AI Open 3 (2022) 71–90

Contents lists available at ScienceDirect

AI Open

journal homepage: www.keaipublishing.com/en/journals/ai-open

Data augmentation approaches in natural language processing: A survey

Bohan Li, Yutai Hou, Wanxiang Che *

Harbin Institute of Technology, Harbin, China

ARTICLE INFO

Keywords:
Machine learning
Data augmentation
Natural language processing

ABSTRACT

As an effective strategy, data augmentation (DA) alleviates data scarcity scenarios where deep learning techniques may fail. It is widely applied in computer vision then introduced to natural language processing and achieves improvements in many tasks. One of the main focuses of the DA methods is to improve the diversity of training data, thereby helping the model to better generalize to unseen testing data. In this survey, we frame DA methods into three categories based on the **diversity** of augmented data, including paraphrasing, noising, and sampling. Our paper sets out to analyze DA methods in detail according to the above categories. Further, we also introduce their applications in NLP tasks as well as the challenges. Some useful resources are provided in Appendix A.

1. Introduction

Data augmentation refers to methods used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. Such methods alleviate data scarcity scenarios where deep learning techniques may fail, so DA has received active interest and demand recently. Data augmentation is widely applied in the field of computer vision (Shorten and Khoshgoftaar, 2019), such as flipping and rotation, then introduced to natural language processing (NLP). Different to images, natural language is discrete, which makes the adoption of DA methods more difficult and underexplored in NLP.

Large numbers of DA methods have been proposed recently, and a survey of existing methods is beneficial so that researchers could keep up with the speed of innovation. Liu et al. (2020b) and Feng et al. (2021) both present surveys that give a bird's eye view of DA for NLP. They directly divide the categories according to the methods. These categories thus tend to be too limited or general, e.g., *back-translation* and *model-based techniques*. Bayer et al. (2021) post a survey on DA for text classification only. In this survey, we will provide an inclusive overview of DA methods in NLP. One of our main goals is to show the nature of DA, i.e., *why data augmentation works*. To facilitate this, we category DA methods according to the **diversity** of augmented data, since improving training data diversity is one of the main thrusts of DA effectiveness. We frame DA methods into three categories, including paraphrasing, noising, and sampling.[1,2]

Specifically, *paraphrasing*-based methods generate the paraphrases of the original data as the augmented data. This category brings limited changes compared with the original data. *Noising*-based methods add more continuous or discrete noises to the original data and involve more changes. *Sampling*-based methods master the distribution of the original data to sample new data as augmented data. With the help of artificial heuristics and trained models, such methods can sample brand new data rather than changing existing data and therefore generate even more diverse data.

Our paper sets out to analyze DA methods in detail according to the above categories. In addition, we also introduce their applications in NLP tasks as well as the challenges. The rest of the paper is structured as follows:

- Section 2 presents a comprehensive review of the three categories and analyzes every single method in those categories. We also introduce the characteristics of the methods, e.g., the granularity and the level.
- Section 3 refers to a summary of common strategies and tricks to improve the quality of augmented data, including method stacking, optimization, and filtering strategies.
- Section 4 analyzes the application of the above methods in NLP tasks. We also show the development of DA methods through a timeline.
- Section 5 introduces some related topics of data augmentation, including pre-trained language models, contrastive learning, similar

* Corresponding author.
E-mail addresses: bhli@ir.hit.edu.cn (B. Li), ythou@ir.hit.edu.cn (Y. Hou), car@ir.hit.edu.cn (W. Che).

[1] This paper is available at: https://arxiv.org/abs/2110.01852 as well as https://www.sciencedirect.com/science/article/pii/S2666651022000080.

[2] We will provide further resources at: https://github.com/BohanLi0110/NLP-DA-Papers.

https://doi.org/10.1016/j.aiopen.2022.03.001
Received 27 September 2021; Accepted 19 March 2022
Available online 24 March 2022

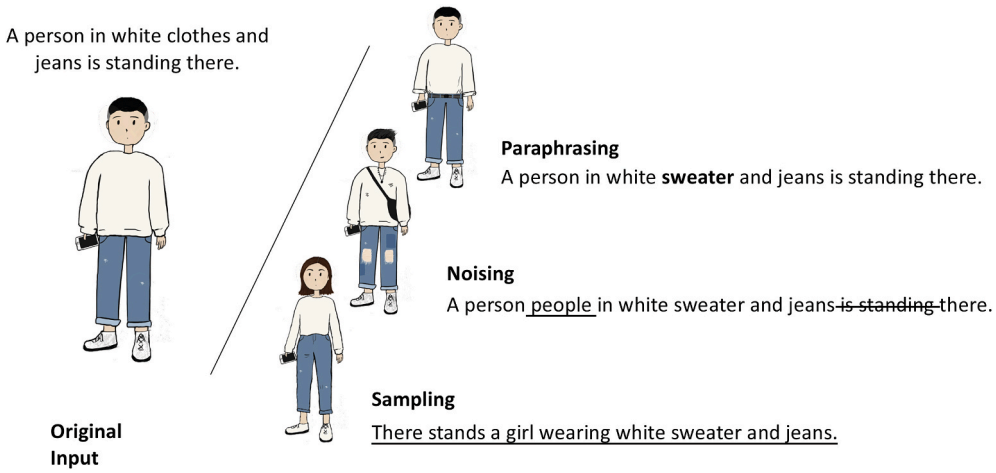

Fig. 1. Data augmentation techniques include three categories. The examples of the original data and augmented data are on the left and right, respectively. As we can see, the **diversity** of *paraphrasing*, *noising*, and *sampling* increases in turn compared to the original input.

data manipulation methods, generative adversarial networks, and adversarial attacks. We aim to connect data augmentation with other topics and show their difference at the same time.
- Section 6 lists some challenges we observe in NLP data augmentation, including theoretical narrative and generalized methods. These points also reveal the future development direction of data augmentation.
- Section 7 concludes the paper.

2. Data augmentation methods in NLP

Data Augmentation aims at generating additional, synthetic training data in insufficient data scenes. Data augmentation ranges from simple techniques like rule-based methods to learnable generation-based methods, and all the above methods essentially guarantee the validity of the augmented data (Raille et al., 2020). That is to say, DA methods need to make sure that the augmented data is valid for the task, i.e., be considered to be part of the same distribution of the original data (Raille et al., 2020). For example, similar semantics in machine translation and the same label in text classification as the original data.

On the basis of validity, augmented data is also expected to be diverse to improve model generalization on downstream tasks. This involves the **diversity** of augmented data. In this survey, we novelly divide DA methods into three categories according to the diversity of their augmented data: paraphrasing, noising, and sampling.

- The paraphrasing-based methods generate augmented data that has limited semantic difference from the original data, based on proper and restrained changes to sentences. The augmented data convey very similar information as the original form.
- The noising-based methods add discrete or continuous noise under the premise of guaranteeing validity. The point of such methods is to improve the robustness of the model.
- The sampling-based methods master the data distributions and sample novel data within them. Such methods output more diverse data and satisfy more needs of downstream tasks based on artificial heuristics and trained models.

As shown in the examples and diagrams in Fig. 1, the paraphrasing, noising, and sampling-based methods provide more diversity in turn. The specific classification is shown in Fig. 2, and we will further introduce them in this section.

2.1. Paraphrasing-based methods

As common phenomena in natural language, paraphrases are alternative ways to convey the same information as the original form Barzilay and McKeown, (2001) and Madnani and Dorr, (2010). Naturally, the generation of paraphrases is a suitable scheme for data augmentation. Paraphrasing consists of several levels, including lexical paraphrasing, phrase paraphrase, and sentence paraphrase (Fig. 3). Therefore, the paraphrasing-based DA techniques introduced below can also be included into these three levels.

2.1.1. Thesauruses

Some works replace words in the original text with their synonyms and hypernyms,[3] so as to obtain a new way of expression while keeping the semantics of the original text as unchanged as possible. As shown in Fig. 4, thesauruses like WordNet (Miller, 1995) contain such lexical triplets of words and are often used as external resources.

Zhang et al. (2015) are the first to apply thesaurus in data augmentation. They use a thesaurus derived from WordNet,[4] which sorts the synonyms of words according to their similarity. For each sentence, they retrieve all replaceable words and randomly choose r of them to be replaced. The probability of number r is determined by a geometric distribution with parameter p in which $P[r] \sim p^r$. Given a word, the index s of its chosen synonym is also determined by a another geometric distribution in which $P[s] \sim p^s$. The method ensures that synonyms that are more similar to the original word are selected with greater probability. Some methods (Mueller and Thyagarajan, 2016; Daval-Frerot and Weis, 2020; Dai et al., 2020) apply a similar method.

A widely used text augmentation method called EDA (**E**asy **D**ata **A**ugmentation Techniques) (Wei et al., 2019) also replaces the original words with their synonyms using WordNet: they randomly choose n words, which are not stop words, from the original sentence.[5] Each of these words is replaced with a random synonym. Zhang et al. (2020a) apply a similar method in extreme multi-label classification.

In addition to synonyms, Coulombe (2018a) propose to use hypernyms to replace the original words. They also recommend the parts of speech of the augmented word in order of increasing difficulty: adverbs,

[3] Replacing a word with an antonym or a hyponym (more specific word) is usually not a semantically invariant transformation (Coulombe, 2018b).

[4] The thesaurus is obtained from the Mytheas component used in LibreOffice project.

[5] n is proportional to the length of the sentence.

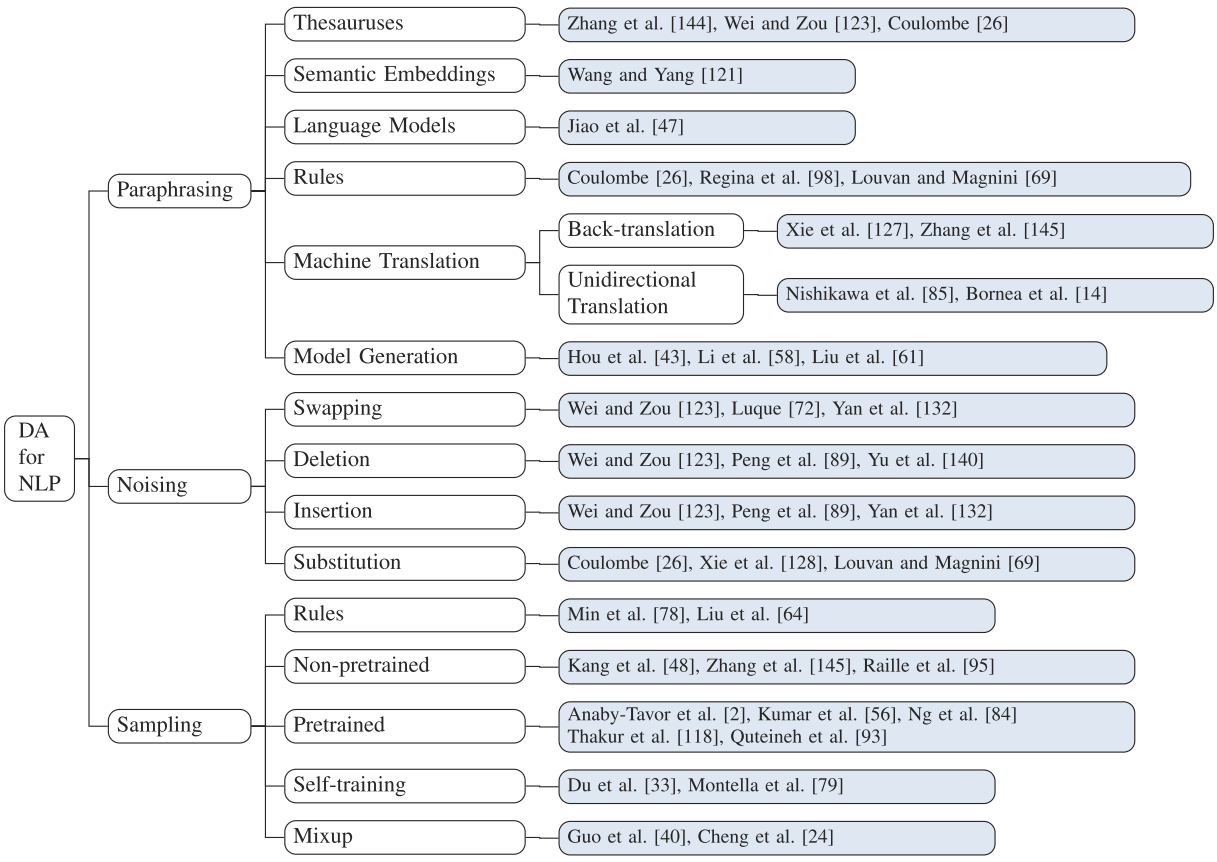

Fig. 2. Taxonomy of NLP DA methods.

adjectives, nouns, and verbs. Zuo et al. (2020) use WordNet and VerbNet (Schuler, 2005) to retrieve synonyms, hypernyms, and words of the same category.

Thesauruses
Advantage(s):
1. Easy to use.
Limitation(s):
1. The scope and part of speech of augmented words are limited.
2. This method cannot resolve the ambiguity problem.
3. Sentence semantics may be affected if there are too many substitutions.

2.1.2. Semantic embeddings

This method overcomes the limitations of replacement range and parts of speech in the thesaurus-based method. It uses pre-trained word embeddings, such as Glove, Word2Vec, FastText, etc., and replaces the original word in the sentence with its closest neighbor in embedding space, as shown in Fig. 5.[6]

In the Twitter message classification task, Wang and Yang (2015) pioneer to use both word embeddings and frame embeddings instead of discrete words.[7] As for word embeddings, each original word in the tweet is replaced with one of its k-nearest-neighbor words using cosine similarity. For example, "Being late is terrible" becomes "Being behind are bad". As for frame semantic embeddings, the authors semantically parse 3.8 million tweets and build a continuous bag-of-frame model to represent each semantic frame using Word2Vec (Mikolov et al., 2013). The same data augmentation approach as words is then applied to semantic frames.

Compared to Wang and Yang (2015), Liu et al. (2020e) only use word embeddings to retrieve synonyms. In the meanwhile, they edit the retrieving result with a thesaurus for balance. Ramirez-Echavarria et al. (2020) create the dictionary of embeddings for selection.

2.1.3. Language models

Pretrained language models have become mainstream models in recent years due to their excellent performance. Masked language models (MLMs) such as BERT and RoBERTa can predict masked words in text based on context, which can be used for text data augmentation (as shown in Fig. 6). Moreover, this approach alleviates the ambiguity problem since MLMs consider the whole context.

Semantic Embeddings
Advantage(s):
1. Easy to use.
2. Higher replacement hit rate and more comprehensive replacement range.
Limitation(s):
1. This method cannot resolve the ambiguity problem.[7]
2. Sentence semantics may be affected if there are too many substitutions.

