of data augmentation method G-DAUG^c [55].

C. Generative Data Augmentation

Deep generative models such as VAE [41], GAN [42], GPT-2 [43], GPT-3 [44], BART [45], and T5 [46] are employed in generative data augmentation methods to generate new sentences conditioned on the label. The semantic label of the augmented data obtained through generative data augmentation is thus determined by the given label and does not always maintain the same semantic label as the original data. Early generative data augmentation is typically performed on condition VAE [47], [48], [49], [50]; GANs [51], [52]; and a bidirectional RNN language model [53]. Furthermore, the benefits of developing pretrained language models (PLMs) [9], [54], two promising paradigms for data augmentation in NLP are proposed.

The first approach involves finetuning the PLMs using task-

specific data and then using the finetuned language model to generate new sentences. For example, [56], [57] use masked language modeling (MLM) mechanisms from BERT and BART, respectively, to produce new synthetic data by masking random words in the original sentences. Yang et al. [55] and Anaby-Tavor et al. [58] employ PLMs GPT-2 as the generator to capture the semantic information expressed implicitly in their training dataset to generate new synthetic sentences. With the help of pretrained language models, the novel data augmentation framework G-DAUG^c [55] (shown in Fig. 2) produces synthetic samples and chooses the most informative and varied samples for data augmentation. FliDA [59] generate augmented data using word substitution based on the pretrained T5, with a classifier to choose label-flipped data. C^{3}DA [60] adopts the T5 model as a text generator and produces new sentences based on given aspect words or sentiment labels (e.g., positive and negative) to enrich the dataset for aspect-based sentiment analysis. Despite these advancements, these PLMs could be overfitted with a small amount of task-specific data and fail to achieve excellent results.

The second type of approach utilizes the prompts, combined with the off-the-shelf PLM, to generate sentences directly without any task-specific fine-tuning. Wang et al. [61], for example, proposed PromDA, a data augmentation built on top of the T5-large model [46]. Specifically, PromDA keeps the parameters of the PLM frozen and trains only the soft prompt prepend at the beginning of the sentence, significantly reducing training resources. GPT3Mix [62] synthesizes hyper-realistic sentences from a variety of real samples by utilizing the large-scale language models of GPT-3 and the discrete prompt. Chen et al. [63] propose a label-guided data augmentation method that exploits the enriched label semantic information for data augmentation in a fashion similar to prompt-tuning. Liu et al. [64] devise a label-conditioned word substitution technique and a question-answering-based prompting approach for data augmentation. The label-conditioned technique aims to create a label-consistent example by capturing potential word-label dependencies, while the question-answering-based prompting approach focuses on generating new training data from unannotated text. Specific details of these two methods can be shown in Fig. 3. Moreover, Prompt-BERT [65] adds different discrete prompt templates to the same sentence and uses PLMs like BERT, RoBERTa to obtain different sentence representations to generate augmented examples.

III. Application Scenarios

In this section, we will discuss some application scenarios for data augmentation. Significantly, data augmentation mainly serves to raise the number of training data in low-resource scenarios, generate positive samples in contrastive learning, and synthesize unseen class samples in few-shot learning.

A. Low-resource Setting

Recent advances in large-scale neural language models [35], [9] have led to excellent performance in various NLP tasks, including machine translation [66], [67], [68] and NER [69],

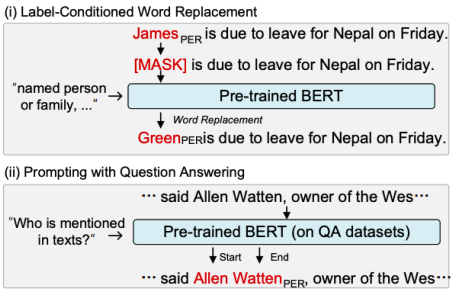

Fig. 3: Label-conditioned and prompting with question answering augmentation methods [64].

[70], [71], their accuracy largely depends on the accessibility of extensive sets of human-annotated training data. However, annotating data is time-consuming and expensive, thus data augmentation is extremely important in low-resource settings.

