Abstract

作为一种有效的策略，数据增强(data augmentation, DA)缓解了深度学习技术可能失败的数据稀缺场景。它被广泛应用于计算机视觉，然后引入到自然语言处理中，在许多任务中取得了改进。提高训练数据的多样性是DA方法的一个主要重点，从而帮助模型更好地 泛化到未见测试数据。在这项综述中，根据扩充数据的多样性将数据增强方法分为三类，包括paraphrasing, noising, and sampling。本文根据上述分类对DA方法进行了详细的分析。此外，我们还介绍了它们在自然语言处理任务中的应用以及面临的挑战。

1. Introduction

数据扩充是指通过添加已经存在的数据的轻微修改的副本或根据现有数据新创建的人造数据来增加数据量的方法。这些方法缓解了深度学习技术可能失败的数据稀缺场景，因此DA最近受到了 极大的关注和需求。数据增强技术被广泛应用于计算机视觉[1]领域，如翻转、旋转等，并引入到自然语言处理(NLP)中。与图像不同，自然语言是离散的，这使得DA方法在自然语言处理中的应用更加困难和缺乏探索。

在这篇综述中，我们将全面概述自然语言处理中的DA方法。我们的主要目标之一是展示DA的本质，即为什么数据增强有效。为了实现这一点，我们根据增强数据的多样性对数据分析方法进行分类，因为提高训练数据的多样性是数据分析有效性的主要推动力之一。我们将DA方法分为三类，包括转述、噪声和抽样（paraphrasing, noising, and sampling.）。

具体而言，基于paraphrasing的方法将原数据的paraphrases生成为扩充数据。与原始数据相比，这类数据带来的变化有限。基于噪声的方法在原始数据中加入了更多的连续或离散噪声，涉及到更多的变化。基于抽样的方法掌握了原始数据的分布，将采样生成的新数据作为扩展数据。在人工启发式和训练模型的帮助下，这种方法可以采样生成全新的数据，而不是改变现有的数据，从而产生更多样化的数据。

2.Data Augmentation Method in NLP

数据增强的目的是在不充分的数据场景中生成额外的、人造的训练数据。数据增强从简单的基于规则的方法到可学习的基于生成的方法，这些方法都从本质上保证了增强数据的有效性。也就是说，DA方法需要确保增强的数据对于任务是有效的，即被认为是原始数据相同分布的一部分。例如，机器翻译时语义相似，文本分类时标签与原始数据相同。

基于有效性， 增强数据需要保证多样性，使下游方法具有更好的泛化能力。这涉及到增强数据的多样性。不同的多样性涉及到不同的方法和相应的增强效果。在这项调查中，我们根据其增强数据的多样性，新奇地将DA方法分为三类:转述、噪声和抽样（paraphrasing, noising, and sampling）。

- 基于复述的方法通过对句子进行适当而克制的改动，生成与原数据语义差异有限的增强数据。增强后的数据传递的信息与原始形式非常相似。
- 基于噪声的方法在保证有效性的前提下添加离散或连续噪声。这些方法的重点是提高模型的鲁棒性。
- 基于抽样的方法掌握数据分布，并在其中抽取新的点。这种方法基于人工启发式和训练模型，输出更多样化的数据，满足下游任务的更多需求。

转述、噪声和基于抽样的方法 产生的增强数据的多样性依次增强。

2.1 基于转述(Paraphrasing)的方法

转述是自然语言中常见的一种表达方式，是表达与原文相同信息的另一种方式。自然，意译的生成是数据扩充的合适解决方案。意译可以发生在几个层次，包括词汇意译、短语意译和句子意译。因此，意译生成的数据增强技术也包括这三种重写。

2.1.1 Thesauruses(同义词)

有些工作将原文中的词语替换成其真实的同义词和超义词，从而在尽可能保持原文语义不变的同时获得一种新的表达方式。如图4所示，像WordNet这样的词库包含单词的词汇三元组，通常用作外部资源。