Wu et al. (2019) fine-tune on pre-trained BERT to perform conditional MLM task. They alter the segmentation embeddings to label embeddings, which are learned corresponding to the annotated labels on labeled datasets. They use this fine-tuned conditional BERT to augment sentences. Specifically, a few words in a labeled sentence are randomly mask then filled by the conditional BERT.

Jiao et al. (2020) use both word embeddings and masked language models to obtain augmented data. They apply the tokenizer of BERT to

[6] Static word embeddings such as Wrod2Vec have only one representation for each word.

[7] The frame embeddings refer to the continuous embeddings of semantic frames (Baker et al., 1998).

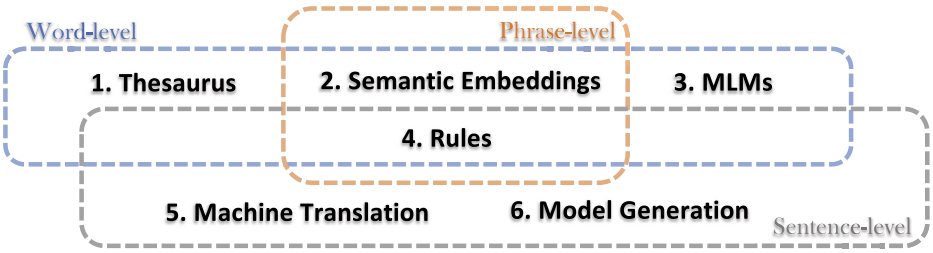

Fig. 3. Data augmentation techniques by paraphrasing include three levels: word-level, phrase-level, and sentence-level.

tokenize words into multiple word pieces. Each word piece is replaced with probability 0.4. If a word piece is not a complete word ("est" for example), it is replaced by its K-nearest-neighbor words in the Glove embedding space. If the word piece is a complete word, the authors replace it with [MASK] and employ BERT to predict *K* Words to fill in the blank. Regina et al. (2020); Tapia-Téllez and Escalante (2020); Lowell et al. (2021), and Palomino and Luna (2020) apply methods that are similar to Jiao et al. (2020). They mask multiple words in a sentence and generate new sentences by filling these masks to generate more varied sentences. In addition, RNNs are also used for replacing the original word based on the context (Kobayashi 2018; Fadaee et al., 2017).

Language Models
Advantage(s):
1. This approach alleviates the ambiguity problem.
2. This method considers context semantics.
Limitation(s):
1. Still limited to the word level.
2. Sentence semantics may be affected if there are too many substitutions.

2.1.4. Rules

This method requires some heuristics about natural language to ensure the maintaining of sentence semantics, as shown in Fig. 7.

On the one hand, some works rely on existing dictionaries or fixed heuristics to generate word-level and phrase-level paraphrases. Coulombe (2018a) introduce the use of regular expressions to transform the form without changing sentence semantics, such as the abbreviations and prototypes of verbs, modal verbs, and negation. For example, replace "is not" with "isn't". Similarly, Regina et al. (2020) use word-pair dictionaries to perform replacements between the expanded form and the abbreviated form.

On the other hand, some works generate sentence-level paraphrases

Fig. 4. Paraphrasing by using thesauruses.

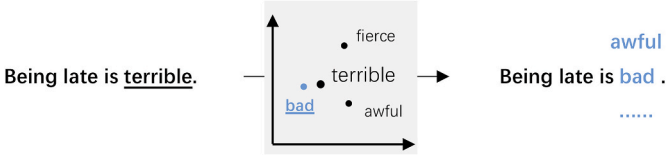

Fig. 5. Paraphrasing by using semantic embeddings.

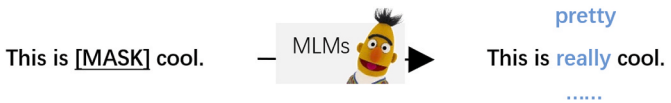

Fig. 6. Paraphrasing by using language models.

Fig. 7. Paraphrasing by using rules.

for original sentences with some rules, e.g. dependency trees. Coulombe (2018a) use a syntactic parser to build a dependency tree for the original sentence. Then the dependency tree is used for syntax transformation. For example, replace "Sally embraced Peter excitedly." with "Peter was embraced excitedly by Sally.". Dehouck and Gómez-Rodríguez (2020) apply a similar method. Louvan and Magnini (2020) crop particular fragments on the dependency tree to create a smaller sentence. They also rotate the target fragment around the root of the dependency parse structure, without harming the original semantics.

Rules
Advantage(s):
1. Easy to use.
2. This method preserves the original sentence semantics.
Limitation(s):
1. This method requires artificial heuristics.
2. Low coverage and limited variation.

2.1.5. Machine translation

Translation is a natural means of paraphrasing. With the development of machine translation models and the availability of online APIs, machine translation is popular as an augmentation method in many tasks, as shown in Fig. 8.

2.1.5.1. Back-translation. This method means that the original text is translated into other languages, and then translated back to obtain the augmented text in the original language. Different from word-level methods, back-translation does not directly replace individual words but rewrites the whole sentence in a generated way.

Xie et al. (2020); Yu et al. (2018), and Fabbri et al. (2021) use English-French translation models (in both directions) to perform back-translation on each sentence and obtain their paraphrases. Lowell et al. (2021) also introduce this method as one of the unsupervised data augmentation methods. Zhang et al. (2020c) leverage back-translation to obtain the formal expression of the original data in the style transfer task.

In addition to some trained machine translation models, some cloud translation API services like Google and DeepL are common tools for back-translation and are applied by some works like Coulombe (2018a); Luque (2019); Ibrahim et al. (2020); Daval-Frerot and Weis (2020); Longpre et al. (2020); Rastogi et al. (2020); Regina et al. (2020);

Fig. 8. Paraphrasing by machine translation.

Perevalov and Both (2020); Aroyehun and Gelbukh (2018).[8]

Some works add additional features based on vanilla back-translation. Nugent et al. (2021) propose a range of softmax temperature settings to ensure diversity while preserving semantic meaning. Qu et al. (2021) combine back-translation with adversarial training, to synthesize diverse and informative augmented examples by organically integrating multiple transformations. Zhang et al. (2020c) employ a discriminator to filter the sentences in the back-translation results. This method greatly improves the quality of the augmented data as a threshold.

2.1.5.2. Unidirectional translation. Different from back-translation, the unidirectional translation method directly translates the original text into other languages once, without translating it back to the original language. This method usually occurs in a multilingual scene.

In the task of unsupervised cross-lingual word embeddings (CLWEs), Nishikawa et al. (2020) build pseudo-parallel corpus with an unsupervised machine translation model. The authors first train unsupervised machine translation (UMT) models using the source/target training corpora and then translate the corpora using the UMT models. The machine-translated corpus is used together with the original corpus to learn monolingual word embeddings for each language independently. Finally, the learned monolingual word embeddings are mapped to a shared CLWE space. This method both facilitates the structural similarity of two monolingual embedding spaces and improves the quality of CLWEs in the unsupervised mapping method.

Bornea et al. (2021); Barrière and Balahur (2020), and Perevalov and Both (2020) translate the original English corpus into several other languages and obtain multiplied data. Correspondingly, they use multilingual models.

Machine Translation
Advantage(s):
1. Easy to use.
2. Wide range of applications.
3. This approach guarantees correct syntax and unchanged semantics.

Limitation(s):
1. Poor controllability and limited diversity because of the fixed machine translation models.

2.1.6. Model generation

Some methods employ Seq2Seq models to generate paraphrases directly. Such models output more diverse sentences given proper training objects, as shown in Fig. 9.

Hou et al. (2018) propose a Seq2Seq data augmentation model for the language understanding module of task-based dialogue systems. They feed the delexicalized input utterance and the specified diverse rank k (e.g. 1, 2, 3) into the Seq2Seq model as the input to generate a new utterance. Similarly, Hou et al. (2021) encodes the concatenated multiple input utterances by an L-layer transformer. The proposed model uses duplication-aware attention and diverse-oriented regularization to generate more diverse sentences.

In the task of aspect term extraction, Li et al. (2020) adopt Transformer as the basic structure. The masked original sentences as well as their label sequences are used to train a model M that reconstructs the masked fragment as the augmented data.[9] Kober et al. (2021) use GAN to generate samples that are very similar to the original data. Liu et al. (2020a) employ a pre-trained model to provide prior information to the proposed Transformer-based model. Then the proposed model could generate both context-relevant answerable questions and unanswerable questions.

Fig. 9. Paraphrasing by model generation.

Model Generation
Advantage(s):
1. Wide range of applications.
2. Strong application.

Limitation(s):
1. Require for training data.
2. High training difficulty.

2.2. Noising-based methods

The focus of paraphrasing is to make the semantics of the augmented data as similar to the original data as possible. In contrast, the noising-based methods add faint noise that does not seriously affect the semantics, so as to make it appropriately deviate from the original data. Humans greatly reduce the impact of weak noise on semantic understanding through their grasp of linguistic phenomena and prior knowledge, but this noise can pose challenges for models. Thus, this method not only expands the amount of training data but also improves model robustness (see Fig. 10).

2.2.1. Swapping

The semantics of natural language are sensitive to text order, while slight order change is still readable for humans (Wang et al., 1999). Therefore, the random swapping between words even sentences within a reasonable range can be used as a data augmentation method.

Wei and Zou (2019) randomly choose two words in the sentence and swap their positions. This process is repeated n times, in which n is proportional to the sentence length l. Longpre et al. (2020); Rastogi et al. (2020), and Zhang et al. (2020a) also apply the same method. Dai and Adel (2020) split the token sequence into segments according to labels, then randomly choose some segments to shuffle the order of the tokens inside, with the label order unchanged.

In addition to word-level swapping, some works also propose sentence-level even instance-level swapping. In the task of tweet sentiment analysis, Luque (2019) divide tweets into two halves. They randomly sample and combine first halves with second halves that have the same label. Although the data generated in this way may be ungrammatical and semantically unsound, it still carries relatively complete semantics and emotional polarity compared to individual words. Yan et al. (2019) perform sentence-level random swapping on legal documents classification. Since sentences independently contain relatively complete semantics comparing to words, the sentence order in the legal document has little effect on the meaning of the original text. Consequently, the authors shuffle the sentences to obtain the augmented text.

2.2.2. Deletion

This method means randomly deleting words in a sentence or deleting sentences in a document.

As for word-level deletion, Wei and Zou (2019) randomly remove each word in the sentence with probability p. Longpre et al. (2020); Rastogi et al. (2020), and Zhang et al. (2020a) also apply the same method. In the task of spoken language understanding, Peng et al.

[8] The above Cloud Translation API services and their links are: Google (https://cloud.google.com/translate/docs/apis) and DeepL (https://www.deepl.com/translator).

[9] Half of the words in original sentences whose sequence labels are not 'O' are masked.

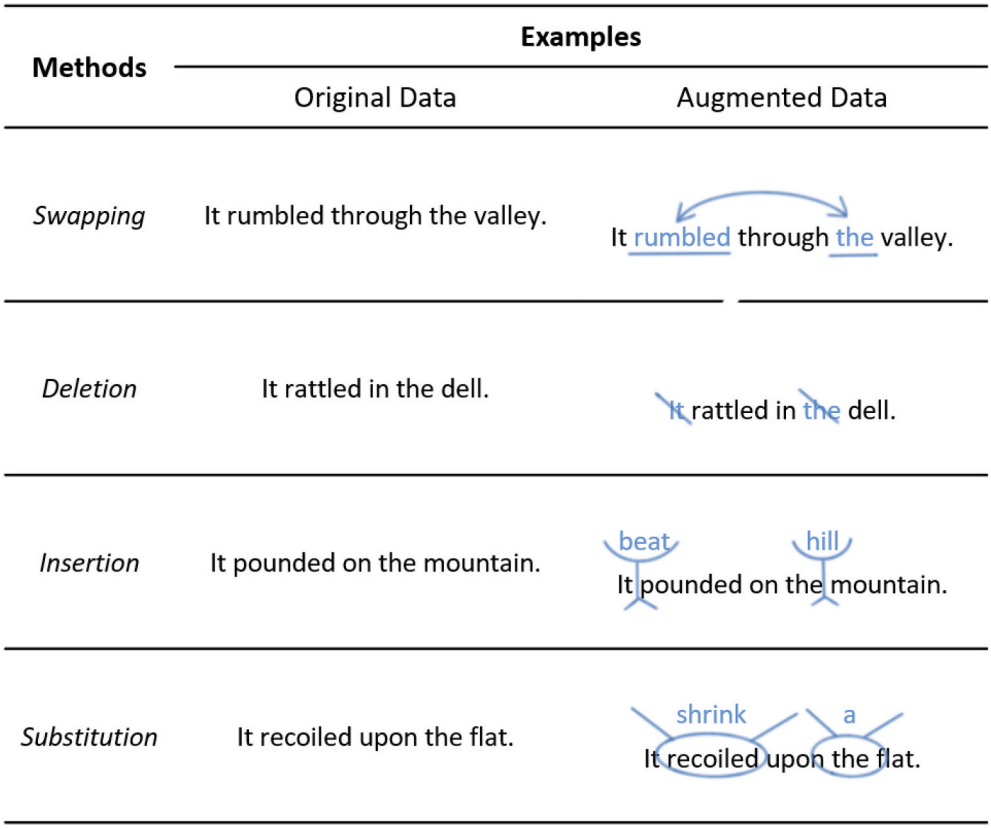

Fig. 10. The example of five noising-based methods.

(2020) augment input dialogue acts by deleting slot values to obtain more combinations.

As for sentence-level deletion, Yan et al. (2019) randomly delete each sentence in a legal document according to a certain probability. They do this because there exist many irrelevant statements and deleting them will not affect the understanding of the legal case. Yu et al. (2019) employ the attention mechanism to determine the objective of both word-level and sentence-level random deletion.

2.2.3. Insertion

This method means randomly inserting words into a sentence or inserting sentences into a document.

As for word-level insertion, Wei and Zou (2019) select a random synonym of a random word in a sentence that is not a stop word, then insert that synonym into a random position in the sentence. This process is repeated *n* times. In the task of spoken language understanding, Peng et al. (2020) augment input dialogue acts by inserting slot values to obtain more combinations.

In legal documents classification, since documents with the same label may have similar sentences, Yan et al. (2019) employ sentence-level random insertion. They randomly select sentences from other legal documents with the same label to get augmented data.

> Random insertion introduces new noisy information that may change the original label. Tips to avoid this problem:
> 1. Word level: use label-independent external resources.
> 2. Sentence level: use other samples with the same labels as the original data.

2.2.4. Substitution

This method means randomly replacing words or sentences with other strings. Different from the above paraphrasing methods, this method usually avoids using strings that are semantically similar to the original data.

Some works implement substitution through existing outer resources. Coulombe (2018a) and Regina et al. (2020) introduce a list of the most common misspellings in English to generate augmented texts containing common misspellings.[10] For example, “across” is easily misspelled as “across”. Xie et al. (2017) borrow from the idea of “word-dropout” and improve generalization by reducing the information in the sentence. This work uses “_” as a placeholder to replace random words, indicating that the information at that position is empty.