Inspired by the work in CV, Fadaee et al. [66] present a new data augmentation method that targets low-frequency words and produces novel sentence pairs with uncommon words in a novel synthetic context to enrich the training corpus for machine translation. Xia et al. [67] propose a generic data augmentation framework that generates parallel corpora via back-translating English to low-resource language or high-resource language as pivoting to a related high-resource language, improving the performance of low-resource translation. To improve the performance of low-resource machine translation, Li et al. [68] extend the training data via creating diverse pseudo-parallel data from the source and target sides.

Zhou et al. [71] employ masked entity language modeling (MELM) for data augmentation to obtain augmented data and alleviate the data scarcity in low-resource NER tasks. Liu et al. [70] use back translation to generate multilingual labeled data. These augmented data allow the NER model to learn the different linguistic features for cross-lingual NER tasks. For low-resource tagging tasks like NER and part-of-speech (POS) tagging, Ding et al. [69] develop a novel augmentation method using a language model trained on linearized labeled sentences to produce high-quality synthetic data. For low-resource natural language understanding (NLU) tasks, Wang et al. [61] develop synthetic data using generative augmentation techniques, reducing the effort for humans to annotate data while maintaining the quality of the synthetic data produced.

B. Contrastive Learning

Contrastive learning is a metric learning algorithm that learns useful representations by encouraging positive pairs to be closer and negative pairs to be further away. Positive pairs are typically different views of the anchor and can be generated by various data augmentations. Negative pairs are usually the remaining in-batch samples. As a result, data augmentation is essential to contrastive learning, and effective data augmentation will significantly enhance contrastive learning's performance. Notably, data augmentation can not only help generate positive pairs that share the same semantic label but also negative samples that are semantically dissimilar.

The most popular text data augmentation in contrastive learning for NLP is the dropout augmentation [72], [73], which is also the most recent state-of-the-art data augmentation strategy. Dropout augmentation utilizes the dropout in the embedding and attention layers of BERT to encode the same sentence twice to obtain two different sentence representations as positive pairs. Additionally, back translation [74], synonym replacement [29],token shuffle, and feature cutoff [32] are also used to produce positive pairs for sentence representation learning. By combining various data augmentations, Qu et al. [75] create diverse augmented examples, which are then combined with the contrastive learning objective to enhance NLU tasks. Wang et al. [60] employ generative data augmentation and contrastive learning to improve sentiment analysis.

In addition to generating positive samples, data augmentation can be used to create negative samples for contrastive learning. For example, CLINE [29] replaces words in sentences with antonyms to create negative samples for feature extraction with a triplet contrastive loss objective. MixCSE [76] produces hard negative samples by mixing the features of positive samples and negative samples randomly to further improve the performance of contrastive learning. To differentiate and uncouple semantic similarity from textual similarity, SNCSE [60] uses the Spacy[1] to perform sentence parsing to obtain the syntactic tree, lexical labels, and label stems of the sentence and then utilizes this information to transform the sentence into a syntactically correct and semantically-opposite sentence as soft negative samples. FlipDA [59] adopts generative data augmentation to generate a label-flipped augmented sample automatically, which can be considered negative samples of contrastive learning.

C. Few-shot Learning

Few-shot learning is a technique for extracting information from a small number of examples. Data augmentation techniques can assist few-shot learning by introducing different kinds of examples. Chao et al. [77] devise a novel data augmentation to address the problems of imbalanced data distribution and small samples of rare classes in few-shot learning. Arthaud et al. [78] use contextual augmentation to create new samples to train a pretrained machine translation model that can accurately translate previously unseen words on the basis of a few examples. Chen et al. [63] propose PromptDA generative augmentation to obtain multiple label words for few-shot text classification tasks. According to Wei et al. [79], data augmentation improves curriculum learning in triplet networks for few-shot text classification tasks. FlipDA [59] aims to produce label-flipped data as they found label-flipped data to be more effective than label-preserved data in enhancing the performance of few-shot learning.

