POLYPUBLIE

A survey of data augmentation approaches for NLP

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar Anbil Parthipan, Soroush Vosoughi, Teruko Mitamura, Eduard Hovy

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

Date	2021
Type	Communication de conférence / Conference or Workshop Item Communication écrite/Paper
Version	Version officielle de l'éditeur / Published version
URL de PolyPublie **PolyPublie URL**	https://publications.polymtl.ca/54293/
Évaluation par les pairs **Peer reviewed**	Révisé par les pairs / Refereed
Conditions d'utilisation **Terms of use**	Creative Commons Attribution 4.0 International (CC BY)
Référence **Citation**	Feng, S. Y., Gangal, V., Wei, J., Anbil Parthipan, S. C., Vosoughi, S., Mitamura, T., & Hovy, E. (août 2021). A survey of data augmentation approaches for NLP [Communication écrite]. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), En ligne. https://doi.org/10.18653/v1/2021.findings-acl.84

Publication officielle
Official publication

Nom de la conférence **Conference name**	59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)
Date et lieu **Date and location**	2021-08-01 - 2021-08-06, En ligne
Maison d'édition **Publisher**	Association for Computational Linguistics
URL officielle **Official URL**	https://doi.org/10.18653/v1/2021.findings-acl.84
Mention légale **Legal notice**	©2021 Association for Computational Linguistics Materials published in or after 2016 are licensed on a Creative Commons Attribution 4.0 International License.

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal
https://publications.polymtl.ca

A Survey of Data Augmentation Approaches for NLP

Steven Y. Feng[*1] **Varun Gangal**[*1] **Jason Wei**[†2] **Sarath Chandar**[3]
Soroush Vosoughi[4] **Teruko Mitamura**[1] **Eduard Hovy**[1]
[1]Carnegie Mellon University, [2]Google Research
[3]Mila - Quebec AI Institute, [4]Dartmouth College
{syfeng,vgangal,teruko,hovy}@cs.cmu.edu
jasonwei@google.com sarath.chandar@mila.quebec
soroush@dartmouth.edu

Abstract

Data augmentation has recently seen increased interest in NLP due to more work in low-resource domains, new tasks, and the popularity of large-scale neural networks that require large amounts of training data. Despite this recent upsurge, this area is still relatively underexplored, perhaps due to the challenges posed by the discrete nature of language data. In this paper, we present a comprehensive and unifying survey of data augmentation for NLP by summarizing the literature in a structured manner. We first introduce and motivate data augmentation for NLP, and then discuss major methodologically representative approaches. Next, we highlight techniques that are used for popular NLP applications and tasks. We conclude by outlining current challenges and directions for future research. Overall, our paper aims to clarify the landscape of existing literature in data augmentation for NLP and motivate additional work in this area. We also present a GitHub repository with a paper list that will be continuously updated at https://github.com/styfeng/DataAug4NLP.

1 Introduction

Data augmentation (DA) refers to strategies for increasing the diversity of training examples without explicitly collecting new data. It has received active attention in recent machine learning (ML) research in the form of well-received, general-purpose techniques such as UDA (Xie et al., 2020) (3.1) and MIXUP (Zhang et al., 2017) (3.2). These are often first explored in computer vision (CV), and DA's adaptation for natural language processing (NLP) seems secondary and comparatively underexplored, perhaps due to challenges presented by the discrete nature of language, which rules out continuous noising and makes it hard to maintain invariance.

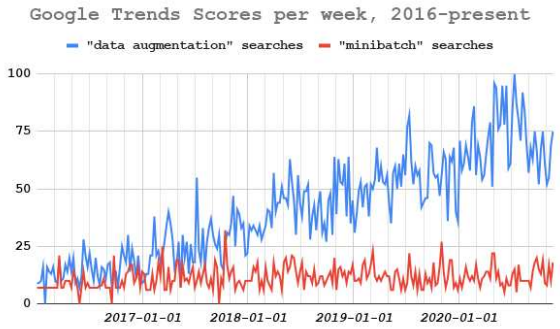


Figure 1: Weekly Google Trends scores for the search term *"data augmentation"*, with a control, uneventful ML search term (*"minibatch"*) for comparison.

Despite these challenges, there has been increased interest and demand for DA for NLP. As NLP grows due to off-the-shelf availability of large pretrained models, there are increasingly more tasks and domains to explore. Many of these are low-resource, and have a paucity of training examples, creating many use-cases for which DA can play an important role. Particularly, for many non-classification NLP tasks such as span-based tasks and generation, DA research is relatively sparse despite their ubiquity in real-world settings.

Our paper aims to sensitize the NLP community towards this growing area of work, which has also seen increasing interest in ML overall (as seen in Figure 1). As interest and work on this topic continue to increase, this is an opportune time for a paper of our kind to (i) give a bird's eye view of DA for NLP, and (ii) identify key challenges to effectively motivate and orient interest in this area. To the best of our knowledge, this is the first survey to take a detailed look at DA methods for NLP.[1]

This paper is structured as follows. Section

[*] Equal contribution by the two authors.
[†] AI Resident.

[1] Liu et al. (2020a) present a smaller-scale text data augmentation survey that is concise and focused. Our work serves as a more comprehensive survey with larger coverage and is more up-to-date.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 968–988
August 1–6, 2021. ©2021 Association for Computational Linguistics

2 discusses what DA is, its goals and trade-offs, and why it works. Section 3 describes popular methodologically representative DA techniques for NLP—which we categorize into rule-based (3.1), example interpolation-based (3.2), or model-based (3.3). Section 4 discusses useful NLP applications for DA, including low-resource languages (4.1), mitigating bias (4.2), fixing class imbalance (4.3), few-shot learning (4.4), and adversarial examples (4.5). Section 5 describes DA methods for common NLP tasks including summarization (5.1), question answering (5.2), sequence tagging tasks (5.3), parsing tasks (5.4), grammatical error correction (5.5), neural machine translation (5.6), data-to-text NLG (5.7), open-ended and conditional text generation (5.8), dialogue (5.9), and multimodal tasks (5.10). Finally, Section 6 discusses challenges and future directions in DA for NLP. Appendix A lists useful blog posts and code repositories.

Through this work, we hope to emulate past papers which have surveyed DA methods for other types of data, such as images (Shorten and Khoshgoftaar, 2019), faces (Wang et al., 2019b), and time series (Iwana and Uchida, 2020). We hope to draw further attention, elicit broader interest, and motivate additional work in DA, particularly for NLP.

2 Background

What is data augmentation? Data augmentation (DA) encompasses methods of increasing training data diversity without directly collecting more data. Most strategies either add slightly modified copies of existing data or create synthetic data, aiming for the augmented data to act as a regularizer and reduce overfitting when training ML models (Shorten and Khoshgoftaar, 2019; Hernández-García and König, 2020). DA has been commonly used in CV, where techniques like *cropping*, *flipping*, and *color jittering* are a standard component of model training. In NLP, where the input space is discrete, how to generate effective augmented examples that capture the desired invariances is less obvious.

What are the goals and trade-offs? Despite challenges associated with text, many DA techniques for NLP have been proposed, ranging from rule-based manipulations (Zhang et al., 2015) to more complicated generative approaches (Liu et al., 2020b). As DA aims to provide an alternative to collecting more data, an ideal DA technique should be both easy-to-implement and improve model performance. Most offer trade-offs between these two.

Rule-based techniques are easy-to-implement but usually offer incremental performance improvements (Li et al., 2017; Wei and Zou, 2019; Wei et al., 2021b). Techniques leveraging trained models may be more costly to implement but introduce more data variation, leading to better performance boosts. Model-based techniques customized for downstream tasks can have strong effects on performance but be difficult to develop and utilize.

Further, the distribution of augmented data should neither be too similar nor too different from the original. This may lead to greater overfitting or poor performance through training on examples not representative of the given domain, respectively. Effective DA approaches should aim for a balance.

Kashefi and Hwa (2020) devise a KL-Divergence-based unsupervised procedure to *preemptively* choose among DA heuristics, rather than a typical "run-all-heuristics" comparison, which can be very time and cost intensive.