Some works use task-related resources or generate random strings for substitution. Xie et al. (2020) and Xie et al. (2017) replace the original words with other words in the vocabulary, and they use the TF-IDF value and the unigram frequency to choose words from the vocabulary, respectively. Lowell et al. (2021) and Daval-Frerot and Weis (2020) also explore this method as one of unsupervised data augmentation methods. Wang et al. (2018) propose a method that randomly replaces words in the input and target sentences with other words in the vocabulary. In NER, Dai and Adel (2020) replace the original token with a random token in the training set with the same label. Qin et al. (2020) propose a multi-lingual code-switching method that replaces original words in the source language with words of other languages. In the task of task-oriented dialogue, random substitution is a useful way to generate augmented data. Peng et al. (2020) augment input dialogue acts by replacing slot values to obtain more combinations in spoken language understanding. In slot filling, Louvan and Magnini (2020) do slot substitution according to the slot label. Song et al. (2021) augment the training data for dialogue state tracking by copying user utterances and replace the corresponding real slot values with generated random

[10] A list of common spelling errors in English can be obtained from the online resources of Oxford Dictionaries: https://en.oxforddictionaries.com/spelling/common-misspellings.

strings.

Random substitution introduces new noisy information that may change the original label. Tips to avoid this problem:
1. Word level: use label-independent external resources.
2. Sentence level: use other samples with the same labels as the original data.

Noising
Advantage(s):
1. Noising-based methods improve model robustness.
Disadvantage(s):
1. Poor interpretability.
2. Limited diversity for every single method.

2.3. Sampling-based methods

Sampling-based methods grasp the data distribution and sample new data within it. Similar to paraphrasing-based models, they also involve rules and trained models to generate augmented data. The difference is that the sampling-based methods are task-specific and require task information like labels and data format.[11] Such methods not only ensure validity but also increase diversity. They satisfy more needs of downstream tasks based on artificial heuristics and trained models, and can be designed according to specific task requirements. Thus, they are usually more flexible and difficult than the former two categories.

2.3.1. Rules

This method uses some rules to directly generate new augmented data. Heuristics about natural language and the corresponding labels are sometimes required to ensure the validity of the augmented data. The model structure is as shown in Fig. 11(a). Different from the above rule-based paraphrasing method, this method constructs valid but not guaranteed to be similar to the original data (even different labels).

Min et al. (2020) swap the subject and object of the original sentence, and convert predicate verbs into passive form. For example, inverse "This small collection contains 16 El Grecos." into "16 El Grecos contain this small collection.". The labels of new samples are determined by rules. Liu et al. (2020c) apply data augmentation methods in the task of solving math word problems (MWPs). They filter out some irrelevant numbers. Then some rules are used to construct new data based on the idea of double-checking, e.g., constructing augmented data describing $distance = time \times speed$ by reusing the original data describing $time = distance/speed$. The output equations of this method are computationally right. Given the training set of Audio-Video Scene-Aware Dialogue that provides 10 question-answer pairs for each video, Mou et al. (2020) shuffle the first n pairs as dialogue history and take the $n+1$-th question as what needs to be answered. In natural language inference, Kang et al. (2018) apply external resources like PPDB and artificial heuristics to construct new sentences. Then they combine the new sentences with original sentences as augmented pairs according to rules, for example, *if A entails B and B entails C, then A entails C*. Kober et al. (2021) define some rules to construct positive and negative pairs using adjective-noun (AN) and noun-noun (NN) compounds. For example, given $< car, car >$, they construct $< fastcar, car >$ as a positive sample and $< fastcar, redcar >$ as a negative sample. Shakeel et al. (2020) construct both paraphrase annotations and non-paraphrase annotations through three properties including reflexivity, symmetry, and transitive extension. Yin et al. (2020) use two kinds of rules including symmetric consistency and transitive consistency, as well as logic-guided DA methods to generate DA samples.

Rules
Advantage(s):
1. Easy to use.
Limitation(s):
1. Require for artificial heuristics.
2. Low coverage and limited variation.

2.3.2. Non-pretrained models

Some methods use non-pretrained models to generate augmented data. Such methods usually entail the idea of **back translation (BT)** (Sennrich et al., 2016a),[12] which is to train a target-to-source Seq2Seq model and use the model to generate source sentences from target sentences, i.e., constructing pseudo-parallel sentences (Zhang et al., 2020c). Such Seq2Seq model learns the internal mapping between the distributions of the target and the source, as shown in Fig. 11(b). This is different from the model generation based paraphrasing method because the augmented data of the paraphrasing method shares similar semantics with the original data.

Sennrich et al. (2016b) train an English-to-Chinese NMT model using existing parallel corpus, and use the target English monolingual corpus to generate Chinese corpus through the above English-to-Chinese model. Kang et al. (2018) train a Seq2Seq model for each label (*entailment*, *contradiction*, and *neutral*) and then generate new data using the Seq2Seq model given a sentence and a specific label. Chen et al. (2020d) adopt the Tranformer architecture and map the "rewrite utterance → request utterance" to the machine translation process. Moreover, they enforce the optimization process of the Seq2Seq generation with a policy gradient technique for controllable rewarding. Zhang et al. (2020c) use Transformer as the encoder and transfer the knowledge from Grammatical Error Correction to Formality Style Transfer. Raille et al. (2020) create the Edit-transformer, a Transformer-based model works cross-domain. Yoo et al. (2019) propose a novel VAE model to output the semantic slot sequence and the intent label given an utterance.

Non-pretrained Models
Advantage(s):
1. Strong diversity.
2. Strong application.
Limitation(s):
1. Require training data.
2. High training difficulty.

2.3.3. Pretrained models

In recent years, large-scale language models have achieved great success by acquiring rich linguistic knowledge through pretraining. Thus, they are naturally used as augmentation tools, as shown in Fig. 11 (c).

Anaby-Tavor et al. (2020) propose a data augmentation method named LAMBDA. They generate labeled augmented sentences with GPT-2, which is fine-tuned on the training set in advance. Then the augmented sentences are filtered by a classifier to ensure the data quality. Kumar et al. (2020) applies a similar method without the classifier for filtering.

Some works adopt masked language models to obtain augmented data. Ng et al. (2020) use the masked language model to construct a corruption model and a reconstruction model. Given the input data points, they initially generate data far away from the original data manifold with the corruption model. Then the reconstruction model is used to pull the data point back to the original data manifold as the final augmented data.

Some works adopt auto-regressive models to obtain augmented data.

[11] Recall that paraphrasing-based methods are task-independent and only require the original sentence as input.

[12] Note that the idea of back translation here is DIFFERENT from the above paraphrasing method called "back-translation" in Section 2.1.5.

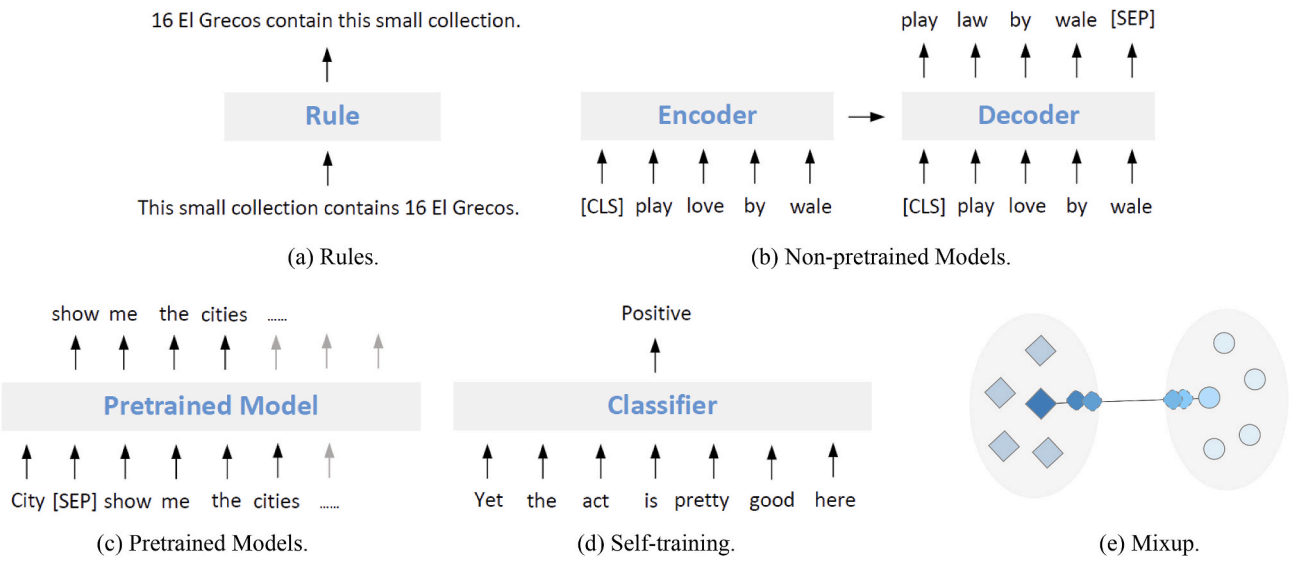

Fig. 11. Sampling-based models.

Peng et al. (2020) use the pre-trained SC-GPT and SC-GPT-NLU to generate utterances and dialogue acts respectively. The results are filtered to ensure the data quality. Abonizio and Junior (2020) fine-tune DistilBERT (Sanh et al., 2019) on original sentences to generate synthetic sentences. Especially, GPT-2 is a popular model used for generating augmented data. Quteineh et al. (2020) use label-conditioned GPT-2 to generate augmented data. Tarján et al. (2020) generate augmented data with GPT-2 and retokenize them into statistically derived subwords to avoid the vocabulary explosion in a morphologically rich language. Zhang et al. (2020a) use GPT-2 to generate substantially diversified augmented data in extreme multi-label classification.

Pretrained Models
Advantage(s):
1. Strong diversity.
2. Strong application.
Limitation(s):
1. Require training data.

2.3.4. Self-training

In some scenarios, unlabeled raw data is easy to obtain. Thus, converting such data into valid data would greatly increase the amount of data, as shown in Fig. 11(d).

Thakur et al. (2021) first fine-tune BERT on the original data, then use the fine-tuned BERT to label unlabeled sentence pairs. Such augmented data, as well as the gold data, are used to train SBERT together. Miao et al. (2020) further introduce data distillation into the self-training process. They output the label of unlabeled data by the iteratively updated teacher model. Yang et al. (2021) apply a similar self-training method in question answering; a cross-attention-based teacher model is used to determine the label of each QA pair. Du et al. (2021) introduce SentAugment, a data augmentation method that computes task-specific query embeddings from labeled data to retrieve sentences from a bank of billions of unlabeled sentences crawled from the web.

Some methods directly transfer exsiting models from other tasks to generate pseudo-parallel corpus. Montella et al. (2020) make use of Wikipedia to leverage a massive sentences. Then they use Stanford OpenIE package to extract the triplets given Wikipedia sentences. For example, given "*Barack Obama was born in Hawaii.*", the returned triples by Stanford OpenIE are $< BarackObama; was; born >$ and $< BarackObama; wasbornin; Hawaii >$ Such mappings are flipped as the augmented data of RDF-to-text tasks. Perevalov and Both (2020) apply a similar method. Since BERT does well on object-property (OP) relationship prediction and object-affordance (OA) relationship prediction, Zhao et al. (2020) directly use a fine-tuned BERT to predict the label of OP and OA samples.

Self-training
Advantage(s):
1. Easier than generative models.
2. Suitable for data-sparse scenarios.
Disadvantage(s):
1. Require for unlabeled data.

2.3.5. Mixup

This method uses virtual embeddings instead of generated natural language form text as augmented samples. The existing data is used as the basis to sample in the virtual vector space, and the sampled data may have different labels than the original data.

The idea of Mixup first appears in the image field by Zhang et al. (2018). Inspired by this work, Guo et al. (2019) propose two variants of Mixup for sentence classification. The first one called wordMixup conducts sample interpolation in the word embedding space, and the second one called senMixup interpolates the hidden states of sentence encoders. The interpolated new sample through wordMixup as well as senMixup, and their common interpolated label are obtained as follows:

$$\widetilde{B}_t^{ij} = \lambda B_t^i + (1-\lambda) B_t^j, \tag{1}$$

$$\widetilde{B}_{\{k\}}^{ij} = \lambda f(B^i)_{\{k\}} + (1-\lambda) f(B^j)_{\{k\}}, \tag{2}$$

$$\widetilde{y}^{ij} = \lambda y^i + (1-\lambda) y^j, \tag{3}$$

in which $B_t^i, B_t^j \in R^{N\times d}$ denote the t-th word in two original sentences, and $f(B^i)$, $f(B^j)$ denote the hidden layer sentence representation. Moreover, y^i, y^j are the corresponding original labels.

Mixup is widely applied in many works recently. Given the original samples, Cheng et al. (2020) firstly construct their adversarial samples following Cheng et al. (2019), and then apply two Mixup strategies named P_{adv} and P_{aut}: the former interpolates between adversarial samples, and the latter interpolates between the two corresponding original samples. Similarly, Sun et al. (2020); Bari et al. (2020), and Si et al.

(2020) all apply such Mixup method for text classification. Sun et al. (2020) propose Mixup-Transformer which combines Mixup with transformer-based pre-trained architecture. They test its performance on text classification datasets. Chen et al. (2020b) introduce Mixup into NER, proposing both Intra-LADA and InterLADA.

1. Mixup introduces continuous noise instead of discrete noise, it could generate augmented data between different labels.
2. This method is less interpretable and more difficult than the above noising-based methods.

2.4. Analysis

As shown in Table 1, we compare the above DA methods by various aspects.

- It is easy to find that nearly all paraphrasing-based and noising-based methods are not learnable, except for *Seq2Seq* and *Mixup*. However, most sampling-based methods are learnable except for the *rule*-based ones. Learnable methods are usually more complex than non-learnable ones, thus sampling-based methods generate more diverse and fluent data than the former two.
- Among all learnable methods, *Mixup* is the only **online** one. That is to say, the DA process is during model training. Thus, *Mixup* is the only one that outputs cross-label and discrete embedding from augmented data.
- Comparing *Learnable* and *Resource*, we could see that most non-learnable methods require external knowledge resources which go beyond the original dataset and task definition. Commonly used resources include semantic thesauruses like WordNet and PPDB, handmade resources like misspelling dictionary in Coulombe (2018a), and artificial heuristics like the ones in Min et al. (2020) and Kang et al. (2018).
- Through *Learnable*, *Ext.Know* and *Pretrain*, it can be seen that in addition to artificial heuristics, DA requires other external interventions to generate valid new data. This includes model training objectives, external knowledge resources, and knowledge implicit in pretrained language models.
- Comparing *Learnable* and *Task-related*, we could see that all paraphrasing-based and noising-based methods except model generation are not task-related. They generate augmented data given only original data without labels or task definition. However, all sampling-based methods are task-related because heuristics and model training are adopted to satisfy the needs of specific tasks.
- Comparing *Level* and *Task-related*, we could see that they are relevant. The paraphrasing-based methods are at the text level. The same is true for noising-based methods, except for Mixup, which augments both embeddings and labels. All sampling-based methods are at the text and label level since the labels are also considered and constructed during augmentation.
- Comparing *Learnable* and *Granularity*, we could see that almost all non-learnable methods could be used for word-level and phrase-level DA, but all learnable methods could only be applied for sentence-level DA. Although learnable methods generate high-quality augmented sentences, unfortunately, they do not work for document augmentation because of their weaker processing ability for documents. Thus, document augmentation still relies on simple non-learnable methods, which is also a current situation we have observed in our research.