IV. Downstream Tasks

In this section, we discuss some common NLP tasks involving data augmentation, i.e., sentence representation learning, text classification, question answering, and sequence tagging tasks.

[1]https://github.com/explosion/spaCy

A. Sentence Representation

Learning sentence representations has long been a fundamental and important research direction in NLP. Sentence representation aims to learn key semantic and syntactic information about sentences. Most existing work on sentence representation learning involving data augmentation is based on contrastive learning, a metric learning method that performs well in learning representations.

For example, [80] considers any two integrations of word deletion, span deletion, span swap, and synonym replacement to form a stronger augmentation for sentence representation learning. SimCSE [72] achieves excellent performance in the seven semantic textual similarity (STS) tasks using dropout augmentation. ESimCSE [31] argues that all the positive sentence embeddings constructed by SimCSE have the same length, which may mislead the model into viewing this as a distinctive feature to differentiate positives from negative instances. To address this issue, they propose a novel data augmentation method, word repetition, along with dropout augmentation, to improve the performance of sentence learning. ConSERT [32] selects randomly two data augmentation approaches for contrastive representation learning: token shuffling, token cutoff, feature cutoff, and dropout [81], with token shuffling and feature cutoff yielding the best results for positive pairs.

B. Text Classification

Text classification is the most simple and fundamental NLP task. It aims to train a text classifier that can automatically analyze text and then assign a predefined label based on the content of the text. Text classification covers a wide range of tasks such as sentiment analysis, topic detection, text matching, etc. Simple EDA augmentation [30], and AEDA augmentation [33] can both be used to produce augmented samples to improve the performance of text classification. [28] substitute the unimportant words with the predicted words generated by Auto-Encoding model or Seq2Seq model without altering the aspect-level polarity for data augmentation to improve aspect-based sentiment analysis.

For few-shot text classification, [82] investigates data augmentation methods that work in the feature space and combine supervised and unsupervised representation learning methods to improve classification performance. MEDA [83] is proposed based on meta-learning, this data augmentation framework is made up of one ball generator and one meta-learner, with the ball generator being used to increase the amount of shots per class via producing more examples, allowing the meta-learner to be trained with both original and augmented examples. Experimental results show that MEDA greatly improves the performance of meta-learning in the classification of a small number of texts.

For contrastive text classification, [60] proposes cross-channel data augmentation to raise the number of training samples and also to provide more diverse samples with multi-aspects. It employs contrastive learning to learn and capture the sentiment representations of various aspects to improve the performance of aspect-based sentiment analysis.

C. Question Answering

Question answering is the task of providing appropriate answers to given questions. It retrieves the answers to questions from a given text, which is very useful for searching for answers in documents. [84] demonstrates that the SQuAD benchmarks for reading comprehension significantly improve when contextual paraphrases are produced through back translation. [85] explores back translation based on query and context paraphrases for domain-agnostic question answering. [86] centers on data augmentation using distant supervision techniques to construct datasets that more closely resemble the types of passages readers see when reasoning to address open domain question answering. [87] propose XLDA, a cross-lingual data augmentation technique that enhances the performance of model on the SQuAD question answering task by substituting a section of the input text with the translation in different language. [88] uses labeled training data, in conjunction with logical and linguistic knowledge for augmentation, significantly improving a range of question-answering tasks. In order to improve zero-shot cross-lingual question answering, [89] makes use of question generation models to generate samples in other languages. While [90] employs back translation to convert question-answer pairings into multiple different languages to enhance the performance of cross-lingual open-retrieval question answering.

D. Sequence Tagging

Sequence tagging is a problem where the model sees a sequence of words or tokens and is expected to output a tag for each word in the sequence. To put it another way, the model is anticipated to tag the entire sequence with a suitable tag drawn from a pre-existing tag dictionary. Applications of sequence tagging in NLU include named entity recognition (NER) and part of speech (POS) tagging. NER is an information extraction technique designed to identify named entities in a given sequence of text tokens (words). POS tagging is a text data processing technique that tags words in a sentence with proper POS based on their semantic and contextual content.