Interpretation of DA Dao et al. (2019) note that *"data augmentation is typically performed in an ad-hoc manner with little understanding of the underlying theoretical principles"*, and claim the typical explanation of DA as *regularization* to be insufficient. Overall, there indeed appears to be a lack of research on *why* exactly DA works. Existing work on this topic is mainly surface-level, and rarely investigates the theoretical underpinnings and principles. We discuss this challenge more in §6, and highlight some of the existing work below.

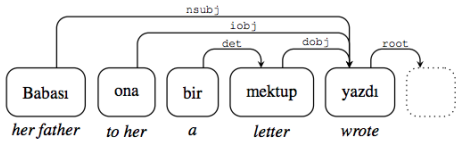
Bishop (1995) show training with noised examples is reducible to Tikhonov regularization (subsumes L2). Rajput et al. (2019) show that DA can increase the positive margin for classifiers, but only when augmenting exponentially many examples for common DA methods.

Dao et al. (2019) think of DA transformations as kernels, and find two ways DA helps: averaging of features and variance regularization. Chen et al. (2020d) show that DA leads to variance reduction by averaging over orbits of the group that keep the data distribution approximately invariant.

3 Techniques & Methods

We now discuss some methodologically representative DA techniques which are relevant to all tasks via the extensibility of their formulation.[2]

[2]Table 1 compares several DA methods by various aspects relating to their applicability, dependencies, and requirements.



(a) Dependency analysis

(1) Babası yazdı *(Her father he-wrote)*
(2) Ona yazdı *(He-wrote to her)*
(3) Bir mektup yazdı *(He-wrote a letter)*

(b) Sentence Cropping

(1) Babası yazdı bir mektup ona (SVOIO)
(2) Yazdı babası ona bir mektup (VSIOO)
(3) Bir mektup yazdı babası ona (OVSIO)
(4) Ona bir mektup yazdı babası (IOOVS)

(c) Sentence Rotating

Figure 2: *Dependency tree morphing* DA applied to a Turkish sentence, Şahin and Steedman (2018)

3.1 Rule-Based Techniques

Here, we cover DA primitives which use easy-to-compute, predetermined transforms sans model components. *Feature space DA* approaches generate augmented examples in the model's feature space rather than input data. Many few-shot learning approaches (Hariharan and Girshick, 2017; Schwartz et al., 2018) leverage estimated feature space "analogy" transformations between examples of known classes to augment for novel classes (see §4.4). Paschali et al. (2019) use iterative affine transformations and projections to maximally "stretch" an example along the class-manifold.

Wei and Zou (2019) propose EASY DATA AUGMENTATION (EDA), a set of token-level random perturbation operations including *random insertion, deletion*, and *swap*. They show improved performance on many text classification tasks. UDA (Xie et al., 2020) show how supervised DA methods can be exploited for unsupervised data through consistency training on $(x, DA(x))$ pairs.

For paraphrase identification, Chen et al. (2020b) construct a signed graph over the data, with individual sentences as nodes and pair labels as signed edges. They use balance theory and transitivity to infer augmented sentence pairs from this graph. Motivated by image cropping and rotation, Şahin and Steedman (2018) propose *dependency tree morphing*. For dependency-annotated sentences, children of the same parent are swapped (à la rotation) or some deleted (à la cropping), as seen in Figure 2. This is most beneficial for language families with rich case marking systems (e.g. *Baltic* and *Slavic*).

3.2 Example Interpolation Techniques

Another class of DA techniques, pioneered by MIXUP (Zhang et al., 2017), interpolates the inputs and labels of two or more real examples. This class of techniques is also sometimes referred to as *Mixed Sample Data Augmentation* (MSDA). Ensuing work has explored interpolating inner components (Verma et al., 2019; Faramarzi et al., 2020), more general mixing schemes (Guo, 2020), and adding adversaries (Beckham et al., 2019).

Another class of extensions of MIXUP which has been growing in the vision community attempts to fuse raw input image pairs together into a single input image, rather than improve the continuous interpolation mechanism. Examples of this paradigm include CUTMIX (Yun et al., 2019), CUTOUT (DeVries and Taylor, 2017) and COPY-PASTE (Ghiasi et al., 2020). For instance, CUTMIX replaces a small sub-region of Image A with a patch sampled from Image B, with the labels mixed in proportion to sub-region sizes. There is potential to borrow ideas and inspiration from these works for NLP, e.g. for multimodal work involving both images and text (see *"Multimodal challenges"* in §6).

A bottleneck to using MIXUP for NLP tasks was the requirement of continuous inputs. This has been overcome by mixing embeddings or higher hidden layers (Chen et al., 2020c). Later variants propose speech-tailored mixing schemes (Jindal et al., 2020b) and interpolation with adversarial examples (Cheng et al., 2020), among others.

SEQ2MIXUP (Guo et al., 2020) generalizes MIXUP for sequence transduction tasks in two ways - the "hard" version samples a binary mask (from a Bernoulli with a $\beta(\alpha, \alpha)$ prior) and picks from one of two sequences at each token position, while the "soft" version softly interpolates between sequences based on a coefficient sampled from $\beta(\alpha, \alpha)$. The "soft" version is found to outperform the "hard" version and earlier interpolation-based techniques like SWITCHOUT (Wang et al., 2018a).

3.3 Model-Based Techniques

Seq2seq and language models have also been used for DA. The popular BACKTRANSLATION method (Sennrich et al., 2016) translates a sequence into another language and then back into the original language. Kumar et al. (2019a) train seq2seq models with their proposed method DiPS which learns to generate diverse paraphrases of input text using a modified decoder with a submodular objective,

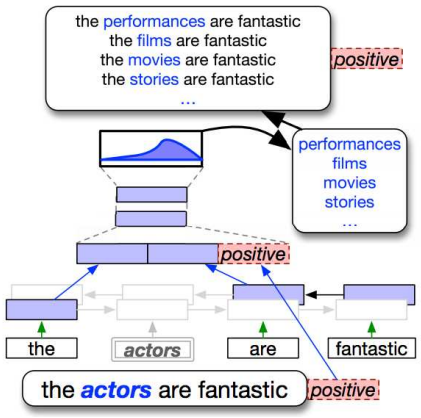


Figure 3: *Contextual Augmentation*, Kobayashi (2018)

and show its effectiveness as DA for several classification tasks. Pretrained language models such as RNNs (Kobayashi, 2018) and transformers (Yang et al., 2020) have also been used for augmentation.

Kobayashi (2018) generate augmented examples by replacing words with others randomly drawn according to the recurrent language model's distribution based on the current context (illustration in Figure 3). Yang et al. (2020) propose G-DAUG^c which generates synthetic examples using pretrained transformer language models, and selects the most informative and diverse set for augmentation. Gao et al. (2019) advocate retaining the full distribution through "soft" augmented examples, showing gains on machine translation.

Nie et al. (2020) augment word representations with a context-sensitive attention-based mixture of their semantic neighbors from a pretrained embedding space, and show its effectiveness for NER on social media text. Inspired by denoising autoencoders, Ng et al. (2020) use a corrupt-and-reconstruct approach, with the corruption function $q(x'|x)$ masking an arbitrary number of word positions and the reconstruction function $r(x|x')$ unmasking them using BERT (Devlin et al., 2019). Their approach works well on domain-shifted test sets across 9 datasets on sentiment, NLI, and NMT.

Feng et al. (2019) propose a task called SEMANTIC TEXT EXCHANGE (STE) which involves adjusting the overall semantics of a text to fit the context of a new word/phrase that is inserted called the *replacement entity (RE)*. They do so by using a system called SMERTI and a masked LM approach. While not proposed directly for DA, it can be used as such, as investigated in Feng et al. (2020).

Rather than starting from an existing example and modifying it, some model-based DA approaches directly estimate a generative process from the training set and sample from it. Anaby-Tavor et al. (2020) learn a label-conditioned generator by finetuning GPT-2 (Radford et al., 2019) on the training data, using this to generate candidate examples per class. A classifier trained on the original training set is then used to select top k candidate examples which confidently belong to the respective class for augmentation. Quteineh et al. (2020) use a similar label-conditioned GPT-2 generation method, and demonstrate its effectiveness as a DA method in an active learning setup.