3. Strategies and tricks

The three types of DA methods including paraphrasing, noising, and sampling, as well as their characteristics, have been introduced above. In practical applications, the effect of the DA method is influenced by many factors. In this chapter, we present these factors to inspire our readers to use some strategies and tricks for selecting and constructing suitable DA methods.

3.1. Method stacking

The methods in Section 2 are not mandatory to be applied alone. They could be combined for better performance. Common combinations include:

3.1.1. The same type of methods

Some works combine different paraphrasing-based methods and obtain different paraphrases, to increase the richness of augmented data. For example, Liu et al. (2020e) use both thesauruses and semantic embeddings, and Jiao et al. (2020) use both semantic embeddings and MLMs. As for noising-based methods, the former unlearnable ways are usually used together like Peng et al. (2020). It is because these methods are simple, effective, and complementary. Some methods also adopt different sources of noising or paraphrasing like Regina et al. (2020) and Xie et al. (2017). The combination of different resources could also improve the robustness of the model.

3.1.2. Unsupervised methods

In some scenarios, the simple and task-independent unsupervised DA methods could meet the demand. Naturally, they are grouped together and widely used. Wei and Zou (2019) introduce a DA toolkit called EDA that consists of synonym replacement, random insertion, random swap, and random deletion. EDA is very popular and used for many tasks like Longpre et al. (2020) and Rastogi et al. (2020). UDA by Xie et al. (2020) includes back-translation and unsupervised noising-based methods; it is also used in many tasks like Daval-Frerot and Weis (2020).

3.1.3. Multi-granularity

Some works apply the same method at different levels to enrich the augmented data with changes of different granularities and improve the robustness of the model. For example, Wang and Yang (2015) train both word embeddings and frame embeddings by Word2Vec; Guo et al. (2019) apply Mixup at the word and sentence level, and Yu et al. (2019) use a series of noising-based methods at both the word and the sentence level.

3.2. Optimization

The optimization process of DA methods directly influences the quality of augmented data. We introduce it through four angles: the use of augmented data, hyperparameters, training strategies, and training objects.

3.2.1. The use of augmented data

The way of using augmented data directly influences the final effect. From the perspective of data quality, the augmented data could be used to pre-train a model if it is not of high quality; otherwise, it could be used to train a model directly. From the perspective of data amount, if the amount of the augmented data is much higher than the original data, they are usually not directly used together for model training. Instead, some common practices include (1) oversampling the original data before training the model (2) pre-training the model with the augmented data and fine-tuning it on the original data.

3.2.2. Hyperparameters

All the above methods involve hyperparameters that largely affect the augmentation effect. We list some common hyperparameters in Fig. 12:

3.2.3. Training strategies

Some works apply training strategies based on the basic data augmentation methods. For example, Qu et al. (2021) combine

Table 1
Characteristics of different DA methods. *Learnable* denotes whether the methods involve model training; *online* and *offline* denote whether the DA process is during or after model training. *Ext.Know* denotes to whether the methods require external knowledge resources to generate augmented data. *Pretrain* denotes whether the methods require a pre-trained model. *Task-related* denotes whether the methods consider the label information, task format, and task requirements to generate augmented data. *Level* denotes the depth and extent to which elements of the instance/data are modified by the DA; *t*, *e*, and *l* denote text, embedding, and label, respectively. *Granularity* indicates the extent to which the method could augment; *w*, *p*, and *s* denote word, phrase, and sentence, respectively.

		Learnable	Ext.Know	Pretrain	Task-related	Level	Granularity
Paraphrasing	Thesauruses	–	✓	–	–	*t*	*w*
	Semantic Embeddings	–	✓	–	–	*t*	*w*, *p*
	Language Models	–	–	✓	–	*t*	*w*
	Rules	–	✓	–	–	*t*	*w*, *p*, *s*
	Machine Translation	–	–	–	–	*t*	*s*
	Model Generation	offline	–	–	✓	*t*	*s*
Noising	Swapping	–	–	–	–	*t*	*w*, *p*, *s*
	Deletion	–	–	–	–	*t*	*w*, *p*, *s*
	Insertion	–	✓	–	–	*t*	*w*, *p*, *s*
	Substitution	–	✓	–	–	*t*	*w*, *p*, *s*
Sampling	Rules	–	✓	–	✓	*t*, *l*	*w*, *p*, *s*
	Non-pretrained	offline	–	–	✓	*t*, *l*	*s*
	Pretrained	offline	–	✓	✓	*t*, *l*	*s*
	Self-training	offline	–	–	✓	*t*, *l*	*s*
	Mixup	**online**	–	–	✓	*e*, *l*	*s*

back-translation with adversarial training. Similarly, Quteineh et al. (2020) transform the basic pre-trained model into an optimization problem[13] to maximize the usefulness of the generated output. Hu et al. (2019) and Liu et al. (2020d) use pre-trained language models to generate augmented data, and transfer such progress into reinforcement learning. Some works (Rastogi et al., 2020; Shehnepoor et al., 2020) take the idea of Generative Adversarial Networks to generate challenging augmented data.

3.2.4. Training objects

Training objects are essential for model training, especially for the learnable DA methods. Nugent et al. (2021) propose a range of softmax temperature settings to ensure diversity while preserving semantic meaning. Hou et al. (2021) use duplication-aware attention and diverse-oriented regularization to generate more diverse sentences. Cheng et al. (2020) employ curriculum learning to encourage the model to focus on the difficult training examples.

3.3. Filtering

Sometimes the progress of data augmentation inevitably introduces some noise even errors, thus filtering mechanisms are introduced to avoid this problem.

Some works filter input data in the initial stage to avoid inappropriate input affecting the augmentation effect. A typical example is sentence length, i.e., filter sentences that are too short (Li et al., 2020). Liu et al. (2020c) filter out irrelevant numbers without augmenting them in solving Math Word Problems, to ensure the generated data is computationally right.

In addition, some works filter the synthetic augmented data at the end-stage. This is usually achieved through a model. For example, Zhang et al. (2020c) employ a discriminator to filter the back-translation results. Anaby-Tavor et al. (2020) and Peng et al. (2020) both apply a classifier to filter the augmented sentences generated by pre-trained models to ensure the data quality.

4. Applications on NLP tasks

Although many data augmentation methods have emerged in recent years, it is difficult to directly compare their performances. This is because different tasks, evaluation metrics, datasets, model architectures, and experimental settings make direct comparisons meaningless. Therefore, based on the work introduced above, we analyze the data augmentation methods from the perspective of different NLP tasks including text classification, text generation, and structured prediction (Che et al., 2021).

- Text classification is the simplest and most basic natural language processing problem. That is, for a piece of text input, output the category to which the text belongs, where the category is a predefined closed set.[14]
- Text generation, as the name implies, is to generate the corresponding text given the input data. The most classic example is machine translation.
- The structured prediction problem is usually unique to NLP. Different from the text classification, there are strong correlation and format requirements between the output categories in the structured prediction problem.

In this section, we try to analyze the features as well as the development status of DA in these tasks. Some statistical results are shown in Tables 2 and 3.

DA methods are applied more widely in text classification than other NLP tasks in general and in each category. Moreover, each individual DA method could be applied to text classification. Such application advantage is because of the simple form of text classification: given the input text, it directly investigates the model's understanding of semantics by label prediction. Therefore, it is relatively simple for data augmentation to only consider retaining the semantics of words that are important for classification.

As for text generation, it prefers sampling-based methods to bring more semantic diversity. And structured prediction prefers paraphrasing-based methods because it is sensitive to data format. Thus, it has higher requirements for data validity.

By comparing each DA method, we can see that simple and effective unsupervised methods like machine translation, thesaurus-based paraphrasing, and random substitution, are quite popular. In addition, learnable methods like paraphrasing-based model generation and sampling-based pretrained models, also gain a lot of attention because of their diversity and effectiveness.

We also show the development process of the DA method on three

[13] Monte Carlo Tree Search.

[14] Text matching tasks such as Natural Language Inference can also be transformed into text classification.

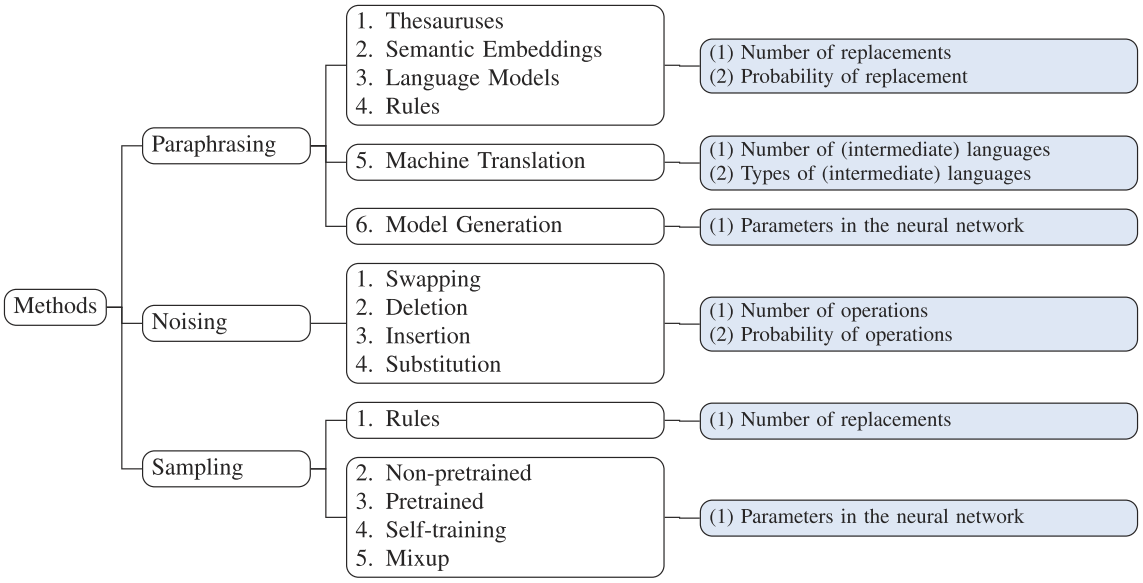

Fig. 12. Hyperparameters that affect the augmentation effect in each DA method.

types of tasks through a timeline (Table 3). On the whole, the number of applications of DA in these tasks has increased these years. Text classification is the first task to use DA, and the number of corresponding papers is also larger than the other two tasks. In terms of text generation and structured prediction, DA is receiving increasing attention. Paraphrasing-based methods have always been a popular method. In recent years, sampling-based methods show clear momentum in text classification and text generation, because they bring more gains to powerful pretrained language models than paraphrasing-based methods. However, people still tend to use paraphrasing and noising-based methods in structured prediction.

5. Related topics

How does data augmentation relate to other learning methods? In this section, we connect data augmentation with other similar topics.

5.1. Pretrained language models

The training of most pre-trained language models (PLMs) is based on self-supervised learning. Self-supervised learning mainly uses auxiliary tasks to mine its supervised information from large-scale unsupervised data, and trains the network through this constructed supervised information, so that it can learn valuable representations for downstream tasks. From this perspective, PLMs also introduce more training data into downstream tasks, in an implicit way. On the other hand, the general large-scale unsupervised data of PLMs may be out-of-domain for specific tasks. Differently, the task-related data augmentation methods essentially focus on specific tasks.

5.2. Contrastive learning

Contrastive learning is to learn an embedding space in which similar samples are close to each other while dissimilar ones are far apart. It focuses on learning the common features between similar samples and distinguishing the differences between dissimilar ones. The first step of contrastive learning is applying data augmentation to construct similar samples with the same label, and the second step is to randomly choose instances as the negative samples. Thus, contrastive learning is one of the applications of data augmentation.

5.3. Other data manipulation methods

In addition to DA, there are some other data manipulation methods to improve model generalization (Kukacka et al., 2017; Hu et al., 2019). *Oversampling* is usually used in data imbalance scenarios. It simply samples original data from the minority group as new samples, instead of generating augmented data. *Data cleaning* is additionally applied to the original data to improve data quality and reduce data noise. It usually includes lowercasing, stemming, lemmatization, etc. *Data weighting* assigns different weights to different samples according to their importance during training, without generating new data. *Data synthesis* provides entire labeled artificial examples instead of augmented data generated by models or rules.

5.4. Generative adversarial networks

Generative Adversarial Networks (GANs) are first introduced by Goodfellow et al. (2014). As a type of semi-supervised method, GANs include the generative model, which is mainly used to challenge the discriminator of GANs, while the generative models in some DA methods are directly used to augment training data. Moreover, the generative model of GANS is applied as a DA method in some scenes like Rastogi et al. (2020); Morris et al. (2020); Shehnepoor et al. (2020); Kober et al. (2021); Zhou et al. (2020); Cao and Lee (2020), and have demonstrated to be effective for data augmentation purposes.

5.5. Adversarial attacks

Adversarial attacks are techniques to generate adversarial examples attacking a machine learning model, i.e., causing the model to make a mistake. Some works use DA methods like code-switch substitution to generate adversarial examples as consistency regularization (Zheng et al., 2021).

6. Challenges and opportunities

Data augmentation has seen a great process over the last few years, and it has provided a great contribution to large-scale model training as well as the development of downstream tasks. Despite the process, there are still challenges to be addressed. In this section, we discuss some of these challenges and future directions that could help advance the field.

Table 2
The application of DA methods in NLP tasks. Note that if a paper involves multiple methods, we count it multiple times.