Sahin et al. [4] use sentence cropping and sentence rotating to generate synthetic data for POS tagging. [69] leverage the generative augmentation with LSTM as a sentence generator on the given label for NER and POS tagging. [27] employs label-wise token replacement and synonym replacement for NER. With the help of MELM, [71] creates novel entities in high-quality augmented data, enhancing NER performance by supplying rich entity regularity knowledge. [70] translates the training data into other languages to produce augmented data in multiple languages for cross-lingual NER. [64] designs two generative data augmentation startegies for low-resource NER using the prompting approach along with the BERT model.

V. Challenges

Data augmentation has made great progress in the past few years. Despite these successes, there are still challenges that can be explored further. In this section, we discuss these challenges and suggest directions for future research.

A. Theoretical Explanation of Text Data Augmentation

The effectiveness of data augmentation in NLP has been demonstrated in a large number of experiments [30], [71], [88], [70], [34], but few studies have theoretically investigated how data augmentation works. Several recent studies [91], [92] have investigated and analyzed how data augmentation helps capture features. However, these studies have focused on images because image data can be represented by sparse coding models [93] or spike covariance models [94]. However, because the text is discrete, comparable theoretical studies in NLP are still lacking.

B. Trade-off between Computing Resources and Augmentation Effects

With the advancement of data augmentation, a variety of data augmentation methods have been proposed, especially generative augmentation strategies based on large-scale pretrained language models. These generative augmentation approaches usually show better performance than random augmentation methods in improving the model, as they are designed for specific tasks. For instance, recent studies from [60], [61] show that the generative augmentation method using large-scale PLM as the generator is obviously superior to augmentation methods like EDA and back translation in aspect-based sentiment analysis [60] and low resource NLU [61] tasks. Despite success in improving model performance, these generative augmentation approaches based on large-scale PLMs typically consume more computational resources and time. Therefore, the development of data augmentation strategies that are effective and consume little computational resources could be considered in the future.

C. Generative Augmentation without Label

The majority of current generative augmentation methods perform well in producing high-quality augmented samples. However, these generative augmentation methods usually require a label or prompt to help the generator generate appropriate sentences. This limits the application of these generative augmentation methods in the unsupervised text domain. The text data augmentation strategy proposed in Prompt-BERT [65] prepends different prompt templates at the beginning of the same sentence and then feeds them to the sentence encoder to obtain sentence representations as augmented samples. This augmentation method is a generative augmentation method that uses prompts and does not use labels. Future work could therefore consider how to develop unsupervised generative augmentation methods from this perspective.

VI. Conclusion

In this paper, we present a comprehensive and brief survey of recent data augmentation approaches in NLP. We discuss the benefits of data augmentation and common representative methods for textual data augmentation techniques, and classify these methods into three categories: semantic-invariant augmentation, random augmentation, and generative augmentation. In addition, we conclude the main application

scenarios and downstream application tasks for data augmentation. Finally, we outline the challenges in the field of textual data augmentation and show that there is still a lot of room to be further exploited. Overall, we hope that this paper will provide a novel perspective on current text data augmentation techniques and inspire more effective data augmentation approaches to be devised.

REFERENCES

[1] H. Qi, M. Brown, and D. G. Lowe, "Low-shot learning with imprinted weights," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5822–5830.

[2] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*. PMLR, 2016, pp. 1842–1850.

[3] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[4] G. G. Şahin and M. Steedman, "Data augmentation via dependency tree morphing for low-resource languages," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Oct.-Nov. 2018, pp. 5004–5009.

[5] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[6] H. Guo, Y. Mao, and R. Zhang, "Augmenting data with mixup for sentence classification: An empirical study," *arXiv preprint arXiv:1905.08941*, 2019.

[7] J. Chen, Z. Yang, and D. Yang, "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*. Association for Computational Linguistics, July 2020, pp. 2147–2157.

[8] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Oct. 2014, pp. 1746–1751.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*. Association for Computational Linguistics, June 2019, pp. 4171–4186.