Other approaches include syntactic or controlled paraphrasing (Iyyer et al., 2018; Kumar et al., 2020), document or story-level paraphrasing (Gangal et al., 2021), augmenting misclassified examples (Dreossi et al., 2018), BERT cross-encoder labeling of new inputs (Thakur et al., 2021), and guided generation using large-scale generative language models (Liu et al., 2020b,c). Models can also learn to combine together simpler DA primitives (Cubuk et al., 2018; Ratner et al., 2017) or add human-in-the-loop (Kaushik et al., 2020, 2021).

4 Applications

In this section, we discuss several DA methods for some common NLP applications.[2]

4.1 Low-Resource Languages

Low-resource languages are an important and challenging application for DA, typically for neural machine translation (NMT). Techniques using external knowledge such as WordNet (Miller, 1995) may be difficult to use effectively here.[3] There are ways to leverage high-resource languages for low-resource languages, particularly if they have similar linguistic properties. Xia et al. (2019) use this approach to improve low-resource NMT.

Li et al. (2020b) use *backtranslation* and self-learning to generate augmented training data. Inspired by work in CV, Fadaee et al. (2017) generate additional training examples that contain low-frequency (rare) words in synthetically created contexts. Qin et al. (2020) present a DA framework to generate multi-lingual code-switching data to fine-tune multilingual-BERT. It encourages the alignment of representations from source and multiple target languages once by mixing their context information. They see improved performance across 5 tasks with 19 languages.

[3]Low-resource language challenges discussed more in §6.

DA Method	Ext.Know	Pretrained	Preprocess	Level	Task-Agnostic
SYNONYM REPLACEMENT (Zhang et al., 2015)	✓	×	tok	Input	✓
RANDOM DELETION (Wei and Zou, 2019)	×	×	tok	Input	✓
RANDOM SWAP (Wei and Zou, 2019)	×	×	tok	Input	✓
BACKTRANSLATION (Sennrich et al., 2016)	×	✓	Depends	Input	✓
SCPN (Wieting and Gimpel, 2017)	×	✓	const	Input	✓
SEMANTIC TEXT EXCHANGE (Feng et al., 2019)	×	✓	const	Input	✓
CONTEXTUALAUG (Kobayashi, 2018)	×	✓	-	Input	✓
LAMBADA (Anaby-Tavor et al., 2020)	×	✓	-	Input	×
GECA (Andreas, 2020)	×	×	tok	Input	×
SEQMIXUP (Guo et al., 2020)	×	×	tok	Input	×
SWITCHOUT (Wang et al., 2018b)	×	×	tok	Input	×
EMIX (Jindal et al., 2020a)	×	×	-	Emb/Hidden	✓
SPEECHMIX (Jindal et al., 2020b)	×	×	-	Emb/Hidden	Speech/Audio
MIXTEXT (Chen et al., 2020c)	×	×	-	Emb/Hidden	✓
SIGNEDGRAPH (Chen et al., 2020b)	×	×	-	Input	×
DTREEMORPH (Şahin and Steedman, 2018)	×	×	dep	Input	✓
Sub^2 (Shi et al., 2021)	×	×	dep	Input	Substructural
DAGA (Ding et al., 2020)	×	×	tok	Input+Label	×
WN-HYPERS (Feng et al., 2020)	✓	×	const+KWE	Input	✓
SYNTHETIC NOISE (Feng et al., 2020)	×	×	tok	Input	✓
UEDIN-MS (DA part) (Grundkiewicz et al., 2019)	✓	×	tok	Input	✓
NONCE (Gulordava et al., 2018)	✓	×	const	Input	✓
XLDA (Singh et al., 2019)	×	✓	Depends	Input	✓
SEQMIX (Zhang et al., 2020)	×	✓	tok	Input+Label	×
SLOT-SUB-LM (Louvan and Magnini, 2020)	×	✓	tok	Input	✓
UBT & TBT (Vaibhav et al., 2019)	×	✓	Depends	Input	✓
SOFT CONTEXTUAL DA (Gao et al., 2019)	×	✓	tok	Emb/Hidden	✓
DATA DIVERSIFICATION (Nguyen et al., 2020)	×	✓	Depends	Input	✓
DIPS (Kumar et al., 2019a)	×	✓	tok	Input	✓
AUGMENTED SBERT (Thakur et al., 2021)	×	✓	-	Input+Label	Sentence Pairs

Table 1: Comparing a selection of DA methods by various aspects relating to their applicability, dependencies, and requirements. *Ext.Know*, *KWE*, *tok*, *const*, and *dep* stand for External Knowledge, keyword extraction, tokenization, constituency parsing, and dependency parsing, respectively. *Ext.Know* refers to whether the DA method requires external knowledge (e.g. WordNet) and *Pretrained* if it requires a pretrained model (e.g. BERT). *Preprocess* denotes preprocessing required, *Level* denotes the depth at which data is modified by the DA, and *Task-Agnostic* refers to whether the DA method can be applied to different tasks. See Appendix B for further explanation.

4.2 Mitigating Bias

Zhao et al. (2018) attempt to mitigate gender bias in coreference resolution by creating an augmented dataset identical to the original but biased towards the underrepresented gender (using gender swapping of entities such as replacing *"he"* with *"she"*) and train on the union of the two datasets. Lu et al. (2020) formally propose COUNTERFACTUAL DA (CDA) for gender bias mitigation, which involves causal interventions that break associations between gendered and gender-neutral words. Zmigrod et al. (2019) and Hall Maudslay et al. (2019) propose further improvements to CDA. Moosavi et al. (2020) augment training sentences with their corresponding predicate-argument structures, improving the robustness of transformer models against various types of biases.

4.3 Fixing Class Imbalance

Fixing class imbalance typically involves a combination of undersampling and oversampling. SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) (Chawla et al., 2002), which generates augmented minority class examples through interpolation, still remains popular (Fernández et al., 2018). MULTILABEL SMOTE (MLSMOTE) (Charte et al., 2015) modifies SMOTE to balance classes for multi-label classification, where classifiers predict more than one class at the same time. Other techniques such as EDA (Wei and Zou, 2019) can possibly be used for oversampling as well.

4.4 Few-Shot Learning

DA methods can ease few-shot learning by adding more examples for novel classes introduced in the few-shot phase. Hariharan and Girshick (2017) use learned analogy transformations $\phi(z_1, z_2, x)$ between example pairs from a non-novel class $z_1 \rightarrow z_2$ to generate augmented examples $x \rightarrow x'$ for novel classes. Schwartz et al. (2018) generalize this to beyond just linear offsets, through their "Δ-network" autoencoder which learns the distribution

$P(z_2|z_1, C)$ from all $y^*_{z_1} = y^*_{z_2} = C$ pairs, where C is a class and y is the ground-truth labelling function. Both these methods are applied only on image tasks, but their theoretical formulations are generally applicable, and hence we discuss them.

Kumar et al. (2019b) apply these and other DA methods for few-shot learning of novel intent classes in task-oriented dialog. Wei et al. (2021a) show that data augmentation facilitates curriculum learning for training triplet networks for few-shot text classification. Lee et al. (2021) use T5 to generate additional examples for data-scarce classes.

4.5 Adversarial Examples (AVEs)

Adversarial examples can be generated using innocuous label-preserving transformations (e.g. paraphrasing) that fool state-of-the-art NLP models, as shown in Jia et al. (2019). Specifically, they add sentences with distractor spans to passages to construct AVEs for span-based QA. Zhang et al. (2019d) construct AVEs for paraphrase detection using word swapping. Kang et al. (2018) and Glockner et al. (2018) create AVEs for textual entailment using WordNet relations.

5 Tasks

In this section, we discuss several DA works for common NLP tasks.[2] We focus on non-classification tasks as classification is worked on by default, and well covered in earlier sections (e.g. §3 and §4). Numerous previously mentioned DA techniques, e.g. (Wei and Zou, 2019; Chen et al., 2020b; Anaby-Tavor et al., 2020), have been used or can be used for text classification tasks.

5.1 Summarization

Fabbri et al. (2020) investigate *backtranslation* as a DA method for few-shot abstractive summarization with the use of a consistency loss inspired by UDA. Parida and Motlicek (2019) propose an iterative DA approach for abstractive summarization that uses a mix of synthetic and real data, where the former is generated from Common Crawl. Zhu et al. (2019) introduce a query-focused summarization (Dang, 2005) dataset collected using Wikipedia called WIKIREF which can be used for DA. Pasunuru et al. (2021) use DA methods to construct two training datasets for Query-focused Multi-Document Summarization (QMDS) called QMDSCNN and QMDSIR by modifying CNN/DM (Hermann et al., 2015) and mining search-query logs, respectively.