		Text	Text	Structure
		Classification	Generation	Prediction
Paraphrasing	Thesauruses	Zhang et al. (2015); Wei and Zou (2019); Liu et al. (2020e); Coulombe (2018a); Daval-Frerot and Weis (2020); Longpre et al. (2020); Zhang et al. (2020a); Zuo et al. (2020); Liu and Yu (2020); Kovatchev et al. (2021)	–	Daval-Frerot and Weis (2020); Dai and Adel (2020)
	Embeddings	Wang and Yang (2015); Liu et al. (2020e); Kovatchev et al. (2021)	–	–
	Language Models	Regina et al. (2020); Tapia-Téllez and Escalante (2020); Kobayashi (2018); Wu et al. (2019); Bari et al. (2021)	Fadaee et al. (2017)	–
	Rules	Regina et al. (2020); Coulombe (2018a); Louvan and Magnini (2020)	–	Sahin and Steedman (2019); Andreas (2020)
	Machine Translation	Daval-Frerot and Weis (2020); Longpre et al. (2020); Regina et al. (2020); Xie et al. (2020); Ibrahim et al. (2020); Rastogi et al. (2020); Aroyehun and Gelbukh (2018); Coulombe (2018a); Luque (2019); Barrière and Balahur (2020); Lun et al. (2020); Liu and Yu (2020); Risch and Krestel (2018)	Zhang et al. (2020c); Fabbri et al. (2021)	Daval-Frerot and Weis (2020); Yu et al. (2018); Bornea et al. (2021); Longpre et al. (2019)
	Model Generation	Liu et al. (2020a); Kober et al. (2021); Xu et al. (2020); Guo et al. (2020); Zhao et al. (2019)	Liu et al. (2020a); Wan et al. (2020); Kumar et al. (2019); Li et al. (2019)	Liu et al. (2020a); Hou et al. (2018); Hou et al. (2021); Li et al. (2020); Yoo et al. (2020); Yin et al. (2020)
Noising	Swapping	Wei and Zou (2019); Longpre et al. (2020); Zhang et al. (2020a); Rastogi et al. (2020); Yan et al. (2019); Luque (2019); Du and Black (2018); Kovatchev et al. (2021)	–	Dai and Adel (2020)
	Deletion	Wei and Zou (2019); Longpre et al. (2020); Zhang et al. (2020a); Rastogi et al. (2020); Yan et al. (2019); Yu et al. (2019); Chen et al. (2021a)	Peng et al. (2020)	–
	Insertion	Wei and Zou (2019); Longpre et al. (2020); Zhang et al. (2020a); Rastogi et al. (2020); Kovatchev et al. (2021)	Peng et al. (2020)	–
	Substitution	Daval-Frerot and Weis (2020); Regina et al. (2020); Xie et al. (2020); Coulombe (2018a); Lun et al. (2020)	Xie et al. (2017); Wang et al. (2018); Peng et al. (2020)	Daval-Frerot and Weis (2020); Louvan and Magnini (2020); Dai and Adel (2020); Shi et al. (2021)
Sampling	Rules	Min et al. (2020); Kang et al. (2018); Kober et al. (2021); Shakeel et al. (2020); Lun et al. (2020); Xu et al. (2016); Chen et al. (2021b); Jiang et al. (2021)	Mou et al. (2020); Zhang et al. (2020b); Asai and Hajishirzi (2020); Bergmanis and Goldwater (2019)	Zmigrod et al. (2019)
	Non-pretrained	Kang et al. (2018); Raille et al. (2020); Yoo et al. (2019); Zhou et al. (2020); Niu and Bansal (2019)	Zhang et al. (2020c); Chen et al. (2020d); Yao et al. (2020); Chen et al. (2020a); Sennrich et al. (2016b)	Yoo et al. (2019); Liu et al. (2021)
	Pretrained	Zhang et al. (2020a); Ng et al. (2020); Kumar et al. (2020); Quteineh et al. (2020); Liu et al. (2020d); Anaby-Tavor et al. (2020); Abonizio and Junior (2020); Staliunaite et al. (2021); Dong et al. (2021)	Peng et al. (2020); Ng et al. (2020); Tarján et al. (2020)	Peng et al. (2020); Riabi et al. (2021)
	Self-training	Du et al. (2021); Yang et al. (2021); Perevalov and Both (2020); Miao et al. (2020)	Montella et al. (2020); Xu et al. (2021)	Yang et al. (2021)
	Mixup	Guo et al. (2019); Sun et al. (2020); Si et al. (2020; 2021)	Cheng et al. (2020)	Chen et al. (2020b)

6.1. *Theoretical narrative*

At this stage, there appears to be a lack of systematic probing work and theoretical analysis of DA methods in NLP. The few related works are of DA in the image domain, considering data augmentation as encoding a priori knowledge about data or task invariance (Dao et al., 2019), variance reduction (Chen et al., 2020c) or regularization methods (Wu et al., 2020). In NLP, Most previous works propose new methods or prove the effectiveness of the DA method on downstream tasks, but do not explore the reasons and laws behind it, e.g., from the perspective of mathematics. The discrete nature of natural language makes theoretical narrative essential since narrative helps us understand the nature of DA, without being limited to determining effectiveness through experiments.

6.2. *More exploration on pretrained language models*

In recent years, pre-trained language models have been widely applied in NLP, which contain rich knowledge through self-supervision on a huge scale of corpora. There are works using pre-trained language models for DA, but most of them are limited to [MASK] completion (Wu et al., 2019), direct generation after fine-tuning (Zhang et al., 2020a), or self-training (Du et al., 2021). Is DA still helpful in the era of pre-trained language models? Or, how to further use the information in pre-trained models to generate more diverse and high-quality data with less cost? There are some initial explorations in these directions (Zhou et al., 2022; Liu et al., 2022), while we still look forward to more works in the future.

6.3. *Few-shot scenarios*

In few-shot scenarios, models are required to achieve performance which rivals that of traditional machine learning models, yet the amount of training data is extremely limited. DA methods provide a direct solution to the problem. However, most current works in few-shot scenarios are paraphrasing-based methods (Fabbri et al., 2021). Such methods ensure the validity of the augmented data, but also lead to insufficient semantic diversity. Mainstream pretrained language models obtain rich semantic knowledge by language modeling. Such knowledge even covers to some extent the semantic information introduced by traditional paraphrasing-based DA methods. In other words, the improvement space that traditional DA methods bring to pretrained language models has been greatly compressed. Therefore, it is an interesting question how to provide models with fast generalization and problem solving capability by generating high quality augmented data in few-shot scenarios.

Table 3
Timeline of DA methods applied in three kinds of NLP tasks. The time for each paper is based on its first arXiv version (if exists) or estimated submission time. denotes paraphrasing-based methods; N denotes noising-based methods; S denotes sampling-based methods.

	Text Classification	Text Generation	Structured Prediction
2015.09	Zhang et al. (2015) P Wang and Yang (2015) P		
2015.11		Sennrich et al. (2016b) S	
2016.01	Xu et al. (2016) S		
…			
2017.03		Xie et al. (2017) N	
2017.05		Fadaee et al. (2017) P	
…			
2018.04			Yu et al. (2018) P
2018.05	Kang et al. (2018) S		
2018.06	Kobayashi (2018) P		
2018.07			Hou et al. (2018) P
2018.08	Aroyehun and Gelbukh (2018) P Risch and Krestel (2018) P	Wang et al. (2018) N	
2018.09	Yoo et al. (2019) S		Yoo et al. (2019) S
2018.10	Du and Black (2018) N		Sahin and Steedman (2019) P
2018.12	Coulombe (2018a) P, N Wu et al. (2019) P		
2019.01	Wei and Zou (2019) P, N		
2019.04	Xie et al. (2020) P, N		
2019.05	Guo et al. (2019) S	Guo et al. (2019) N	
2019.06		Xia et al. (2019) S Bergmanis and Goldwater (2019) S Kumar et al. (2019) P	
2019.07	Yu et al. (2019) N	Li et al. (2019) P	Zmigrod et al. (2019) S
2019.08			Yin et al. (2020) P
2019.09	Luque (2019) P, N Yan et al. (2019) N		
2019.11	Anaby-Tavor et al. (2020) S Malandrakis et al. (2019) P Niu and Bansal (2019) S Zhao et al. (2019) P		Longpre et al. (2019) P
2019.12	Shakeel et al. (2020) S		

(*continued on next page*)

Table 3 (*continued*)

	Text Classification	Text Generation	Structured Prediction
2020.01			Yoo et al. (2020) P
2020.03	Kumar et al. (2020) S Raille et al. (2020) S		
2020.04	Lun et al. (2020) P, N, S	Peng et al. (2020) N, S	Li et al. (2020) P Peng et al. (2020) S
2020.05	Kober et al. (2021) P, S Cao and Lee (2020) S	Zhang et al. (2020c) P, S	
2020.06	Liu et al. (2020e) P Qin et al. (2020) N	Cheng et al. (2020) S	Qin et al. (2020) S
2020.07	Min et al. (2020) S Rastogi et al. (2020) P, N Regina et al. (2020) P, N Asai and Hajishirzi (2020) S	Chen et al. (2020a) S Tarján et al. (2020) S Mou et al. (2020) S	(Andreas, 2020) P
2020.09	Ng et al. (2020) S Zhang et al. (2020a) P, N, S	Ng et al. (2020) S Zhang et al. (2020b) S	Yang et al. (2021) S
2020.10	Barrière and Balahur (2020) P Louvan and Magnini (2020) P Tapia-Téllez and Escalante (2020) P Sun et al. (2020) S Abonizio and Junior (2020) S Zuo et al. (2020) P	Fabbri et al. (2021) P	Liu et al. (2020a) P Louvan and Magnini (2020) N Chen et al. (2020b) S Dai and Adel (2020) P, N Riabi et al. (2021) S
2020.11	Longpre et al. (2020) P, N Quteineh et al. (2020) S		
2020.12	Miao et al. (2020) S Daval-Frerot and Weis (2020) P,	Wan et al. (2020) P Yao et al. (2020)	Bornea et al. (2021) P Hou et al. (2021) P

(*continued on next page*)

Table 3 (*continued*)

	Text Classification	Text Generation	Structured Prediction
	P		
	Liu et al. (2020d)	Montella et al. (2020)	Daval-Frerot and Weis (2020)
	S	S	P,
			N
	Perevalov and Both (2020)	Chen et al. (2020d)	
	S	S	
	Si et al. (2020)		
	S		
	Xu et al. (2020)		
	P		
	Liu and Yu (2020)		
	P		
	Guo et al. (2020)		
	P		
	Si et al. (2021)		
	S		
2021.01	Shi et al. (2021)		Shi et al. (2021)
	N		N
	Staliunaite et al. (2021)		
	S		
	Dong et al. (2021)		
	S		
2021.06	Chen et al. (2021b)	Xu et al. (2021)	
	S	S	
	Chen et al. (2021a)		
	N		
	Jiang et al. (2021)		
	S		
	Kovatchev et al. (2021)		
	P,		
	N		
2021.08	Bari et al. (2021)		Liu et al. (2021)
	P		S

6.4. Retrieval augmentation

Retrieval-augmented language models integrate retrieval into pre-training and downstream usage.[15] Retrieval augmentation makes models much more parameter-efficient, as they need to store less knowledge in their parameters and can instead retrieve it (Singh et al., 2021; Yogatama et al., 2021). It also enables efficient domain adaptation by simply updating the data used for retrieval (Khandelwal et al., 2019). Recently, the size of the retrieval corpora has achieved explosive growth (Borgeaud et al., 2021) and models have been equipped with the ability to query the web for answering questions (Komeili et al., 2021; Nakano et al., 2021). In the future, there may be different forms of retrieval to leverage different kinds of information such as common sense knowledge, factual relations, linguistic information, etc. Retrieval augmentation could also be combined with more structured forms of knowledge retrieval, such as methods from knowledge base population and open information extraction.

6.5. More generalized methods for NLP

Natural language is most different from image or sound in that its representation is discrete. At the same time, NLP includes specific tasks such as structured prediction that are not available in other modalities. Therefore, unlike general methods such as *clipping* for image augmentation or *speed perturbation* for audio augmentation, there is currently no DA method that can be effective for all NLP tasks. This means that there is still a gap for DA methods between different NLP tasks. With the development of pre-trained models, this seems to have some possibilities. Especially the proposal of T5 (Raffel et al., 2020) and GPT3 (Brown et al., 2020), as well as the emergence of prompting learning further verify that the formalization of tasks in natural language can be independent of the traditional categories, and a more generalized model could be obtained by unifying task definitions.

6.6. Working with long texts and low resources languages

The existing methods have made significant progress in short texts and common languages. However, limited by model capabilities, DA methods on long texts still struggle with the simplest methods of paraphrasing and noising (Liu et al., 2020e; Yan et al., 2019; Yu et al., 2019) (as shown in Table 1). At the same time, limited by data resources, augmentation methods of low resource languages are scarce (Kumar et al., 2020), although they have more demand for data augmentation. Obviously, exploration in these two directions is still limited, and they could be promising directions.

7. Conclusion

In this paper, we presented a comprehensive and structured survey of data augmentation for natural language processing. In order to inspect the nature of DA, we framed DA methods into three categories according to the **diversity** of augmented data, including paraphrasing, noising, and sampling. Such categories help to understand and develop DA methods. We also introduced the characteristics of DA methods and their applications in NLP tasks, then analyzed them through a timeline. In addition, we introduced some tricks and strategies so that researchers and practitioners can refer to obtain better model performance. Finally,

[15] See https://ruder.io/ml-highlights-2021/for further information.

we distinguish DA with some related topics and outlined current challenges as well as opportunities for future research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Key R&D Program of China via grant 2020AAA0106501 and the National Natural Science Foundation of China (NSFC) via grant 61976072 and 62176078.

Appendix A. Related Resources

There are some **popular resources** that provides helpful information or APIs of DA.

- A Visual Survey of Data Augmentation in NLP (Blog)
- EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks (Repo)
- Unsupervised Data Augmentation (Repo)
- Unsupervised Data Augmentation (Pytorch) (Repo)
- nlpaug: Data Augmentation in NLP (Repo)
- TextAttack: Generating Adversarial Examples for NLP Models (Repo)
- AugLy: A Data Augmentations Library for Audio, Image, Text, and Video (Repo)
- NL-Augmenter: A Collaborative Repository of Natural Language Transformations (Repo)

In addition to English, there are resources in **other languages** such as:

- Turkish: nlpaug: Data Augmentation in NLP (Repo)
- Chinese: NLP Data Augmentation with EDA, BERT, BART, and back-translation (Repo)
- Chinese: ChineseNLPDataAugmentation4Paddle: NLP Data Augmentation with EDA and BERT Contextual Augmentation, Customized for PaddleNLP (Repo)
- Chinese: Tencent AI Lab Embedding Corpus for Chinese Words and Phrases (Link)

References

Abonizio, H.Q., Junior, S.B., 2020. Pre-trained data augmentation for text classification. October 20-23, 2020. In: Cerri, R., Prati, R.C. (Eds.), Intelligent Systems - 9th Brazilian Conference, BRACIS 2020. Proceedings, Part I, Springer, Rio Grande, Brazil, pp. 551–565. https://doi.org/10.1007/978-3-030-61377-8_38, 10.1007/978-3-030-61377-8_38.

Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., Zwerdling, N., 2020. Do not have enough data? deep learning to the rescue. February 7-12, 2020. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020. AAAI Press, New York, NY, USA, pp. 7383–7390. https://aaai.org/ojs/index.php/AAAI/article/view/6233.

Andreas, J., 2020. Good-enough compositional data augmentation. Online. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 7556–7566. https://aclanthology.org/2020.acl-main.676, 10.18653/v1/2020.acl-main.676.