[10] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *Journal of big Data*, vol. 8, no. 1, pp. 1–34, 2021.

[11] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A survey on data augmentation for text classification," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–39, 2022.

[12] J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang, "An empirical survey of data augmentation for limited data learning in nlp," *arXiv preprint arXiv:2106.07499*, 2021.

[13] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for NLP," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Aug. 2021, pp. 968–988.

[14] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in natural language processing: A survey," *AI Open*, 2022.

[15] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020.

[16] J. Chen, Y. Wu, and D. Yang, "Semi-supervised models via data augmentationfor classifying interactive affective responses," *arXiv preprint arXiv:2004.10972*, 2020.

[17] A. Kumar, S. Bhattamishra, M. Bhandari, and P. Talukdar, "Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation," 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)*. Association for Computational Linguistics, June 2019, pp. 3609–3619.

[18] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2016, pp. 86–96.

[19] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Oct.-Nov. 2018, pp. 489–500.

[20] A. Prakash, S. A. Hasan, K. Lee, V. Datla, A. Qadir, J. Liu, and O. Farri, "Neural paraphrase generation with stacked residual LSTM networks," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Dec. 2016, pp. 2923–2934.

[21] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text 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 1: Long Papers)*. Association for Computational Linguistics, July 2015, pp. 1681–1691.

[22] A. Gupta, A. Agarwal, P. Singh, and P. Rai, "A deep generative framework for paraphrase generation," in *Proceedings of the aaai conference on artificial intelligence*, vol. 32, no. 1, 2018.

[23] O. Kolomiyets, S. Bethard, and M.-F. Moens, "Model-portability experiments for textual temporal analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, vol. 2. ACL; East Stroudsburg, PA, 2011, pp. 271–276.

[24] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[25] W. Y. Wang and D. Yang, "That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2557–2563.

[26] J. Raiman and J. Miller, "Globally normalized reader," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Sept. 2017, pp. 1059–1069.

[27] X. Dai and H. Adel, "An analysis of simple data augmentation for named entity recognition," in *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Dec. 2020, pp. 3861–3867.

[28] T.-W. Hsu, C.-C. Chen, H.-H. Huang, and H.-H. Chen, "Semantics-preserved data augmentation for aspect-based sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2021, pp. 4417–4422.

[29] D. Wang, N. Ding, P. Li, and H. Zheng, "CLINE: Contrastive learning with semantic negative examples for natural language understanding," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 2332–2342.

[30] J. Wei and K. Zou, "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)*. Association for Computational Linguistics, Nov. 2019, pp. 6382–6388.

[31] X. Wu, C. Gao, L. Zang, J. Han, Z. Wang, and S. Hu, "ESimCSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding," in *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Oct. 2022, pp. 3898–3907.

[32] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, "ConSERT: A contrastive framework for self-supervised sentence representation transfer," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 5065–5075.

[33] A. Karimi, L. Rossi, and A. Prati, "AEDA: An easier data augmentation technique for text classification," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Nov. 2021, pp. 2748–2754.

[34] L. Sun, C. Xia, W. Yin, T. Liang, P. Yu, and L. He, "Mixup-transformer: Dynamic data augmentation for NLP tasks," in *Proceedings of the 28th*

International Conference on Computational Linguistics. International Committee on Computational Linguistics, Dec. 2020, pp. 3436–3440.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. Liu, "Freelb: Enhanced adversarial training for natural language understanding," in *International Conference on Learning Representations*, 2020.

[37] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, "SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 2177–2190.

[38] Y. Cheng, L. Jiang, W. Macherey, and J. Eisenstein, "AdvAug: Robust adversarial augmentation for neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 5961–5970.

[39] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[40] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[41] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[43] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[44] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[45] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[46] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer." *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[47] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Aug. 2016, pp. 10–21.

[48] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *International conference on machine learning*. PMLR, 2017, pp. 1587–1596.

[49] K. Guu, T. B. Hashimoto, Y. Oren, and P. Liang, "Generating sentences by editing prototypes," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 437–450, 2018.