5.2 Question Answering (QA)

Longpre et al. (2019) investigate various DA and sampling techniques for domain-agnostic QA including paraphrasing by *backtranslation*. Yang et al. (2019) propose a DA method using distant supervision to improve BERT finetuning for open-domain QA. Riabi et al. (2020) leverage Question Generation models to produce augmented examples for zero-shot cross-lingual QA. Singh et al. (2019) propose XLDA, or CROSS-LINGUAL DA, which substitutes a portion of the input text with its translation in another language, improving performance across multiple languages on NLI tasks including the SQuAD QA task. Asai and Hajishirzi (2020) use logical and linguistic knowledge to generate additional training data to improve the accuracy and consistency of QA responses by models. Yu et al. (2018) introduce a new QA architecture called QANet that shows improved performance on SQuAD when combined with augmented data generated using backtranslation.

5.3 Sequence Tagging Tasks

Ding et al. (2020) propose DAGA, a two-step DA process. First, a language model over sequences of tags and words linearized as per a certain scheme is learned. Second, sequences are sampled from this language model and de-linearized to generate new examples. Şahin and Steedman (2018), discussed in §3.1, use *dependency tree morphing* (Figure 2) to generate additional training examples on the downstream task of part-of-speech (POS) tagging.

Dai and Adel (2020) modify DA techniques proposed for sentence-level tasks for named entity recognition (NER), including label-wise token and synonym replacement, and show improved performance using both recurrent and transformer models. Zhang et al. (2020) propose a DA method based on MIXUP called SEQMIX for active sequence labeling by augmenting queried samples, showing improvements on NER and Event Detection.

5.4 Parsing Tasks

Jia and Liang (2016) propose DATA RECOMBINATION for injecting task-specific priors to neural semantic parsers. A synchronous context-free grammar (SCFG) is induced from training data, and new "recombinant" examples are sampled. Yu et al. (2020) introduce GRAPPA, a pretraining approach for table semantic parsing, and generate synthetic question-SQL pairs via an SCFG. Andreas (2020)

use *compositionality* to construct synthetic examples for downstream tasks like semantic parsing. Fragments of original examples are replaced with fragments from other examples in similar contexts.

Vania et al. (2019) investigate DA for low-resource dependency parsing including *dependency tree morphing* from Şahin and Steedman (2018) (Figure 2) and modified *nonce* sentence generation from Gulordava et al. (2018), which replaces content words with other words of the same POS, morphological features, and dependency labels.

5.5 Grammatical Error Correction (GEC)

Lack of parallel data is typically a barrier for GEC. Various works have thus looked at DA methods for GEC. We discuss some here, and more can be found in Table 2 in Appendix C.

There is work that makes use of additional resources. Boyd (2018) use German edits from Wikipedia revision history and use those relating to GEC as augmented training data. Zhang et al. (2019b) explore multi-task transfer, or the use of annotated data from other tasks.

There is also work that adds synthetic errors to noise the text. Wang et al. (2019a) investigate two approaches: token-level perturbations and training error generation models with a filtering strategy to keep generations with sufficient errors. Grundkiewicz et al. (2019) use *confusion sets* generated by a spellchecker for noising. Choe et al. (2019) learn error patterns from small annotated samples along with *POS-specific noising*.

There have also been approaches to improve the diversity of generated errors. Wan et al. (2020) investigate noising through editing the latent representations of grammatical sentences, and Xie et al. (2018) use a neural sequence transduction model and beam search noising procedures.

5.6 Neural Machine Translation (NMT)

There are many works which have investigated DA for NMT. We highlighted some in §3 and §4.1, e.g. (Sennrich et al., 2016; Fadaee et al., 2017; Xia et al., 2019). We discuss some further ones here, and more can be found in Table 3 in Appendix C.

Wang et al. (2018a) propose SWITCHOUT, a DA method that randomly replaces words in both source and target sentences with other random words from their corresponding vocabularies. Gao et al. (2019) introduce SOFT CONTEXTUAL DA that softly augments randomly chosen words in a sentence using a contextual mixture of multiple related words over the vocabulary. Nguyen et al. (2020) propose DATA DIVERSIFICATION which merges original training data with the predictions of several forward and backward models.

5.7 Data-to-Text NLG

Data-to-text NLG refers to tasks which require generating natural language descriptions of structured or semi-structured data inputs, e.g. game score tables (Wiseman et al., 2017). Randomly perturbing game score values without invalidating overall game outcome is one DA strategy explored in game summary generation (Hayashi et al., 2019).

Two popular recent benchmarks are E2E-NLG (Dušek et al., 2018) and WebNLG (Gardent et al., 2017). Both involve generation from structured inputs - meaning representation (MR) sequences and triple sequences, respectively. Montella et al. (2020) show performance gains on WebNLG by DA using Wikipedia sentences as targets and parsed OpenIE triples as inputs. Tandon et al. (2018) propose DA for E2E-NLG based on permuting the input MR sequence. Kedzie and McKeown (2019) inject Gaussian noise into a trained decoder's hidden states and sample diverse augmented examples from it. This sample-augment-retrain loop helps performance on E2E-NLG.

5.8 Open-Ended & Conditional Generation

There has been limited work on DA for open-ended and conditional text generation. Feng et al. (2020) experiment with a suite of DA methods for finetuning GPT-2 on a low-resource domain in attempts to improve the quality of generated continuations, which they call GENAUG. They find that WN-HYPERS (WordNet hypernym replacement of keywords) and SYNTHETIC NOISE (randomly perturbing non-terminal characters in words) are useful, and the quality of generated text improves to a peak at $\approx$ 3x the original amount of training data.

5.9 Dialogue

Most DA approaches for dialogue focus on task-oriented dialogue. We outline some below, and more can be found in Table 4 in Appendix C.

Quan and Xiong (2019) present sentence and word-level DA approaches for end-to-end task-oriented dialogue. Louvan and Magnini (2020) propose LIGHTWEIGHT AUGMENTATION, a set of word-span and sentence-level DA methods for low-resource slot filling and intent classification.

Hou et al. (2018) present a seq2seq DA framework to augment dialogue utterances for dialogue language understanding (Young et al., 2013), including a *diversity rank* to produce diverse utterances. Zhang et al. (2019c) propose MADA to generate diverse responses using the property that several valid responses exist for a dialogue context.

There is also DA work for spoken dialogue. Hou et al. (2018), Kim et al. (2019), Zhao et al. (2019), and Yoo et al. (2019) investigate DA methods for dialogue and *spoken language understanding* (SLU), including generative latent variable models.

5.10 Multimodal Tasks

DA techniques have also been proposed for multimodal tasks where aligned data for multiple modalities is required. We look at ones that involve language or text. Some are discussed below, and more can be found in Table 5 in Appendix C.

Beginning with speech, Wang et al. (2020) propose a DA method to improve the robustness of downstream dialogue models to speech recognition errors. Wiesner et al. (2018) and Renduchintala et al. (2018) propose DA methods for end-to-end *automatic speech recognition* (ASR).

Looking at images or video, Xu et al. (2020) learn a cross-modality matching network to produce synthetic image-text pairs for multimodal classifiers. Atliha and Šešok (2020) explore DA methods such as synonym replacement and contextualized word embeddings augmentation using BERT for *image captioning*. Kafle et al. (2017), Yokota and Nakayama (2018), and Tang et al. (2020) propose methods for *visual QA* including question generation and adversarial examples.

6 Challenges & Future Directions

Looking forward, data augmentation faces substantial challenges, specifically for NLP, and with these challenges, new opportunities for future work arise.

Dissonance between empirical novelties and theoretical narrative: There appears to be a conspicuous lack of research on *why* DA works. Most studies might show empirically that a DA technique works and provide some intuition, but it is currently challenging to measure the goodness of a technique without resorting to a full-scale experiment. A recent work in vision (Gontijo-Lopes et al., 2020) has proposed that affinity (the distributional shift caused by DA) and diversity (the complexity of the augmentation) can predict DA performance, but it is unclear how these results might translate to NLP.