Zhang等人率先将词库应用于数据增强。他们使用一个来自WordNet的辞典，该辞典根据单词的相似度对同义词进行分类。对于每个句子，他们检索所有可替换词，并随机选择r个可替换词。数r的概率是由一个参数为p的几何分布决定的，其中$P[r] \sim p^r$ 。根据给定的单词，选择的同义词的索引s也由$P[s] \sim p^s$的另一个几何分布决定，这种方法确保了选择与原单词更接近的同义词的概率更大。

一种被广泛使用的文本增强方法EDA (Easy Data augmentation Techniques)也会使用WordNet用它们的同义词替换原来的单词：他们从句子中随机选择n个不是停用词的单词，并用随机选择的同义词替换这些单词，而不是按照几何分布。

除了同义词外，Coulombe等人建议使用上下义来替代原词。他们还推荐了词汇替换的候选词汇类型，按照难度增加的顺序:副词、形容词、名词和动词。Zuo等人使用WordNet和VerbNet检索同义词、hypernym和同类词。

Thesauruses
Advantage(s):
1. Easy to use.

Limitation(s):
1. The scope and part-of-speech of replacement words are limited.
2. This method cannot solve the problem of ambiguity.
3. The sentence semantics may be affected if too many replacements occur.

2.1.2 Semantic Embeddings(语义编码)

该方法克服了基于同义词法的替换范围和词性的限制。它使用预先训练好的词向量，如Glove、Word2Vec、FastText等，用向量空间中最接近原始词的词替换它们，如图5所示。

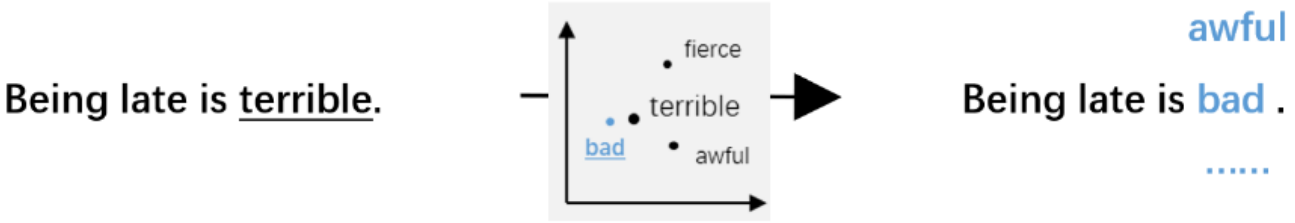


Figure 5: Paraphrasing by using semantic embeddings.

在推特消息分类任务中，Wang等人首创了同时使用word embeddings和frame embeddings来代替离散词。在word embeddings方面，利用余弦相似度将推文中每个原始单词替换为k近邻中的一个单词。例如，“Being late is terrible”变成了“Being behind are bad”。在frame embeddings方面，作者对380万条推文进行了语义解析，并使用Word2Vec[48]构建了一个连续的袋框架模型来表示每个语义框架。然后将与单词相同的数据增强方法应用于语义框架。与Wang et al.[8]相比，Liu et al.[49]只使用词嵌入来检索同义词。同时，他们用词典对检索结果进行编辑，以达到平衡。RamirezEchavarria等人[50]创建了用于选择的嵌入字典。

Semantic Embeddings
Advantage(s):
1. Easy to use.
2. Higher replacement hit rate and wider replacement range.

Limitation(s):
1. This method cannot solve the problem of ambiguity.
2. The sentence semantics may be affected if too many replacements occur.