Aroyehun, S.T., Gelbukh, A.F., 2018. Aggression detection in social media: using deep neural networks, data augmentation, and pseudo labeling. August 25, 2018. In: Kumar, R., Ojha, A.K., Zampieri, M., Malmasi, S. (Eds.), Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 90–97. https://aclanthology.org/W18-4411/.

Asai, A., Hajishirzi, H., 2020. Logic-guided data augmentation and regularization for consistent question answering. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. Association for Computational Linguistics, pp. 5642–5650. https://doi.org/10.18653/v1/2020.acl-main.499. Online, July 5-10, 2020, 10.18653/v1/2020.acl-main.499.

Baker, C.F., Fillmore, C.J., Lowe, J.B., 1998. The Berkeley FrameNet project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, vol 1. Association for Computational Linguistics, Montreal, Quebec, Canada, pp. 86–90. https://aclanthology.org/P98-1013, 10.3115/980845.980860.

Bari, M.S., Mohiuddin, M.T., Joty, S.R., 2020. Multimix: a robust data augmentation strategy for cross-lingual NLP. CoRR abs/2004.13240. https://arxiv.org/abs/2004.13240. arXiv:2004.13240.

Bari, M.S., Mohiuddin, T., Joty, S., 2021. UXLA: a robust unsupervised data augmentation framework for zero-resource cross-lingual NLP. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, pp. 1978–1992. Online. https://aclanthology.org/2021.acl-long.154. doi:10.18653/v1/2021.acl-long.154.

Barrière, V., Balahur, A., 2020. Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation. In: Scott, D., Bel, N., Zong, C. (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020. International Committee on Computational Linguistics, Barcelona, Spain, pp. 266–271. https://doi.org/10.18653/v1/2020.coling-main.23 (Online), December 8-13, 2020, 10.18653/v1/2020.coling-main.23.

Barzilay, R., McKeown, K.R., 2001. Extracting paraphrases from a parallel corpus. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Toulouse, France, pp. 50–57. https://aclanthology.org/P01-1008, 10.3115/1073012.1073020.

Bayer, M., Kaufhold, M., Reuter, C., 2021. A survey on data augmentation for text classification. CoRR abs/2107.03158. https://arxiv.org/abs/2107.03158. arXiv:2107.03158.

Bergmanis, T., Goldwater, S., 2019. Data augmentation for context-sensitive neural lemmatization using inflection tables and raw text (Long and Short Papers). In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4119–4128. https://aclanthology.org/N19-1418, 10.18653/v1/N19-1418.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Driessche, G. v.d., Lespiau, J.B., Damoc, B., Clark, A., et al., 2021. Improving Language Models by Retrieving from Trillions of Tokens arXiv preprint arXiv:2112.04426.

Bornea, M.A., Pan, L., Rosenthal, S., Florian, R., Sil, A., 2021. Multilingual transfer learning for QA using translation as data augmentation. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, the Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, pp. 12583–12591. https://ojs.aaai.org/index.php/AAAI/article/view/17491.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020. NeurIPS 2020. December 6-12, 2020, virtual. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Cao, R., Lee, R.K., 2020. Hategan: adversarial generative-based data augmentation for hate speech detection (Online), December 8-13, 2020. In: Scott, D., Bel, N., Zong, C. (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020. International Committee on Computational Linguistics, Barcelona, Spain, pp. 6327–6338. https://doi.org/10.18653/v1/2020.coling-main.557, 10.18653/v1/2020.coling-main.557.

Chen, G., Chen, Y., Wang, Y., Li, V.O.K., 2020a. Lexical-constraint-aware neural machine translation via data augmentation. In: Bessiere, C. (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, ijcai.org, pp. 3587–3593. https://doi.org/10.24963/ijcai.2020/496, 10.24963/ijcai.2020/496.

Chen, J., Shen, D., Chen, W., Yang, D., 2021a. HiddenCut: simple data augmentation for natural language understanding with better generalizability. Online. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ume 1. Long Papers), Association for Computational Linguistics, pp. 4380–4390. https://aclanthology.org/2021.acl-long.338, 10.18653/v1/2021.acl-long.338.

Chen, J., Wang, Z., Tian, R., Yang, Z., Yang, D., 2020b. Local additivity based data augmentation for semi-supervised NER. In: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, pp. 1241–1251. https://doi.org/10.18653/v1/2020.emnlp-main.95, 10.18653/v1/2020.emnlp-main.95.

Chen, S., Dobriban, E., Lee, J., 2020c. A group-theoretic framework for data augmentation. Adv. Neural Inf. Process. Syst. 33, 21321–21333.

Chen, Y., Kedzie, C., Nair, S., Galuscakova, P., Zhang, R., Oard, D., McKeown, K., 2021b. Cross-language sentence selection via data augmentation and rationale training. Online. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural

Language Processing, ume 1. Long Papers), Association for Computational Linguistics, pp. 3881–3895. https://aclanthology.org/2021.acl-long.300, 10.18653/v1/2021.acl-long.300.

Chen, Y., Lu, S., Yang, F., Huang, X., Fan, X., Guo, C., 2020d. Pattern-aware data augmentation for query rewriting in voice assistant systems. arXiv abs/2012.11468. https://arxiv.org/abs/2012.11468. arXiv:2012.11468.

Cheng, Y., Jiang, L., Macherey, W., 2019. Robust neural machine translation with doubly adversarial inputs. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 4324–4333. https://aclanthology.org/P19-1425, 10.18653/v1/P19-1425.

Cheng, Y., Jiang, L., Macherey, W., Eisenstein, J., 2020. Advaug: robust adversarial augmentation for neural machine translation. Online, July 5-10, 2020. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. Association for Computational Linguistics, pp. 5961–5970. https://doi.org/10.18653/v1/2020.acl-main.529, 10.18653/v1/2020.acl-main.529.

Che, Wanxiang, Guo, Jiang, Cui, Yiming, 2021. Natural Language Processing: Methods Based on Pre-trained Models. Electronic Industry Press.

Coulombe, C., 2018a. Text data augmentation made simple by leveraging NLP cloud apis. In: arXiv:1812.04718. ArXiv abs/1812.04718. http://arxiv.org/abs/1812.04718.

Coulombe, C., 2018b. Text data augmentation made simple by leveraging NLP cloud apis. In: arXiv:1812.04718 arXiv abs/1812.04718. http://arxiv.org/abs/1812.04718.

Xiang Dai and Heike Adel. 2020, An Analysis of Simple Data Augmentation for Named Entity Recognition, Donia Scott, Nuria Bel and Chengqing Zong, Proceedings of the 28th International Conference on Computational Linguistics, {COLING} 2020, Barcelona, Spain (Online), December 8-13, 2020, International Committee on Computational Linguistics, 3861–3867, https://doi.org/10.18653/v1/2020.coling-main.343.

Dao, T., Gu, A., Ratner, A., Smith, V., De Sa, C., Ré, C., 2019. A kernel theory of modern data augmentation. In: International Conference on Machine Learning. PMLR, pp. 1528–1537.

Daval-Frerot, G., Weis, Y., 2020. WMD at SemEval-2020 tasks 7 and 11: assessing humor and propaganda using unsupervised data augmentation (online). In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. International Committee for Computational Linguistics, Barcelona, pp. 1865–1874. https://www.aclweb.org/anthology/2020.semeval-1.246.

Dehouck, M., Gómez-Rodríguez, C., 2020. Data augmentation via subtree swapping for dependency parsing of low-resource languages (Online), December 8-13, 2020. In: Scott, D., Bel, N., Zong, C. (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020. International Committee on Computational Linguistics, Barcelona, Spain, pp. 3818–3830. https://doi.org/10.18653/v1/2020.coling-main.339, 10.18653/v1/2020.coling-main.339.

Dong, X., Zhu, Y., Fu, Z., Xu, D., de Melo, G., 2021. Data augmentation with adversarial training for cross-lingual NLI. Online. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, pp. 5158–5167. https://aclanthology.org/2021.acl-long.401, 10.18653/v1/2021.acl-long.401.

Du, J., Grave, E., Gunel, B., Chaudhary, V., Celebi, O., Auli, M., Stoyanov, V., Conneau, A., 2021. Self-training improves pre-training for natural language understanding. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tür, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, pp. 5408–5418. https://doi.org/10.18653/v1/2021.naacl-main.426, 10.18653/v1/2021.naacl-main.426.

Du, W., Black, A., 2018. Data augmentation for neural online chats response selection. In: Proceedings of the 2018 EMNLP Workshop SCAI: the 2nd International Workshop on Search-Oriented Conversational AI. Association for Computational Linguistics, Brussels, Belgium, pp. 52–58. https://aclanthology.org/W18-5708, 10.18653/v1/W18-5708.

Fabbri, A.R., Han, S., Li, H., Li, H., Ghazvininejad, M., Joty, S.R., Radev, D.R., Mehdad, Y., 2021. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tür, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, pp. 704–717. https://doi.org/10.18653/v1/2021.naacl-main.57, 10.18653/v1/2021.naacl-main.57.

Fadaee, M., Bisazza, A., Monz, C., 2017. Data augmentation for low-resource neural machine translation. July 30 - August 4. In: Barzilay, R., Kan, M. (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, vol. 2. Short Papers, Association for Computational Linguistics, Vancouver, Canada, pp. 567–573. https://doi.org/10.18653/v1/P17-2090, 10.18653/v1/P17-2090.

Feng, S.Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E., 2021. A survey of data augmentation approaches for NLP. Online. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, pp. 968–988. https://aclanthology.org/2021.findings-acl.84, 10.18653/v1/2021.findings-acl.84.

Gao, F., Zhu, J., Wu, L., Xia, Y., Qin, T., Cheng, X., Zhou, W., Liu, T.Y., 2019. Soft contextual data augmentation for neural machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 5539–5544. https://aclanthology.org/P19-1555, 10.18653/v1/P19-1555.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y., 2014. Generative adversarial networks. CoRR abs/1406.2661. http://arxiv.org/abs/1406.2661. arXiv:1406.2661.

Guo, H., Mao, Y., Zhang, R., 2019. Augmenting data with mixup for sentence classification: an empirical study. arXiv abs/1905.08941. http://arxiv.org/abs/1905.08941. arXiv:1905.08941.

Guo, Z., Liu, Z., Ling, Z., Wang, S., Jin, L., Li, Y., 2020. Text classification by contrastive learning and cross-lingual data augmentation for Alzheimer's disease detection. Barcelona, Spain (Online). In: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, pp. 6161–6171. https://aclanthology.org/2020.coling-main.542, 10.18653/v1/2020.coling-main.542.

Hou, Y., Chen, S., Che, W., Chen, C., Liu, T., 2021. C2c-genda: cluster-to-cluster generation for data augmentation of slot filling. February 2-9, 2021. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, the Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event. AAAI Press, pp. 13027–13035. https://ojs.aaai.org/index.php/AAAI/article/view/17540.

Hou, Y., Liu, Y., Che, W., Liu, T., 2018. Sequence-to-sequence data augmentation for dialogue language understanding. August 20-26, 2018. In: Bender, E.M., Derczynski, L., Isabelle, P. (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 1234–1245. https://aclanthology.org/C18-1105/.

Hu, Z., Tan, B., Salakhutdinov, R., Mitchell, T.M., Xing, E.P., 2019. Learning data manipulation for augmentation and weighting. Vancouver, BC, Canada. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, pp. 15738–15749. https://proceedings.neurips.cc/paper/2019/hash/671f0311e2754fcdd37f70a8550379bc-Abstract.html.

Ibrahim, M., Torki, M., El-Makky, N., 2020. AlexU-Back Translation-TL at SemEval-2020 task 12: improving offensive language detection using data augmentation and transfer learning (online). In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. International Committee for Computational Linguistics, Barcelona, pp. 1881–1890. https://aclanthology.org/2020.semeval-1.248, 10.18653/v1/2020.semeval-1.248.

Jiang, Z., Han, J., Sisman, B., Dong, X.L., 2021. CoRI: collective relation integration with data augmentation for open information extraction. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ume 1. Long Papers), Association for Computational Linguistic, pp. 4706–4716. https://aclanthology.org/2021.acl-long.363, 10.18653/v1/2021.acl-long.363.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q., 2020. Tinybert: distilling BERT for natural language understanding. Online Event, 16-20 November 2020. In: Cohn, T., He, Y., Liu, Y. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, pp. 4163–4174. https://doi.org/10.18653/v1/2020.findings-emnlp.372, 10.18653/v1/2020.findings-emnlp.372.

Kang, D., Khot, T., Sabharwal, A., Hovy, E.H., 2018. Adventure: adversarial training for textual entailment with knowledge-guided examples. July 15-20, 2018. In: Gurevych, I., Miyao, Y. (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, vol. 1. Long Papers, Association for Computational Linguistics, Melbourne, Australia, pp. 2418–2428. https://aclanthology.org/P18-1225/, 10.18653/v1/P18-1225.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., Lewis, M., 2019. Generalization through Memorization: Nearest Neighbor Language Models arXiv preprint arXiv:1911.00172.

Sosuke Kobayashi, Marilyn A. Walker, Heng Ji and Amanda Stent, Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), Association for Computational Linguistics, https://doi.org/10.18653/v1/n18-2072.

Kober, T., Weeds, J., Bertolini, L., Weir, D.J., 2021. Data augmentation for hypernymy detection. Online, April 19 - 23, 2021. In: Merlo, P., Tiedemann, J., Tsarfaty, R. (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021. Association for Computational Linguistics, pp. 1034–1048. https://aclanthology.org/2021.eacl-main.89/.

Komeili, M., Shuster, K., Weston, J., 2021. Internet-augmented Dialogue Generation arXiv preprint arXiv:2107.07566.

Kovatchev, V., Smith, P., Lee, M.G., Devine, R.T., 2021. Can vectors read minds better than experts? comparing data augmentation strategies for the automated scoring of children's mindreading ability. August 1-6, 2021. In: Zong, C., Xia, F., Li, W., Navigli, R. (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, Long Papers), Virtual Event, ume 1. Association for Computational Linguistics, pp. 1196–1206. https://doi.org/10.18653/v1/2021.acl-long.96, 10.18653/v1/2021.acl-long.96.

Kukacka, J., Golkov, V., Cremers, D., 2017. Regularization for deep learning: a taxonomy. CoRR abs/1710.10686. http://arxiv.org/abs/1710.10686. arXiv:1710.

Kumar, A., Bhattamishra, S., Bhandari, M., Talukdar, P., 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation (Long and Short Papers). In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ume 1. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 3609–3619. https://aclanthology.org/N19-1363, 10.18653/v1/N19-1363.

Kumar, V., Choudhary, A., Cho, E., 2020. Data augmentation using pre-trained transformer models. arXiv abs/2003.02245. https://arxiv.org/abs/2003.02245. arXiv:2003.02245.