Minimal benefit for pretrained models on in-domain data: With the popularization of large pretrained language models, it has recently come to light that a couple of previously effective DA techniques for certain text classification tasks in English (Wei and Zou, 2019; Sennrich et al., 2016) provide little benefit for models like BERT and RoBERTa, which already achieve high performance on in-domain text classification (Longpre et al., 2020). One hypothesis for this could be that using simple DA techniques provides little benefit when finetuning large pretrained transformers on tasks for which examples are well-represented in the pretraining data, but DA methods could still be effective when finetuning on tasks for which examples are scarce or out-of-domain compared with the training data. Further work could study under which scenarios data augmentation for large pretrained models is likely to be effective.

Multimodal challenges: While there has been increased work in multimodal DA, as discussed in §5.10, effective DA methods for multiple modalities has been challenging. Many works focus on augmenting a single modality or multiple ones separately. For example, there is potential to further explore simultaneous image and text augmentation for image captioning, such as a combination of CutMix (Yun et al., 2019) and caption editing.

Span-based tasks offer unique DA challenges as there are typically many correlated classification decisions. For example, random token replacement may be a locally acceptable DA method but possibly disrupt coreference chains for latter sentences. DA techniques here must take into account dependencies between different locations in the text.

Working in specialized domains such as those with domain-specific vocabulary and jargon (e.g. *medicine*) can present challenges. Many pretrained models and external knowledge (e.g. WordNet) cannot be effectively used. Studies have shown that DA becomes less beneficial when applied to out-of-domain data, likely because the distribution of augmented data can substantially differ from the original data (Zhang et al., 2019a; Herzig et al., 2020; Campagna et al., 2020; Zhong et al., 2020).

Working with low-resource languages may present similar difficulties as specialized domains.

Further, DA techniques successful in the high-resource scenario may not be effective for low-resource languages that are of a different language family or very distinctive in linguistic and typological terms. For example, those which are language isolates or lack high-resource cognates.

More vision-inspired techniques: Although many NLP DA methods have been inspired by analogous approaches in CV, there is potential for drawing further connections. Many CV DA techniques motivated by real-world invariances (e.g. many angles of looking at the same object) may have similar NLP interpretations. For instance, *grayscaling* could translate to toning down aspects of the text (e.g. plural to singular, *"awesome"* $\rightarrow$ *"good"*). Morphing a dependency tree could be analogous to rotating an image, and paraphrasing techniques may be analogous to changing perspective. For example, negative data augmentation (NDA) (Sinha et al., 2021) involves creating out-of-distribution samples. It has so far been exclusively explored for CV, but could be investigated for text.

Self-supervised learning: More recently, DA has been increasingly used as a key component of self-supervised learning, particularly in vision (Chen et al., 2020e). In NLP, BART (Lewis et al., 2020) showed that predicting deleted tokens as a pretraining task can achieve similar performance as the masked LM, and ELECTRA (Clark et al., 2020) found that pretraining by predicting corrupted tokens outperforms BERT given the same model size, data, and compute. We expect future work will continue exploring how to effectively manipulate text for both pretraining and downstream tasks.

Offline versus online data augmentation: In CV, standard techniques such as cropping, color jittering, and rotations are typically done stochastically, allowing for DA to be incorporated elegantly into the training pipeline. In NLP, however, it is unclear how to include a lightweight code module to apply DA stochastically. This is because DA techniques for NLP often leverage external resources (e.g. a word dictionary for token substitution or a translation model for backtranslation) that are not easily transferable across model training pipelines. Thus, a common practice for DA in NLP is simply to generate augmented data offline and store it as additional data to be loaded during training.[4] Future work on a lightweight module for online DA in NLP could be fruitful, though another challenge will be determining when such a module will be helpful, which—compared with CV, where the invariances being imposed are well-accepted—can vary substantially across NLP tasks.

Lack of unification is a challenge for the current literature on data augmentation for NLP, and popular methods are often presented in an auxiliary fashion. Whereas there are well-accepted frameworks for DA for CV (e.g. default augmentation libraries in PyTorch, *RandAugment* (Cubuk et al., 2019)), there are no such "generalized" DA techniques for NLP. Further, we believe that DA research would benefit from the establishment of standard and unified benchmark tasks and datasets to compare different augmentation methods.

Good data augmentation practices would help make DA work more accessible and reproducible to the NLP and ML communities. On top of unified benchmark tasks, datasets, and frameworks/libraries mentioned above, other good practices include making code and augmented datasets publicly available, reporting variation among results (e.g. standard deviation across random seeds), and more standardized evaluation procedures. Further, transparent hyperparameter analysis, explicitly stating failure cases of proposed techniques, and discussion of the intuition and theory behind them would further improve the transparency and interpretability of DA techniques.

7 Conclusion

In this paper, we presented a comprehensive and structured survey of data augmentation for natural language processing (NLP). We provided a background about data augmentation and how it works, discussed major methodologically representative data augmentation techniques for NLP, and touched upon data augmentation techniques for popular NLP applications and tasks. Finally, we outlined current challenges and directions for future research, and showed that there is much room for further exploration. Overall, we hope our paper can serve as a guide for NLP researchers to decide on which data augmentation techniques to use, and inspire additional interest and work in this area. Please see the corresponding GitHub repository at https://github.com/styfeng/DataAug4NLP.

[4] See Appendix D.

References

Diego Alves, Askars Salimbajevs, and Mārcis Pinnis. 2020. Data augmentation for pipeline-based speech translation. In *9th International Conference on Human Language Technologies - the Baltic Perspective (Baltic HLT 2020)*, Kaunas, Lithuania.

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 AAAI*, pages 7383–7390.

Jacob Andreas. 2020. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566, Online. Association for Computational Linguistics.

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5642–5650, Online. Association for Computational Linguistics.

Viktar Atliha and Dmitrij Šešok. 2020. Text augmentation using BERT for image captioning. *Applied Sciences*, 10:5978.

Christopher Beckham, Sina Honari, Vikas Verma, Alex M. Lamb, Farnoosh Ghadiri, R Devon Hjelm, Yoshua Bengio, and Chris Pal. 2019. On adversarial mixup resynthesis. In *Advances in Neural Information Processing Systems*, pages 4346–4357.

Chris M. Bishop. 1995. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116.

Adriane Boyd. 2018. Using Wikipedia edits in low resource grammatical error correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium. Association for Computational Linguistics.

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica Lam. 2020. Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 122–132, Online. Association for Computational Linguistics.

F. Charte, Antonio Rivera Rivas, María José Del Jesus, and Francisco Herrera. 2015. Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Guanhua Chen, Yun Chen, Yong Wang, and Victor O.K. Li. 2020a. Lexical-constraint-aware neural machine translation via data augmentation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3587–3593. International Joint Conferences on Artificial Intelligence Organization. Main track.

Hannah Chen, Yangfeng Ji, and David Evans. 2020b. Finding friends and flipping frenemies: Automatic paraphrase dataset augmentation using graph theory. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4741–4751, Online. Association for Computational Linguistics.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020c. 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.

Shuxiao Chen, Edgar Dobriban, and Jane Lee. 2020d. A group-theoretic framework for data augmentation. *Advances in Neural Information Processing Systems*, 33.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020e. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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.

Mara Chinea-Ríos, Álvaro Peris, and Francisco Casacuberta. 2017. Adapting neural machine translation with parallel synthetic data. In *Proceedings of the Second Conference on Machine Translation*, pages 138–147, Copenhagen, Denmark. Association for Computational Linguistics.

Yo Joong Choe, Jiyeon Ham, Kyubyong Park, and Yeoil Yoon. 2019. A neural grammatical error correction system built on better pre-training and sequential transfer learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–227,

Florence, Italy. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. *Proceedings of ICLR*.

Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. 2018. Autoaugment: Learning augmentation policies from data. *Proceedings of CVPR*.

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2019. Randaugment: Practical data augmentation with no separate search. *CoRR*, abs/1909.13719.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hao T. Dang. 2005. Overview of DUC 2005.

Tri Dao, Albert Gu, Alexander J. Ratner, Virginia Smith, Christopher De Sa, and Christopher Ré. 2019. A kernel theory of modern data augmentation. *Proceedings of Machine Learning Research*, 97:1528.