[50] N. Malandrakis, M. Shen, A. Goyal, S. Gao, A. Sethi, and A. Metallinou, "Controlled text generation for data augmentation in intelligent artificial agents," in *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Association for Computational Linguistics, Nov. 2019, pp. 90–98.

[51] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, June 2018, pp. 1875–1885.

[52] J. Xu, X. Ren, J. Lin, and X. Sun, "Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3940–3949.

[53] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, June 2018, pp. 452–457.

[54] L. Zhuang, L. Wayne, S. Ya, and Z. Jun, "A robustly optimized BERT pre-training approach with post-training," in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Chinese Information Processing Society of China, Aug. 2021, pp. 1218–1227.

[55] Y. Yang, C. Malaviya, J. Fernandez, S. Swayamdipta, R. Le Bras, J.-P. Wang, C. Bhagavatula, Y. Choi, and D. Downey, "Generative data augmentation for commonsense reasoning," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Nov. 2020, pp. 1008–1025.

[56] X. Wu, S. Lv, L. Zang, J. Han, and S. Hu, "Conditional bert contextual augmentation," in *International conference on computational science*. Springer, 2019, pp. 84–95.

[57] V. Kumar, A. Choudhary, and E. Cho, "Data augmentation using pre-trained transformer models," in *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*. Association for Computational Linguistics, Dec. 2020, pp. 18–26.

[58] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, and N. Zwerdling, "Do not have enough data? deep learning to the rescue!" in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7383–7390.

[59] J. Zhou, Y. Zheng, J. Tang, L. Jian, and Z. Yang, "FlipDA: Effective and robust data augmentation for few-shot learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, May 2022, pp. 8646–8665.

[60] B. Wang, L. Ding, Q. Zhong, X. Li, and D. Tao, "A contrastive cross-channel data augmentation framework for aspect-based sentiment analysis," in *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Oct. 2022, pp. 6691–6704.

[61] Y. Wang, C. Xu, Q. Sun, H. Hu, C. Tao, X. Geng, and D. Jiang, "PromDA: Prompt-based data augmentation for low-resource NLU tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, May 2022, pp. 4242–4255.

[62] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, and W. Park, "GPT3Mix: Leveraging large-scale language models for text augmentation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Nov. 2021, pp. 2225–2239.

[63] C. Chen and K. Shu, "Promptda: Label-guided data augmentation for prompt-based few shot learners," *arXiv preprint arXiv:2205.09229*, 2022.

[64] J. Liu, Y. Chen, and J. Xu, "Low-resource ner by data augmentation with prompting," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 4252–4258.

[65] T. Jiang, S. Huang, Z. Zhang, D. Wang, F. Zhuang, F. Wei, H. Huang, L. Zhang, and Q. Zhang, "Promptbert: Improving bert sentence embeddings with prompts," *arXiv preprint arXiv:2201.04337*, 2022.

[66] M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," *arXiv preprint arXiv:1705.00440*, 2017.

[67] M. Xia, X. Kong, A. Anastasopoulos, and G. Neubig, "Generalized data augmentation for low-resource translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2019, pp. 5786–5796.

[68] Y. Li, X. Li, Y. Yang, and R. Dong, "A diverse data augmentation strategy for low-resource neural machine translation," *Information*, vol. 11, no. 5, p. 255, 2020.

[69] B. Ding, L. Liu, L. Bing, C. Kruengkrai, T. H. Nguyen, S. Joty, L. Si, and C. Miao, "DAGA: Data augmentation with a generation approach for low-resource tagging tasks," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Nov. 2020, pp. 6045–6057.

[70] L. Liu, B. Ding, L. Bing, S. Joty, L. Si, and C. Miao, "MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 5834–5846.

[71] R. Zhou, X. Li, R. He, L. Bing, E. Cambria, L. Si, and C. Miao, "MELM: Data augmentation with masked entity language modeling for low-resource NER," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, May 2022, pp. 2251–2262.

[72] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2021, pp. 6894–6910.