2.1.3 Language Model

预训练语言模型因其优异的性能成为近年来的主流模型。Masked language models(MLM)如BERT和BoBERTa通过预训练后，可以在文本中基于上下文预测被mask的单词,可以用于文本数据增强(如图6)。此外,由于MLMs考虑整个上下文，这种方法可以解决歧义问题。

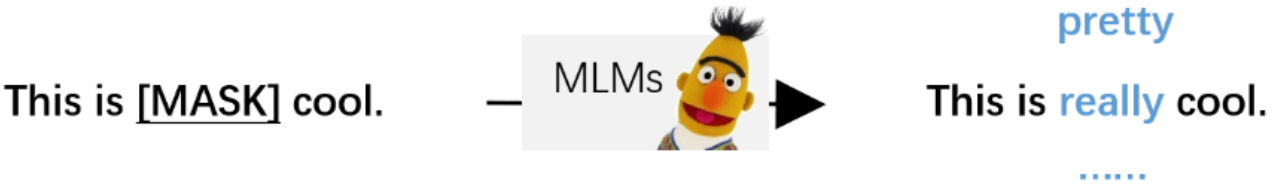


Figure 6: Paraphrasing by using language models.

Jiao等人[9]使用数据增强来获得特定任务的精馏训练数据。它们应用BERT的 分词器将单词分为多个子词，并为每个子词形成候选集。词的替换采用了词的word Embeddings和MLM。具体地说，如果一个词片段不是一个完整的词(例如est)，候选集是由它的k近邻词由Glove组成的。如果词块是一个完整的词，作者将其替换为[MASK]，利用BERT预测K个词，形成候选集。最后，以0.4的概率来确定每个词块是否被候选集中的一个随机词替换。Regina等[10]，Tapia-Téllez等[51]，Lowell等[52]和Palomino等[53]采用了类似的方法。他们在一个句子中 mask多个单词，并通过填充这些被mask的词产生更多不同的句子来生成新

的句子。此外，rnn也 被用来根据上下文替换原词([54,55])。

Language Models
Advantage(s):
1. This method alleviates the problem of ambiguity.
2. This method considers context semantics.

Limitation(s):
1. Still limited to the word level.
2. The sentence semantics might be affected if too many replacements occur.

2.1.4 Rules

该方法需要一些关于自然语言的启发式方法，以确保语句语义的维护，如图7所示。

Figure 7: Paraphrasing by using rules.

一方面，一些工作依赖现有的词典或固定的启发式来生成单词级和短语级的 复述。Coulombe等[7]引入了使用正则表达式在不改变句子语义的情况下进行形式转换，如动词的缩写和原型、情态动词、否定等。例如，将“is not”替换为“isn 't”。类似地，Regina等人[10]根据词对词典执行从扩展到缩写形式的替换，并在一组单词和相应的缩写之间进行反向替换。另一方面，有些工作通过一定的规则，如依存树对原句进行句子级的复述。Coulombe等人[7]首先通过依赖树引入了一种方法。他们使用语法分析器为原始句子构建依存树。然后，复述生成器转换这个依存关系树，在转换语法的指导下创建一个转换后的依存关系树。例如，“Sally embraced Peter excitedly.”被转换为“Peter was embraced excitedly by Sally.”。然后，使用转换后的依存关系树生成作为增强数据的复述。Dehouck等人[56]采用了类似的方法。Louvan等人[11]在依存树上裁剪特定的片段，以创建更小的句子。他们还会在依赖项解析结构的根节点上旋转目标片段，而不会损害原始语义。

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

翻译是一种复述的神经方法。随着机器翻译模型的发展和在线api的可用性，机器翻译作为许多任务中的增强方法受到欢迎，如图8所示。

Figure 8: Paraphrasing by machine translation.