Li, J., Qiu, L., Tang, B., Chen, D., Zhao, D., Yan, R., 2019. Insufficient Data Can Also Rock! Learning to Converse Using Smaller Data with Augmentation. AAAI.

Li, K., Chen, C., Quan, X., Ling, Q., Song, Y., 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. Online, July 5-10, 2020. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. Association for Computational Linguistics, pp. 7056–7066. https://doi.org/10.18653/v1/2020.acl-main.631, 10.18653/v1/2020.acl-main.631.

Liu, A., Swayamdipta, S., Smith, N.A., Choi, Y., 2022. Wanli: Worker and Ai Collaboration for Natural Language Inference Dataset Creation arXiv preprint arXiv:2201.05955.

Liu, C., Yu, D., 2020. BLCU-NLP at SemEval-2020 task 5: data augmentation for efficient counterfactual detecting (online). In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. International Committee for Computational Linguistics, Barcelona, pp. 633–639. https://aclanthology.org/2020.semeval-1.81.

Liu, D., Gong, Y., Fu, J., Yan, Y., Chen, J., Lv, J., Duan, N., Zhou, M., 2020a. Tell me how to ask again: question data augmentation with controllable rewriting in continuous space. In: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, pp. 5798–5810. https://doi.org/10.18653/v1/2020.emnlp-main.467, 10.18653/v1/2020.emnlp-main.467.

Liu, L., Ding, B., Bing, L., Joty, S., Si, L., Miao, C., 2021. MulDA: a multilingual data augmentation framework for low-resource cross-lingual NER. Online. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, vol. 1. Long Papers), Association for Computational Linguistics, pp. 5834–5846. https://aclanthology.org/2021.acl-long.453, 10.18653/v1/2021.acl-long.453.

Liu, P., Wang, X., Xiang, C., Meng, W., 2020b. A survey of text data augmentation. In: 2020 International Conference on Computer Communication and Network Security (CCNS). IEEE, pp. 191–195. https://doi.org/10.1109/CCNS50731.2020.00049.

Liu, Q., Guan, W., Li, S., Cheng, F., Kawahara, D., Kurohashi, S., 2020c. Reverse Operation Based Data Augmentation for Solving Math Word Problems. CoRR abs/2010.01556. https://arxiv.org/abs/2010.01556. arXiv:2010.01556.

Liu, R., Xu, G., Jia, C., Ma, W., Wang, L., Vosoughi, S., 2020d. Data boost: text data augmentation through reinforcement learning guided conditional generation. Online, November 16-20, 2020. In: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020. Association for Computational Linguistics, pp. 9031–9041. https://doi.org/10.18653/v1/2020.emnlp-main.726, 10.18653/v1/2020.emnlp-main.726.

Liu, S., Lee, K., Lee, I., 2020e. Document-level multi-topic sentiment classification of email data with bilstm and data augmentation. Knowl. Base Syst. 197, 105918 https://doi.org/10.1016/j.knosys.2020.105918, 10.1016/j.knosys.2020.105918.

Longpre, S., Lu, Y., Tu, Z., DuBois, C., 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. Association for Computational Linguistics, Hong Kong, China, pp. 220–227. https://aclanthology.org/D19-5829, 10.18653/v1/D19-5829.

Longpre, S., Wang, Y., DuBois, C., 2020. How effective is task-agnostic data augmentation for pretrained transformers? Online. In: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, pp. 4401–4411. https://www.aclweb.org/anthology/2020.findings-emnlp.394, 10.18653/v1/2020.findings-emnlp.394.

Louvan, S., Magnini, B., 2020. Simple is better! lightweight data augmentation for low resource slot filling and intent classification. October 24-26, 2020. In: Nguyen, M.L., Luong, M.C., Song, S. (Eds.), Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation, PACLIC 2020. Association for Computational Linguistics, Hanoi, Vietnam, pp. 167–177. https://aclanthology.org/2020.paclic-1.20/.

Lowell, D., Howard, B.E., Lipton, Z.C., Wallace, B.C., 2021. Unsupervised data augmentation with naive augmentation and without unlabeled data. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event/Punta Cana, Dominican Republic, 7-11 November, 2021. Association for Computational Linguistics, pp. 4992–5001. https://aclanthology.org/2021.emnlp-main.408.

Lun, J., Zhu, J., Tang, Y., Yang, M., 2020. Multiple data augmentation strategies for improving performance on automatic short answer scoring. February 7-12, 2020. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020. AAAI Press, New York, NY, USA, pp. 13389–13396. https://aaai.org/ojs/index.php/AAAI/article/view/7062.

Luque, F.M., 2019. Atalaya at TASS 2019: data augmentation and robust embeddings for sentiment analysis. September 24th, 2019. In: Cumbreras, M.Á.G., Gonzalo, J., Cámara, E.M., Martínez-Unanue, R., Rosso, P., Carrillo-de-Albornoz, J., Montalvo, S., Chiruzzo, L., Collovini, S., Gutiérrez, Y., Zafra, S.M.J., Krallinger, M., Montes-y-Gómez, M., Ortega-Bueno, R., Rosá, A. (Eds.), Proceedings of the Iberian Languages Evaluation Forum Co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019. CEUR-WS.org, Bilbao, Spain, pp. 561–570. http://ceur-ws.org/Vol-2421/TASS_paper_1.pdf.

Madnani, N., Dorr, B.J., 2010. Generating phrasal and sentential paraphrases: a survey of data-driven methods. Comput. Ling. 36, 341–387. https://www.aclweb.org/anthology/J10-3003, 10.1162/coli_a_00002.

Malandrakis, N., Shen, M., Goyal, A.K., Gao, S., Sethi, A., Metallinou, A., 2019. Controlled text generation for data augmentation in intelligent artificial agents. November 4, 2019. In: Birch, A., Finch, A.M., Hayashi, H., Konstas, I., Luong, T., Neubig, G., Oda, Y., Sudoh, K. (Eds.), Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP 2019. Association for Computational Linguistics, Hong Kong, pp. 90–98. https://doi.org/10.18653/v1/D19-5609, 10.18653/v1/D19-5609.

Miao, L., Last, M., Litvak, M., 2020. Twitter data augmentation for monitoring public opinion on COVID-19 intervention measures. Online, December 2020. In: Verspoor, K., Cohen, K.B., Conway, M., de Bruijn, B., Dredze, M., Mihalcea, R., Wallace, B.C. (Eds.), Proceedings of the 1st Workshop on NLP for COVID-19@EMNLP 2020. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.nlpcovid19-2.19, 10.18653/v1/2020.nlpcovid19-2.19.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, 5-8, 2013. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December, Lake Tahoe, Nevada, United States, pp. 3111–3119. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Miller, G.A., 1995. Wordnet: a lexical database for English. Commun. ACM 38, 39–41. http://doi.org/10.1145/219717.219748, 10.1145/219717.219748.

Min, J., McCoy, R.T., Das, D., Pitler, E., Linzen, T., 2020. Syntactic data augmentation increases robustness to inference heuristics. Online, July 5-10, 2020. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. Association for Computational Linguistics, pp. 2339–2352. https://doi.org/10.18653/v1/2020.acl-main.212, 10.18653/v1/2020.acl-main.212.

Montella, S., Fabre, B., Urvoy, T., Heinecke, J., Rojas-Barahona, L.M., 2020. Denoising pre-training and data augmentation strategies for enhanced RDF verbalization with transformers. arXiv abs/2012.00571. https://arxiv.org/abs/2012.00571. arXiv:2012.00571.

Morris, J., Lifland, E., Yoo, J.Y., Grigsby, J., Jin, D., Qi, Y., 2020. TextAttack: a framework for adversarial attacks, data augmentation, and adversarial training in NLP. Online. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, pp. 119–126. https://aclanthology.org/2020.emnlp-demos.16, 10.18653/v1/2020.emnlp-demos.16.

Mou, X., Sigouin, B., Steenstra, I., Su, H., 2020. Multimodal dialogue state tracking by QA approach with data augmentation. arXiv abs/2007.09903. https://arxiv.org/abs/2007.09903. arXiv:2007.09903.

Mueller, J., Thyagarajan, A., 2016. Siamese recurrent architectures for learning sentence similarity. In: Schuurmans, D., Wellman, M.P. (Eds.), Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016. AAAI Press, Phoenix, Arizona, USA, pp. 2786–2792. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12195.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al., 2021. Webgpt: Browser-Assisted Question-answering with Human Feedback arXiv preprint arXiv:2112.09332.

Ng, N., Cho, K., Ghassemi, M., 2020. SSMBA: self-supervised manifold based data augmentation for improving out-of-domain robustness. Online, November 16-20, 2020. In: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020. Association for Computational Linguistics, pp. 1268–1283. https://doi.org/10.18653/v1/2020.emnlp-main.97, 10.18653/v1/2020.emnlp-main.97.

Nishikawa, S., Ri, R., Tsuruoka, Y., 2020. Data augmentation for learning bilingual word embeddings with unsupervised machine translation. CoRR abs/2006.00262. https://arxiv.org/abs/2006.00262. arXiv:2006.00262.

Niu, T., Bansal, M., 2019. Automatically learning data augmentation policies for dialogue tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp. 1317–1323. https://aclanthology.org/D19-1132, 10.18653/v1/D19-1132.

Nugent, T., Stelea, N., Leidner, J.L., 2021. Detecting environmental, social and governance (ESG) topics using domain-specific language models and data augmentation. September 19-24, 2021. In: Andreasen, T., Tré, G.D., Kacprzyk, J., Larsen, H.L., Bordogna, G., Zadrozny, S. (Eds.), Proceedings of the 14th International Conference on Flexible Query Answering Systems (FQAS 2021). Springer, Bratislava, Slovakia, pp. 157–169. https://doi.org/10.1007/978-3-030-86967-0_12.

Palomino, D., Luna, J.O., 2020. Palomino-ochoa at TASS 2020: transformer-based data augmentation for overcoming few-shot learning. September 23th, 2020. In: Cumbreras, M.Á.G., Gonzalo, J., Cámara, E.M., Martínez-Unanue, R., Rosso, P., Zafra, S.M.J., Zambrano, J.A.O., Miranda, A., Zamorano, J.P., Gutiérrez, Y., Rosá, A., Montes-y-Gómez, M., Vega, M.G. (Eds.), Proceedings of the Iberian Languages