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.

Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint*.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach forLow-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Kurt Keutzer, Alberto L. Sangiovanni-Vincentelli, and Sanjit A. Seshia. 2018. Counterexample-guided data augmentation. In *Proceedings of IJCAI*.

Sufeng Duan, Hai Zhao, Dongdong Zhang, and Rui Wang. 2020. Syntax-aware data augmentation for neural machine translation. *arXiv preprint*.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 322–328, Tilburg University, The Netherlands. Association for Computational Linguistics.

Alexander R. Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2020. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. *arXiv preprint*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Alex Falcon, Oswald Lanz, and Giuseppe Serra. 2020. Data augmentation techniques for the video question answering task. *arXiv preprint*.

Mojtaba Faramarzi, Mohammad Amini, Akilesh Badrinaaraayanan, Vikas Verma, and Sarath Chandar. 2020. Patchup: A regularization technique for convolutional neural networks. *arXiv preprint*.

Mariano Felice. 2016. Artificial error generation for translation-based grammatical error correction. *University of Cambridge Technical Report*.

Steven Y. Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020. GenAug: Data augmentation for finetuning text generators. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 29–42, Online. Association for Computational Linguistics.

Steven Y. Feng, Aaron W. Li, and Jesse Hoey. 2019. Keep calm and switch on! Preserving sentiment and fluency in semantic text exchange. 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 2701–2711, Hong Kong, China. Association for Computational Linguistics.

Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V. Chawla. 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61:863–905.

Jennifer Foster and Oistein Andersen. 2009. GenERRate: Generating errors for use in grammatical error detection. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 82–90, Boulder, Colorado. Association for Computational Linguistics.

Varun Gangal, Steven Y. Feng, Eduard Hovy, and Teruko Mitamura. 2021. Nareor: The narrative reordering problem. *arXiv preprint*.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544, Florence, Italy. Association for Computational Linguistics.

Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Paraphrase augmented task-oriented dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 639–649, Online. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. 2020. Simple copy-paste is a strong data augmentation method for instance segmentation. *arXiv preprint*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Raphael Gontijo-Lopes, Sylvia J. Smullin, Ekin D. Cubuk, and Ethan Dyer. 2020. Tradeoffs in data augmentation: An empirical study. *Proceedings of ICLR*.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 45–52, Florence, Italy. Association for Computational Linguistics.

Milan Gritta, Gerasimos Lampouras, and Ignacio Iacobacci. 2021. Conversation graph: Data augmentation, training, and evaluation for non-deterministic dialogue management. *Transactions of the Association for Computational Linguistics*, 9:36–52.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, Florence, Italy. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Demi Guo, Yoon Kim, and Alexander Rush. 2020. Sequence-level mixed sample data augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5547–5552, Online. Association for Computational Linguistics.

Hongyu Guo. 2020. Nonlinear mixup: Out-of-manifold data augmentation for text classification. In *Proceedings of AAAI*, pages 4044–4051.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Bharath Hariharan and Ross Girshick. 2017. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3018–3027.

Hany Hassan, Mostafa Elaraby, and Ahmed Tawfik. 2017. Synthetic data for neural machine translation of spoken-dialects. *arXiv preprint*.

Hiroaki Hayashi, Yusuke Oda, Alexandra Birch, Ioannis Konstas, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Katsuhito Sudoh. 2019. Findings of the third workshop on neural generation and translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 1–14, Hong Kong. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NeurIPS*, pages 1693–1701.

Alex Hernández-García and Peter König. 2020. Data augmentation instead of explicit regularization. *arXiv preprint*.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Mingyue Niu, and Minghao Yang. 2018. Multimodal continuous emotion recognition with data augmentation using recurrent neural networks. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, AVEC'18, page 57–64, New York, NY, USA. Association for Computing Machinery.

Brian Kenji Iwana and Seiichi Uchida. 2020. An empirical survey of data augmentation for time series classification with neural networks. *arXiv preprint*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. 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)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Berlin, Germany. Association for Computational Linguistics.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. 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 4120–4133.

Amit Jindal, Arijit Ghosh Chowdhury, Aniket Didolkar, Di Jin, Ramit Sawhney, and Rajiv Ratn Shah. 2020a. Augmenting NLP models using latent feature interpolations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6931–6936, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Amit Jindal, Narayanan Elavathur Ranganatha, Aniket Didolkar, Arijit Ghosh Chowdhury, Di Jin, Ramit Sawhney, and Rajiv Ratn Shah. 2020b. Speechmix - augmenting deep sound recognition using hidden space interpolations. In *INTERSPEECH*, pages 861–865.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202, Santiago de Compostela, Spain. Association for Computational Linguistics.

Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard Hovy. 2018. AdvEntuRe: Adversarial training for textual entailment with knowledge-guided examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2418–2428, Melbourne, Australia. Association for Computational Linguistics.

Min-Hyung Kang. 2019. Valar nmt : Vastly lacking resources neural machine translation. *Stanford CS224N*.

Omid Kashefi and Rebecca Hwa. 2020. Quantifying the evaluation of heuristic methods for textual data augmentation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 200–208, Online. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Divyansh Kaushik, Amrith Setlur, Eduard H. Hovy, and Zachary Chase Lipton. 2021. Explaining the efficacy of counterfactually augmented data. In *International Conference on Learning Representations*.

Chris Kedzie and Kathleen McKeown. 2019. A good sample is hard to find: Noise injection sampling and self-training for neural language generation models. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 584–593, Tokyo, Japan. Association for Computational Linguistics.

Hwa-Yeon Kim, Yoon-Hyung Roh, and Young-Kil Kim. 2019. Data augmentation by data noising for open-vocabulary slots in spoken language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 97–102, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Kimn. 2020. A syntactic rule-based framework for parallel data synthesis in japanese gec. Massachusetts Institute of Technology.

Sosuke Kobayashi. 2018. 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)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:329–345.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019a. 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)*, pages 3609–3619, Minneapolis, Minnesota. Association for Computational Linguistics.

Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and Wlliam Campbell. 2019b. 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)*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. *arXiv preprint*.

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.

Daniel Li, Te I, Naveen Arivazhagan, Colin Cherry, and Dirk Padfield. 2020a. Sentence boundary augmentation for neural machine translation robustness. *arXiv preprint*.

Yitong Li, Trevor Cohn, and Timothy Baldwin. 2017. Robust training under linguistic adversity. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 21–27, Valencia, Spain. Association for Computational Linguistics.

Yu Li, Xiao Li, Yating Yang, and Rui Dong. 2020b. A diverse data augmentation strategy for low-resource neural machine translation. *Information*, 11(5).

Zhenhao Li and Lucia Specia. 2019. Improving neural machine translation robustness via data augmentation: Beyond back-translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 328–336, Hong Kong, China. Association for Computational Linguistics.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora generation for grammatical error correction. 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 3291–3301, Minneapolis, Minnesota. Association for Computational Linguistics.

Pei Liu, Xuemin Wang, Chao Xiang, and Weiye Meng. 2020a. A survey of text data augmentation. In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pages 191–195.

Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020b. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9031–9041, Online. Association for Computational Linguistics.

Ruibo Liu, Guangxuan Xu, and Soroush Vosoughi. 2020c. Enhanced offensive language detection through data augmentation. *ICWSM Data Challenge*.

Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. 2019. 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*, pages 220–227, Hong Kong, China. Association for Computational Linguistics.

Shayne Longpre, Yu Wang, and Christopher DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? *arXiv preprint arXiv:2010.01764*.

Samuel Louvan and Bernardo Magnini. 2020. Simple is better! lightweight data augmentation for low resource slot filling and intent classification. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 167–177, Hanoi, Vietnam. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. *Gender Bias in Neural Natural Language Processing*, pages 189–202. Springer International Publishing, Cham.

George A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

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.

Nafise Sadat Moosavi, Marcel de Boer, Prasetya Ajie Utama, and Iryna Gurevych. 2020. Improving robustness by augmenting training sentences with predicate-argument structures. *arXiv preprint*.

Xiangyang Mou, Brandyn Sigouin, Ian Steenstra, and Hui Su. 2020. Multimodal dialogue state tracking by qa approach with data augmentation. *AAAI DSTC8 Workshop*.