反向翻译（Back-translation）

这种方法指着将原始文档翻译成其他语言，然后再翻译回来，以获得原始语言的新文本。与单词层次的方法不同，反向翻译不是直接替换单个单词，而是用生成的方式重写整个句子。

Xie等人[12]，Yu等人[57]，Fabbri等人[58]使用英法翻译模型(双向)对每个句子进行反向翻译，获得其复述。Lowell等[52]也将该方法作为一种无监督数据增强方法引入。Zhang等人[13]利用反向翻译在风格转换任务中获得原始数据的形式化表达。

除了一些经过训练的机器翻译模型外，谷歌的云翻译API服务是一种常用的反向翻译工具，被广泛应用。

一些工作添加了基于普通反向翻译的额外功能。Nugent等人[64]提出了一系列softmax温度系数设置，以确保多样性，同时保留语义。Qu等人[65]将反向翻译与对抗性训练相结合，通过有机地集成多种转换来综合多样化和信息性增强示例。Zhang等人[13]利用判别器对反翻译结果中的句子进行过滤。该方法因为阈值的存在大大提高了增强数据的质量。

单向翻译（Unidirectional Translation）

与反向翻译不同，单向翻译方法是直接将原文一次翻译成其他语言，而不需要将其翻译回原文。这种方法通常发生在多语言场景中。

在无监督跨语言词嵌入(CLWEs)任务中，Nishikawa等人利用无监督机器翻译模型建立了伪平行语料库。作者首先利用源/目标训练语料库训练无监督机器翻译(UMT)模型，然后利用 UMT 对语料库进行翻译。

将机器翻译的语料库与原始语料库连接起来，独立学习每种语言的单语词嵌入。最后，将学习到的单语词嵌入映射到共享的CLWE空间。该方法既提高了两种单语嵌入空间的结构相似性，又提高了非监督映射方法中CLWEs的质量。

Bornea等人[15]，Barrire等人[66]和Aleksandr等人[62]将原英语语料库翻译成其他几种语言，获得 增强的数据。相应地，他们使用多语言模型。

Machine Translation

Advantage(s):

1. Easy to use.
2. Strong applicability.
3. This method ensures correct grammar and unchanged semantics.

Limitation(s):

1. Poor controllability and limited diversity because of the fixed machine translation models.

2.1.6 Model Generation

一些方法使用Seq2Seq模型直接生成复述。如果训练对象合适，这些模型会输出更多样化的句子，如图9所示。

Figure 9: Paraphrasing by model generation.

Hou等人[16]为基于任务的对话系统的语言理解模块提出了Seq2Seq数据增强模型。它们将去语言化的输入话语和指定的不同的序号k(如1、2、3)作为输入输入到Seq2Seq模型中，以生成新的话语。类似地，Hou等人[67]通过L层 Transformer对串联的多个输入话语进行编码。该模型利用重复感知注意和面向多样性的正则化来生成更多样化的句子。
在方面术语抽取（Aspect Term Extraction）方面，Li等[17]采用Transformer作为基本结构。利用 Masked原句及其标签序列训练模型M，将 Masked片段重建为增强数据。Kober等[68]使用GAN生成与原始数据非常相似的样本。Liu等人[18]使用一个预训练模型来共享问题嵌入和提出的基于transformer的模型的指导。该模型可以同时生成情境相关的可回答问题和不可回答问题。

Model Generation
Advantage(s):
1. Strong diversity.
2. Strong application.

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

2.2 Noising-based Methods

复述的重点是使增强数据的语义尽可能与原始数据相似。相比之下，基于噪声的方法增加了微弱的噪声，对语义影响不大，从而使其适当地偏离原始数据。人类可以通过对语言现象和先验知识的掌握，极大地减少微弱噪声对语义理解的影响，但这种噪声可能会给模型带来挑战。因

此，该方法不仅扩大了训练数据量，而且提高了模型的鲁棒性。

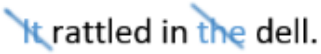
Methods	Examples	
	Original Data	Augmented Data
Swapping	It rumbled through the valley.	
Deletion	It rattled in the dell.	
Insertion	It pounded on the mountain.	
Substitution	It recoiled upon the flat.	
Mixup	Text: B_t^i, B_t^j Label: y^i, y^j	$\tilde{B}_t^{ij} = \lambda B_t^i + (1-\lambda) B_t^j$ $\tilde{y}^{ij} = \lambda y^i + (1-\lambda) y^j$

Figure 10: The example of five noising-based methods.