Evaluation Forum (IberLEF 2020) Co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020). CEUR-WS.org, Málaga, Spain, pp. 171–178. http://ceur-ws.org/Vol-2664/tass_paper1.pdf.
Peng, B., Zhu, C., Zeng, M., Gao, J., 2020. Data augmentation for spoken language understanding via pretrained models. arXiv abs/2004.13952. https://arxiv.org/abs/2004.13952. arXiv:2004.13952.
Perevalov, A., Both, A., 2020. Augmentation-based answer type classification of the SMART dataset. November 5th, 2020. In: Mihindukulasooriya, N., Dubey, M., Gliozzo, A., Lehmann, J., Ngomo, A.N., Usbeck, R. (Eds.), Proceedings of the SeMantic AnsweR Type Prediction Task (SMART) at ISWC 2020 Semantic Web Challenge Co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual Conference. CEUR-WS.org, pp. 1–9. http://ceur-ws.org/Vol-2774/paper-01.pdf.
Qin, L., Ni, M., Zhang, Y., Che, W., 2020. Cosda-ml: multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In: Bessiere, C. (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. IJCAI 2020, ijcai.org, pp. 3853–3860. https://doi.org/10.24963/ijcai.2020/533, 10.24963/ijcai.2020/533.
Qu, Y., Shen, D., Shen, Y., Sajeev, S., Chen, W., Han, J., 2021. Coda: contrast-enhanced and diversity-promoting data augmentation for natural language understanding. In: 9th International Conference on Learning Representations. ICLR 2021, Virtual Event, Austria. May 3-7, 2021, OpenReview.net. https://openreview.net/forum?id=Ozk9MrX1hvA.
Quteineh, H., Samothrakis, S., Sutcliffe, R., 2020. Textual data augmentation for efficient active learning on tiny datasets. Online, November 16-20, 2020. In: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020. Association for Computational Linguistics, pp. 7400–7410. https://doi.org/10.18653/v1/2020.emnlp-main.600, 10.18653/v1/2020.emnlp-main.600.
Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21 (140), 1–140, 67. http://jmlr.org/papers/v21/20-074.html.
Raille, G., Djambazovska, S., Musat, C., 2020. Fast cross-domain data augmentation through neural sentence editing. arXiv abs/2003.10254. https://arxiv.org/abs/2003.10254. arXiv:2003.10254.
Ramirez-Echavarria, D., Bikakis, A., Dickens, L., Miller, R., Vlachidis, A., 2020. On the effects of knowledge-augmented data in word embeddings. CoRR abs/2010.01745. https://arxiv.org/abs/2010.01745. arXiv:2010.01745.
Rastogi, C., Mofid, N., Hsiao, F., 2020. Can we achieve more with less? exploring data augmentation for toxic comment classification. arXiv abs/2007.00875. https://arxiv.org/abs/2007.00875. arXiv:2007.00875.
Regina, M., Meyer, M., Goutal, S., 2020. Text data augmentation: towards better detection of spear-phishing emails. arXiv abs/2007.02033. https://arxiv.org/abs/2007.02033. arXiv:2007.02033.
Riabi, A., Scialom, T., Keraron, R., Sagot, B., Seddah, D., Staiano, J., 2021. Synthetic data augmentation for zero-shot cross-lingual question answering, 7-11 November, 2021. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event/Punta Cana, Dominican Republic. Association for Computational Linguistics, pp. 7016–7030. https://aclanthology.org/2021.emnlp-main.562.
Risch, J., Krestel, R., 2018. Aggression identification using deep learning and data augmentation. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 150–158. https://aclanthology.org/W18-4418.
Sahin, G.G., Steedman, M., 2019. Data augmentation via dependency tree morphing for low-resource languages. arXiv abs/1903.09460. http://arxiv.org/abs/1903.09460. arXiv:1903.09460.
Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR abs/1910.01108. http://arxiv.org/abs/1910.01108. arXiv:1910.01108.
Schuler, K.K., 2005. VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. University of Pennsylvania.
Sennrich, R., Haddow, B., Birch, A., 2016a. Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ume 1. Long Papers), Association for Computational Linguistics, Berlin, Germany, pp. 86–96. https://aclanthology.org/P16-1009, 10.18653/v1/P16-1009.
Sennrich, R., Haddow, B., Birch, A., 2016b. Improving neural machine translation models with monolingual data. ACL 2016, August 7-12, 2016. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ume 1. Long Papers, The Association for Computer Linguistics, Berlin, Germany. https://doi.org/10.18653/v1/p16-1009, 10.18653/v1/p16-1009.
Shakeel, M.H., Karim, A., Khan, I., 2020. A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts. Inf. Process. Manag. 57, 102204 https://doi.org/10.1016/j.ipm.2020.102204, 10.1016/j.ipm.2020.102204.
Shehnepoor, S., Togneri, R., Liu, W., Bennamoun, M., 2020. Gangster: a fraud review detector based on regulated GAN with data augmentation. CoRR abs/2006.06561. https://arxiv.org/abs/2006.06561. arXiv:2006.06561.
Shi, H., Livescu, K., Gimpel, K., 2021. Substructure substitution: structured data augmentation for NLP. Online Event, August 1-6, 2021. In: Zong, C., Xia, F., Li, W., Navigli, R. (Eds.), Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021. Association for Computational Linguistics, pp. 3494–3508. https://doi.org/10.18653/v1/2021.findings-acl.307, 10.18653/v1/2021.findings-acl.307.
Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6, 60. https://doi.org/10.1186/s40537-019-0197-0, 10.1186/s40537-019-0197-0.
Si, C., Zhang, Z., Qi, F., Liu, Z., Wang, Y., Liu, Q., Sun, M., 2020. Better robustness by more coverage: adversarial training with mixup augmentation for robust fine-tuning. arXiv abs/2012.15699. https://arxiv.org/abs/2012.15699. arXiv:2012.15699.
Si, C., Zhang, Z., Qi, F., Liu, Z., Wang, Y., Liu, Q., Sun, M., 2021. Better robustness by more coverage: adversarial and mixup data augmentation for robust finetuning. Online. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, pp. 1569–1576. https://aclanthology.org/2021.findings-acl.137, 10.18653/v1/2021.findings-acl.137.
Singh, D., Reddy, S., Hamilton, W., Dyer, C., Yogatama, D., 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. Adv. Neural Inf. Process. Syst. 34, 25968–25981.
Song, X., Zang, L., Hu, S., 2021. Data augmentation for copy-mechanism in dialogue state tracking. June 16-18, 2021. In: Paszynski, M., Kranzlmüller, D., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M.A. (Eds.), Computational Science - ICCS 2021 - 21st International Conference. Proceedings, Part I, Springer, Krakow, Poland, pp. 736–749. https://doi.org/10.1007/978-3-030-77961-0_59, 10.1007/978-3-030-77961-0_59.
Staliunaite, I., Gorinski, P.J., Iacobacci, I., 2021. Improving commonsense causal reasoning by adversarial training and data augmentation. February 2-9, 2021. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, the Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event. AAAI Press, pp. 13834–13842. https://ojs.aaai.org/index.php/AAAI/article/view/17630.
Sun, L., Xia, C., Yin, W., Liang, T., Yu, P.S., He, L., 2020. Mixup-transformer: dynamic data augmentation for NLP tasks (Online), December 8-13, 2020. In: Scott, D., Bel, N., Zong, C. (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020. International Committee on Computational Linguistics, Barcelona, Spain, pp. 3436–3440. https://doi.org/10.18653/v1/2020.coling-main.305, 10.18653/v1/2020.coling-main.305.
Tapia-Téllez, J.M., Escalante, H.J., 2020. Data augmentation with transformers for text classification. October 12-17, 2020. In: Martínez-Villaseñor, L., Herrera-Alcántara, O., Ponce, H.E., Castro-Espinoza, F. (Eds.), Advances in Computational Intelligence - 19th Mexican International Conference on Artificial Intelligence, MICAI 2020. Proceedings, Part II, Springer, Mexico City, Mexico, pp. 247–259. https://doi.org/10.1007/978-3-030-60887-3_22, 10.1007/978-3-030-60887-3_22.
Tarján, B., Szaszák, G., Fegyó, T., Mihajlik, P., 2020. Deep transformer based data augmentation with subword units for morphologically rich online ASR. arXiv abs/2007.06949. https://arxiv.org/abs/2007.06949. arXiv:2007.06949.
Thakur, N., Reimers, N., Daxenberger, J., Gurevych, I., 2021. Augmented SBERT: data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. Online, June 6-11, 2021. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tür, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021. Association for Computational Linguistics, pp. 296–310. https://doi.org/10.18653/v1/2021.naacl-main.28, 10.18653/v1/2021.naacl-main.28.
Wan, Z., Wan, X., Wang, W., 2020. Improving grammatical error correction with data augmentation by editing latent representation (Online), December 8-13, 2020. In: Scott, D., Bel, N., Zong, C. (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020. International Committee on Computational Linguistics, Barcelona, Spain, pp. 2202–2212. https://doi.org/10.18653/v1/2020.coling-main.200, 10.18653/v1/2020.coling-main.200.
Wang, J., Chen, H.C., Radach, R., Inhoff, A., 1999. Reading Chinese Script: A Cognitive Analysis. Psychology Press.
Wang, W.Y., Yang, D., 2015. That's so annoying!!!: a lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. Lisbon, Portugal, September 17-21, 2015. In: Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y. (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, pp. 2557–2563. https://doi.org/10.18653/v1/d15-1306, 10.18653/v1/d15-1306.
Wang, X., Pham, H., Dai, Z., Neubig, G., 2018. Switchout: an efficient data augmentation algorithm for neural machine translation. October 31 - November 4, 2018. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Association for Computational Linguistics, pp. 856–861. https://doi.org/10.18653/v1/d18-1100, 10.18653/v1/d18-1100.
Jason W. Wei and Kai Zou. 2019, EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, Kentaro Inui, Jing Jiang, Vincent Ng, Xiaojun Wan, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, {EMNLP-IJCNLP} 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 6381–6387, https://doi.org/10.18653/v1/D19-1670.
Wu, S., Zhang, H., Valiant, G., Ré, C., 2020. On the generalization effects of linear transformations in data augmentation. In: International Conference on Machine Learning. PMLR, pp. 10410–10420.
Wu, X., Lv, S., Zang, L., Han, J., Hu, S., 2019. Conditional Bert Contextual Augmentation. ICCS.
Xia, M., Kong, X., Anastasopoulos, A., Neubig, G., 2019. Generalized data augmentation for low-resource translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational

Linguistics, Florence, Italy, pp. 5786–5796. https://aclanthology.org/P19-1579, 10.18653/v1/P19-1579.

Xie, Q., Dai, Z., Hovy, E.H., Luong, T., Le, Q., 2020. Unsupervised data augmentation for consistency training. December 6-12, 2020. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020. virtual. https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html.

Xie, Z., Wang, S.I., Li, J., Lévy, D., Nie, A., Jurafsky, D., Ng, A.Y., 2017. Data noising as smoothing in neural network language models. April 24-26, 2017. In: 5th International Conference on Learning Representations, ICLR 2017. Conference Track Proceedings, Toulon, France. OpenReview.net. https://openreview.net/forum?id=H1VyHY9gg.

Xu, B., Qiu, S., Zhang, J., Wang, Y., Shen, X., de Melo, G., 2020. Data augmentation for multiclass utterance classification - a systematic study (Online), December 8-13, 2020. In: Scott, D., Bel, N., Zong, C. (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020. International Committee on Computational Linguistics, Barcelona, Spain, pp. 5494–5506. https://doi.org/10.18653/v1/2020.coling-main.479, 10.18653/v1/2020.coling-main.479.

Xu, X., Wang, G., Kim, Y.B., Lee, S., 2021. AugNLG: Few-shot natural language generation using self-trained data augmentation. Online. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ume 1. Long Papers), Association for Computational Linguistics, pp. 1183–1195. https://aclanthology.org/2021.acl-long.95, 10.18653/v1/2021.acl-long.95.

Xu, Y., Jia, R., Mou, L., Li, G., Chen, Y., Lu, Y., Jin, Z., 2016. Improved relation classification by deep recurrent neural networks with data augmentation. December 11-16, 2016. In: Calzolari, N., Matsumoto, Y., Prasad, R. (Eds.), COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers. ACL, Osaka, Japan, pp. 1461–1470. https://aclanthology.org/C16-1138/.

Yan, G., Li, Y., Zhang, S., Chen, Z., 2019. Data augmentation for deep learning of judgment documents. October 17-20, 2019, Proceedings, Part II. In: Cui, Z., Pan, J., Zhang, S., Xiao, L., Yang, J. (Eds.), Intelligence Science and Big Data Engineering. Big Data and Machine Learning - 9th International Conference, IScIDE 2019. Springer, Nanjing, China, pp. 232–242. https://doi.org/10.1007/978-3-030-36204-1_19, 10.1007/978-3-030-36204-1_19.

Yang, Y., Jin, N., Lin, K., Guo, M., Cer, D., 2021. Neural retrieval for question answering with cross-attention supervised data augmentation. Short Papers), Virtual Event, August 1-6, 2021. In: Zong, C., Xia, F., Li, W., Navigli, R. (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, ume 2. Association for Computational Linguistics, pp. 263–268. https://doi.org/10.18653/v1/2021.acl-short.35, 10.18653/v1/2021.acl-short.35.

Yao, L., Yang, B., Zhang, H., Chen, B., Luo, W., 2020. Domain transfer based data augmentation for neural query translation (Online), December 8-13, 2020. In: Scott, D., Bel, N., Zong, C. (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020. International Committee on Computational Linguistics, Barcelona, Spain, pp. 4521–4533. https://doi.org/10.18653/v1/2020.coling-main.399, 10.18653/v1/2020.coling-main.399.

Yin, Y., Shang, L., Jiang, X., Chen, X., Liu, Q., 2020. Dialog state tracking with reinforced data augmentation. February 7-12, 2020. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020. AAAI Press, New York, NY, USA, pp. 9474–9481. https://aaai.org/ojs/index.php/AAAI/article/view/6491.

Yogatama, D., de Masson d'Autume, C., Kong, L., 2021. Adaptive semiparametric language models. Trans. Assoc. Comput. Ling. 9, 362–373.

Yoo, K.M., Lee, H., Dernoncourt, F., Bui, T., Chang, W., Lee, S., 2020. Variational hierarchical dialog autoencoder for dialog state tracking data augmentation. Online, November 16-20, 2020. In: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020. Association for Computational Linguistics, pp. 3406–3425. https://doi.org/10.18653/v1/2020.emnlp-main.274, 10.18653/v1/2020.emnlp-main.274.

Yoo, K.M., Shin, Y., Lee, S., 2019. Data augmentation for spoken language understanding via joint variational generation. January 27 - February 1, 2019. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019. AAAI Press, Honolulu, Hawaii, USA, pp. 7402–7409. https://doi.org/10.1609/aaai.v33i01.33017402, 10.1609/aaai.v33i01.33017402.

Yu, A.W., Dohan, D., Luong, M., Zhao, R., Chen, K., Norouzi, M., Le, Q.V., 2018. Qanet: combining local convolution with global self-attention for reading comprehension. April 30 - May 3, 2018. In: 6th International Conference on Learning Representations, ICLR 2018. Conference Track Proceedings, Vancouver, BC, Canada. OpenReview.net. https://openreview.net/forum?id=B14TlG-RW.

Yu, S., Yang, J., Liu, D., Li, R., Zhang, Y., Zhao, S., 2019. Hierarchical data augmentation and the application in text classification. IEEE Access 7, 185476–185485. https://doi.org/10.1109/ACCESS.2019.2960263, 10.1109/ACCESS.2019.2960263.

Zhang, D., Li, T., Zhang, H., Yin, B., 2020a. On data augmentation for extreme multi-label classification. arXiv abs/2009.10778. https://arxiv.org/abs/2009.10778. arXiv:2009.10778.

Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D., 2018. mixup: beyond empirical risk minimization. April 30 - May 3, 2018. In: 6th International Conference on Learning Representations, ICLR 2018. Conference Track Proceedings, Vancouver, BC, Canada. OpenReview.net. https://openreview.net/forum?id=r1Ddp1-Rb.

Zhang, R., Zheng, Y., Shao, J., Mao, X., Xi, Y., Huang, M., 2020b. Dialogue distillation: open-domain dialogue augmentation using unpaired data. Online, November 16-20, 2020. In: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020. Association for Computational Linguistics, pp. 3449–3460. https://doi.org/10.18653/v1/2020.emnlp-main.277, 10.18653/v1/2020.emnlp-main.277.

Zhang, X., Zhao, J.J., LeCun, Y., 2015. Character-level convolutional networks for text classification. December 7-12, 2015, Montreal, Quebec, Canada. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, pp. 649–657. https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Zhang, Y., Ge, T., Sun, X., 2020c. Parallel data augmentation for formality style transfer. Online, July 5-10, 2020. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. Association for Computational Linguistics, pp. 3221–3228. https://doi.org/10.18653/v1/2020.acl-main.294, 10.18653/v1/2020.acl-main.294.

Zhao, Z., Papalexakis, E.E., Ma, X., 2020. Learning physical common sense as knowledge graph completion via BERT data augmentation and constrained tucker factorization. Online, November 16-20, 2020. In: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020. Association for Computational Linguistics, pp. 3293–3298. https://doi.org/10.18653/v1/2020.emnlp-main.266, 10.18653/v1/2020.emnlp-main.266.

Zhao, Z., Zhu, S., Yu, K., 2019. Data augmentation with atomic templates for spoken language understanding. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp. 3637–3643. https://aclanthology.org/D19-1375, 10.18653/v1/D19-1375.

Zheng, B., Dong, L., Huang, S., Wang, W., Chi, Z., Singhal, S., Che, W., Liu, T., Song, X., Wei, F., 2021. Consistency regularization for cross-lingual fine-tuning. Long Papers), Virtual Event, August 1-6, 2021. In: Zong, C., Xia, F., Li, W., Navigli, R. (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, vol 1. Association for Computational Linguistics, pp. 3403–3417. https://doi.org/10.18653/v1/2021.acl-long.264. :10.18653/v1/2021.acl-long.264.

Zhou, J., Zheng, Y., Tang, J., Jian, L., Yang, Z., 2022. FlipDA: effective and robust data augmentation for few-shot learning. Long Papers. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, vol 1. Association for Computational Linguistics, Dublin, Ireland, pp. 8646–8665. https://aclanthology.org/2022.acl-long.592, 10.18653/v1/2022.acl-long.592.

Zhou, Y., Dong, F., Liu, Y., Li, Z., Du, J., Zhang, L., 2020. Forecasting emerging technologies using data augmentation and deep learning. Scientometrics 123, 1–29. https://doi.org/10.1007/s11192-020-03351-6, 10.1007/s11192-020-03351-6.

Zmigrod, R., Mielke, S.J., Wallach, H.M., Cotterell, R., 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. July 28- August 2, 2019. In: Korhonen, A., Traum, D.R., Màrquez, L. (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, ume 1. Long Papers, Association for Computational Linguistics, Florence, Italy, pp. 1651–1661. https://doi.org/10.18653/v1/p19-1161, 10.18653/v1/p19-1161.

Zuo, X., Chen, Y., Liu, K., Zhao, J., 2020. Knowdis: knowledge enhanced data augmentation for event causality detection via distant supervision (Online), December 8-13, 2020. In: Scott, D., Bel, N., Zong, C. (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020. International Committee on Computational Linguistics, Barcelona, Spain, pp. 1544–1550. https://doi.org/10.18653/v1/2020.coling-main.135, 10.18653/v1/2020.coling-main.135.