Diego Moussallem, Mihael Arčan, Axel-Cyrille Ngonga Ngomo, and Paul Buitelaar. 2019. Augmenting neural machine translation with knowledge graphs. *arXiv preprint*.

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.

Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. In *Advances in Neural Information Processing Systems*, volume 33, pages 10018–10029. Curran Associates, Inc.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named entity recognition for social media texts with semantic augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1383–1391, Online. Association for Computational Linguistics.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018. Multi-source neural machine translation with data augmentation. *15th International Workshop on Spoken Language Translation 2018*.

Shantipriya Parida and Petr Motlicek. 2019. Abstract text summarization: A low resource challenge. 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 5994–5998, Hong Kong, China. Association for Computational Linguistics.

Magdalini Paschali, Walter Simson, Abhijit Guha Roy, Muhammad Ferjad Naeem, Rüdiger Göbl, Christian Wachinger, and Nassir Navab. 2019. Data augmentation with manifold exploring geometric transformations for increased performance and robustness. *arXiv preprint*.

Ramakanth Pasunuru, Asli Celikyilmaz, Michel Galley, Chenyan Xiong, Yizhe Zhang, Mohit Bansal, and Jianfeng Gao. 2021. Data augmentation for abstractive query-focused multi-document summarization. *arXiv preprint*.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China. Association for Computational Linguistics.

Wei Peng, Chongxuan Huang, Tianhao Li, Yun Chen, and Qun Liu. 2020. Dictionary-based data augmentation for cross-domain neural machine translation. *arXiv preprint*.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3853–3860. International Joint Conferences on Artificial Intelligence Organization. Main track.

Jun Quan and Deyi Xiong. 2019. Effective data augmentation approaches to end-to-end task-oriented dialogue. In *2019 International Conference on Asian Language Processing (IALP)*, pages 47–52.

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

Shashank Rajput, Zhili Feng, Zachary Charles, Po-Ling Loh, and Dimitris Papailiopoulos. 2019. Does data augmentation lead to positive margin? In *International Conference on Machine Learning*, pages 5321–5330. PMLR.

AJ Ratner, HR Ehrenberg, Z Hussain, J Dunnmon, and C Ré. 2017. Learning to Compose Domain-Specific Transformations for Data Augmentation. *Advances in Neural Information Processing Systems*, 30:3239–3249.

Adithya Renduchintala, Shuoyang Ding, Matthew Wiesner, and Shinji Watanabe. 2018. Multi-Modal Data Augmentation for End-to-end ASR. *Proc. Interspeech 2018*, pages 2394–2398.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2020. Synthetic Data Augmentation for Zero-Shot Cross-Lingual Question Answering. *arXiv preprint*.

Gözde Gül Şahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009, Brussels, Belgium. Association for Computational Linguistics.

Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex M Bronstein. 2018. δ-encoder: an effective sample synthesis method for few-shot object recognition. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2850–2860.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Haoyue Shi, Karen Livescu, and Kevin Gimpel. 2021. Substructure Substitution: Structured Data Augmentation for NLP. *arXiv preprint arXiv:2101.00411*.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering. *arXiv preprint arXiv:1905.11471*.

Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. 2021. Negative data augmentation. In *International Conference on Learning Representations*.

Xiaohui Song, Liangjun Zang, Yipeng Su, Xing Wu, Jizhong Han, and Songlin Hu. 2020. Data Augmentation for Copy-Mechanism in Dialogue State Tracking. *arXiv preprint arXiv:2002.09634*.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.

Shubhangi Tandon, TS Sharath, Shereen Oraby, Lena Reed, Stephanie Lukin, and Marilyn Walker. 2018. TNT-NLG, System 2: Data repetition and meaning representation manipulation to improve neural generation. *E2E NLG Challenge Sfystem Descriptions*.

Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. 2020. Semantic equivalent adversarial data augmentation for visual question answering. In *European Conference on Computer Vision*, pages 437–453. Springer.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *Proceedings of NAACL*.

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving Robustness of Machine Translation with Synthetic Noise. 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 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. 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 1105–1116, Hong Kong, China. Association for Computational Linguistics.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR.

Zhaohong Wan, Xiaojun Wan, and Wenguang Wang. 2020. Improving Grammatical Error Correction with Data Augmentation by Editing Latent Representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2202–2212, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chencheng Wang, Liner Yang, Yun Chen, Yongping Du, and Erhong Yang. 2019a. Controllable Data Synthesis Method for Grammatical Error Correction. *arXiv preprint arXiv:1909.13302*.

Longshaokan Wang, Maryam Fazel-Zarandi, Aditya Tiwari, Spyros Matsoukas, and Lazaros Polymenakos. 2020. Data augmentation for training dialog models robust to speech recognition errors. *arXiv preprint arXiv:2006.05635*.

Xiang Wang, Kai Wang, and Shiguo Lian. 2019b. A survey on face data augmentation. *arXiv preprint arXiv:1904.11685*.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018a. 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.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018b. 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, Chengyu Huang, Soroush Vosoughi, Yu Cheng, and Shiqi Xu. 2021a. Few-shot text classification with triplet networks, data augmentation, and curriculum learning. *Proceedings of NAACL*.

Jason Wei, Chengyu Huang, Shiqi Xu, and Soroush Vosoughi. 2021b. Text augmentation in a multi-task view. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2888–2894, Online. 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.

Max White and Alla Rozovskaya. 2020. A Comparative Study of Synthetic Data Generation Methods for Grammatical Error Correction. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 198–208, Seattle, WA, USA â†' Online. Association for Computational Linguistics.

Matthew Wiesner, Adithya Renduchintala, Shinji Watanabe, Chunxi Liu, Najim Dehak, and Sanjeev Khudanpur. 2018. Low resource multi-modal data augmentation for end-to-end ASR. *CoRR*.

John Wieting and Kevin Gimpel. 2017. Revisiting Recurrent Networks for Paraphrastic Sentence Embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2078–2088, Vancouver, Canada. Association for Computational Linguistics.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in Data-to-Document Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized Data Augmentation for Low-Resource Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. 2018. Noising and Denoising Natural Language: Diverse Backtranslation for Grammar Correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 619–628, New Orleans, Louisiana. Association for Computational Linguistics.

N. Xu, W. Mao, P. Wei, and D. Zeng. 2020. MDA: Multimodal Data Augmentation Framework for Boosting Performance on Image-Text Sentiment/Emotion Classification Tasks. *IEEE Intelligent Systems*, pages 1–1.

Shuyao Xu, Jiehao Zhang, Jin Chen, and Long Qin. 2019. Erroneous data generation for Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 149–158, Florence, Italy. Association for Computational Linguistics.

Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering. *arXiv preprint arXiv:1904.06652*.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. G-daug: Generative data augmentation for commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1008–1025.

Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dialog State Tracking with Reinforced Data Augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9474–9481.

Masashi Yokota and Hideki Nakayama. 2018. Augmenting Image Question Answering Dataset by Exploiting Image Captions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. 2019. Data Augmentation for Spoken Language Understanding via Joint Variational Generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7402–7409.

S. Young, M. Gašić, B. Thomson, and J. D. Williams. 2013. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE*, 101(5):1160–1179.

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*.

Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing. *arXiv preprint arXiv:2009.13845*.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *Proceedings of ICLR*.

Rongzhi Zhang, Yue Yu, and Chao Zhang. 2020. SeqMix: Augmenting Active Sequence Labeling via Sequence Mixup. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8566–8579, Online. Association for Computational Linguistics.

Rui Zhang, Tao Yu, Heyang Er, Sungrok Shim, Eric Xue, Xi Victoria Lin, Tianze Shi, Caiming Xiong, Richard Socher, and Dragomir Radev. 2019a. Editing-based SQL Query Generation for Cross-Domain Context-Dependent Questions. 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 5338–5349, Hong Kong, China. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-Level Convolutional Networks for Text Classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Yi Zhang, Tao Ge, Furu Wei, Ming Zhou, and Xu Sun. 2019b. Sequence-to-sequence Pre-training with Data Augmentation for Sentence Rewriting. *arXiv preprint arXiv:1909.06002*.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2019c. Task-oriented dialog systems that consider multiple appropriate responses under the same context.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019d. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Zijian Zhao, Su Zhu, and Kai Yu. 2019. Data augmentation with atomic templates for spoken language understanding. *arXiv preprint arXiv:1908.10770*.