[73] Y.-S. Chuang, R. Dangovski, H. Luo, Y. Zhang, S. Chang, M. Soljacic, S.-W. Li, S. Yih, Y. Kim, and J. Glass, "DiffCSE: Difference-based contrastive learning for sentence embeddings," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, July 2022, pp. 4207–4218.

[74] Y. Zhang, R. He, Z. Liu, L. Bing, and H. Li, "Bootstrapped unsupervised sentence representation learning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 5168–5180.

[75] Y. Qu, D. Shen, Y. Shen, S. Sajeev, W. Chen, and J. Han, "Co{da}: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding," in *International Conference on Learning Representations*, 2021.

[76] Y. Zhang, R. Zhang, S. Mensah, X. Liu, and Y. Mao, "Unsupervised sentence representation via contrastive learning with mixing negatives," 2022.

[77] X. Chao and L. Zhang, "Few-shot imbalanced classification based on data augmentation," *Multimedia Systems*, pp. 1–9, 2021.

[78] F. Arthaud, R. Bawden, and A. Birch, "Few-shot learning through contextual data augmentation," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Apr. 2021, pp. 1049–1062.

[79] J. Wei, C. Huang, S. Vosoughi, Y. Cheng, and S. Xu, "Few-shot text classification with triplet networks, data augmentation, and curriculum learning," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, June 2021, pp. 5493–5500.

[80] Z. Wu, S. Wang, J. Gu, M. Khabsa, F. Sun, and H. Ma, "Clear: Contrastive learning for sentence representation," *arXiv preprint arXiv:2012.15466*, 2020.

[81] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[82] V. Kumar, H. Glaude, C. de Lichy, and W. Campbell, "A closer look at feature space data augmentation for few-shot intent classification," in *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Association for Computational Linguistics, Nov. 2019, pp. 1–10.

[83] P. Sun, Y. Ouyang, W. Zhang, and X. Dai, "Meda: Meta-learning with data augmentation for few-shot text classification." in *IJCAI*, 2021, pp. 3929–3935.

[84] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "Qanet: Combining local convolution with global self-attention for reading comprehension," *arXiv preprint arXiv:1804.09541*, 2018.

[85] S. Longpre, Y. Lu, Z. Tu, and C. DuBois, "An exploration of data augmentation and sampling techniques for domain-agnostic question answering," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, Nov. 2019, pp. 220–227.

[86] W. Yang, Y. Xie, L. Tan, K. Xiong, M. Li, and J. Lin, "Data augmentation for bert fine-tuning in open-domain question answering," *arXiv preprint arXiv:1904.06652*, 2019.

[87] J. Singh, B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "Xlda: Cross-lingual data augmentation for natural language inference and question answering," *arXiv preprint arXiv:1905.11471*, 2019.

[88] A. Asai and H. Hajishirzi, "Logic-guided data augmentation and regularization for consistent question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 5642–5650.

[89] A. Riabi, T. Scialom, R. Keraron, B. Sagot, D. Seddah, and J. Staiano, "Synthetic data augmentation for zero-shot cross-lingual question answering," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2021, pp. 7016–7030.

[90] C.-C. Hung, T. Green, R. Litschko, T. Tsereteli, S. Takeshita, M. Bombieri, G. Glavaš, and S. P. Ponzetto, "ZusammenQA: Data augmentation with specialized models for cross-lingual open-retrieval question answering system," in *Proceedings of the Workshop on Multilingual Information Access (MIA)*. Association for Computational Linguistics, July 2022, pp. 77–90.

[91] Z. Wen and Y. Li, "Toward understanding the feature learning process of self-supervised contrastive learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 112–11 122.

[92] W. Ji, Z. Deng, R. Nakada, J. Zou, and L. Zhang, "The power of contrast for feature learning: A theoretical analysis," *arXiv preprint arXiv:2110.02473*, 2021.

[93] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[94] Z. Bai and J. Yao, "On sample eigenvalues in a generalized spiked population model," *Journal of Multivariate Analysis*, vol. 106, pp. 167–177, 2012.