Victor Zhong, Mike Lewis, Sida I. Wang, and Luke Zettlemoyer. 2020. Grounded adaptation for zero-shot executable semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6869–6882, Online. Association for Computational Linguistics.

Haichao Zhu, Li Dong, Furu Wei, Bing Qin, and Ting Liu. 2019. Transforming wikipedia into augmented data for query-focused summarization. *arXiv preprint arXiv:1911.03324*.

Ran Zmigrod, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

Appendices

A Useful Blog Posts and Code Repositories

The following blog posts and code repositories could be helpful in addition to the information presented and papers/works mentioned in the body:

- Introduction to popular text augmentation techniques: https://towardsdatascience.com/data-augmentation-in-nlp-2801a34dfc28
- Detailed blog post on various text DA techniques: https://amitness.com/2020/05/data-augmentation-for-nlp/
- Lightweight library for DA on text and audio: https://github.com/makcedward/nlpaug
- `python` framework for adversarial examples: https://github.com/QData/TextAttack

B DA Methods Table - Description of Columns and Attributes

Table 1 in the main body compares a non-exhaustive selection of DA methods along various aspects relating to their applicability, dependencies, and requirements. Below, we provide a more extensive description of each of this table's columns and their attributes.

1. **Ext.Know**: Short for external knowledge, this column is ✓ when the data augmentation process requires knowledge resources which go beyond the immediate input examples and the task definition, such as WordNet (Miller, 1995) or PPDB (Pavlick et al., 2015). Note that we exclude the case where these resources are pretrained models under a separate point (next) for clarity, since these are widespread enough to merit a separate category.

2. **Pretrained**: Denotes that the data augmentation process requires a pretrained model, such as BERT (Devlin et al., 2019) or GPT-2 (Radford et al., 2019).

3. **Preprocess**: Denotes the preprocessing steps, e.g. tokenization (*tok*), dependency parsing (*dep*), etc. required for the DA process. A hyphen (-) means either no preprocessing is required or that it was not explicitly stated.

4. **Level**: Denotes the depth and extent to which elements of the instance/data are modified by the DA. Some primitives modify just the INPUT (e.g. word swapping), some modify both INPUT and LABEL (e.g. negation), while others make changes in the embedding or hidden space (EMBED/HIDDEN) or higher representation layers enroute to the task model.

5. **Task-Agnostic**: This is an approximate, partially subjective column denoting the extent to which a DA method can be applied to different tasks. When we say ✓ here, we don't denote a very rigid sense of the term task-agnostic, but mean that it would possibly easily extend to most NLP tasks as understood by the authors. Similarly, an × denotes being restricted to a specific task (or small group of related tasks) only. There can be other labels, denoting applicability to broad task families. For example, SUBSTRUCTURAL denotes the family of tasks where sub-parts of the input are also valid input examples in their own right, e.g. constituency parsing. SENTENCE PAIRS denotes tasks which involve pairwise sentence scoring such as paraphrase identification, duplicate question detection, and semantic textual similarity.

C Additional DA Works by Task

See Table 2 for additional DA works for GEC, Table 3 for additional DA works for neural machine translation, Table 4 for additional DA works for dialogue, and Table 5 for additional DA works for multimodal tasks. Each work is described briefly.

D Additional Figure



Figure 4: Pedro Domingos' quip about offline data augmentation.

Paper/Work	Brief Description
Lichtarge et al. (2019)	Generate synthetic noised examples of Wikipedia sentences using backtranslation through various languages.
White and Rozovskaya (2020)	Detailed comparative study of the DA for GEC systems UEdin-MS (Grundkiewicz et al., 2019) and Kakao&Brain (Choe et al., 2019).
Foster and Andersen (2009)	Introduces error generation tool called GenERRate which learns to generate ungrammatical text with various errors by using an error analysis file.
Kimn (2020)	Use a set of syntactic rules for common Japanese grammatical errors to generate augmented error-correct sentence pairs for Japanese GEC.
Felice (2016)	Thesis that surveys previous work on error generation and investigates some new approaches using random and probabilistic methods.
Xu et al. (2019)	Noises using five error types: concatenation, misspelling, substitution, deletion, and transposition. Decent performance on the BEA 2019 Shared Task.
Zhang et al. (2019b)	Explore backtranslation and feature discrimination for DA.
Mizumoto et al. (2011)	DA by extracting Japanese GEC training data from the revision log of a language learning SNS.

Table 2: Additional DA works for grammatical error correction (GEC), along with a brief description of each.

Paper/Work	Brief Description
Vaibhav et al. (2019)	Present a *synthetic noise induction* model which heuristically adds social media noise to text, and *labeled backtranslation*.
Hassan et al. (2017)	Present a DA method to project words from closely-related high-resource languages to low-resource languages using word embedding representations.
Cheng et al. (2020)	Propose *AdvAug*, an adversarial augmentation method for NMT, by sampling adversarial examples from a new vicinity distribution and using their embeddings to augment training.
Graça et al. (2019)	Investigate improvements to sampling-based approaches and the synthetic data generated by backtranslation.
Bulte and Tezcan (2019)	Propose DA approaches for NMT that leverage information retrieved from a Translation Memory (TM) and using fuzzy TM matches.
Moussallem et al. (2019)	Propose an NMT model KG-NMT which is augmented by knowledge graphs to enhance semantic feature extraction and hence the translation of entities and terminological expressions.
Peng et al. (2020)	Propose dictionary-based DA (DDA) for cross-domain NMT by synthesizing a domain-specific dictionary and automatically generating a pseudo in-domain parallel corpus.
Li et al. (2020a)	Present a DA method using sentence boundary segmentation to improve the robustness of NMT on ASR transcripts.
Nishimura et al. (2018)	Introduce DA methods for multi-source NMT that fills in incomplete portions of multi-source training data.
Sugiyama and Yoshinaga (2019)	Investigate effectiveness of DA by backtranslation for context-aware NMT.
Li and Specia (2019)	Present DA methods to improve NMT robustness to noise while keeping models small, and explore the use of noise from external data (speech transcripts).
Chinea-Ríos et al. (2017)	Propose DA method to create synthetic data by leveraging the embedding representation of sentences.
Alves et al. (2020)	Propose two methods for pipeline-based speech translation through the introduction of errors through 1. utilizing a speech processing workflow and 2. a rule-based method.
Kang (2019)	Investigate extremely low-resource settings for NMT and a DA approach using a noisy dictionary and language models.
Chen et al. (2020a)	Investigate a DA method for lexically constraint-aware NMT to construct constraint-aware synthetic training data.
Li et al. (2020b)	Propose a diversity DA method for low-resource NMT by generating diverse synthetic parallel data on both source and target sides using a restricted sampling strategy during decoding.
Duan et al. (2020)	Propose syntax-aware DA methods with sentence-specific word selection probabilities using dependency parsing.

Table 3: Additional DA works for neural machine translation (NMT), along with a brief description of each.

Paper/Work	Brief Description
Gao et al. (2020)	Propose a paraphrase augmented response generation (PARG) framework to improve dialogue generation by automatically constructing augmented paraphrased training examples based on dialogue state and act labels.
Gritta et al. (2021)	Introduce a graph-based representation of dialogues called Conversation Graph (ConvGraph) that can be used for DA by creating new dialogue paths.
Yin et al. (2020)	Propose an RL-based DA approach for dialogue state tracking (DST).
Song et al. (2020)	Propose a simple DA algorithm to improve the training of copy-mechanism models for dialogue state tracking (DST).

Table 4: Additional DA works for dialogue, along with a brief description of each.

Paper/Work	Brief Description
Huang et al. (2018)	Propose a DA method for emotion recognition from a combination of audio, visual, and textual modalities.
Mou et al. (2020)	Introduce a DA method for Audio-Video Scene-Aware Dialogue, which involves dialogue containing a sequence of QA pairs about a video.
Falcon et al. (2020)	Investigate DA techniques for video QA including mirroring and horizontal flipping.

Table 5: Additional DA works for multimodal tasks, along with a brief description of each.