2.2.1 Swapping

自然语言的语义对文本顺序信息很敏感，而轻微的顺序变化对人类来说仍然是可读的[69]。因此，在合理的范围内，单词甚至句子之间的随机交换可以作为一种数据扩充方法。

Wei et al.[6]在句子中随机选择两个单词，并互换它们的位置。该工作重复该过程n次，其中n与句子长度l成正比。Longpre等人[60]，Rastogi等人[61]，Zhang等人[44]也采用相同的方法。Dai等[43]首先将token序列按照标签划分为若干段，然后随机选择一些段对其中的token顺序进行打乱，标签顺序不变。

除了字符级交换之外，一些作品还提出了实例级和句子级交换。在推文情感分析任务中，Luque等人[19]将推文分为两部分。他们随机抽样，并将有相同标签的前两半相结合。尽管以这种方式生成的数据可能不符合语法和语义，但与单个单词相比，它仍然具有相对完整的语义和情感极性。Yan等人[20]对法律文件分类进行句子级随机交换。由于句子本身所包含的语义

相对于词语而言较为完整，所以法律文书中的句子顺序对原文意义的影响不大。因此，作者对句子进行了洗牌，以获得扩充的文本。

2.2.2 Deletion

这种方法是指在句子中随机删除单词或在文档中删除句子。
Ces在文档中。在词级删除方面，Wei等[6]以p的概率随机删除句子中的每个词。Longpre等[60]，Rastogi等[61]，Zhang等[44]也采用了相同的方法。在 对话理解任务中，Peng等人通过删除槽值来增强输入对话行为，以获得更多的组合。
在句子级删除方面，Yan等[20]按照一定的概率随机删除法律文件中的每句话。他们这样做是因为有很多不相关的陈述存在，删除它们不会影响对法律案件的理解。Yu等人[22]对单词级和句子级的随机删除都采用了注意力机制。

2.2.3 Insertion

这种方法指在句子中随机插入单词或在文档中插入句子。
在词级插入方面，Wei等人[6]在一个句子中随机选择一个非停用词的同义词，然后将该同义词插入到句子中的任意位置。这项工作重复这个过程n次。在口语语言理解任务中，Peng等人通过插入槽值来增强输入对话行为，以获得更多的组合。
在法律文件分类中，由于相同标签的文件可能有相似的句子，Yan等[20]采用了句子级随机插入。他们从其他法律文件中随机选择相同标签的句子来获得增强数据。

2.2.4 Substitution

这种方法意味着用其他字符串随机替换单词或句子。与 前述复述方法不同，这种方法通常避免使用与原始数据语义相似的字符串。
一些工作通过现有的外部资源实现替换。Coulombe等人[7]和Regina等人[10]引入了英语中最常见的拼写错误的列表，以生成包含常见拼写错误的增强文本。例如，“across”很容易被拼成“accross”。Xie等[23]借鉴了“word-dropout”的思想，通过减少句子中的信息来提高泛化能力。该工作使用“_”作为占位符来替换随机单词，表示该位置的信息为空。Peng等[70]使用pseudo-IND平行语料库嵌入来创建字典和生成增强数据。
有些工作使用与任务相关的资源或生成随机字符串进行替换。Xie et al.[12]和Xie et al.[23]将原词替换为词表中的其他词，分别使用TF-IDF值和unigram频率从词表中选择单词。Lowell et al.[52]和Daval et al.[42]也探索了这种方法作为一种无监督数据增强方法。Wang等[71]提出了一种用词汇表中的其他单词随机替换输入和目标句子中的单词的方法。在NER中，Dai等人[43]用训练集中具有相同标签的随机 token替换原来的token。Qin等[72]提出了一种多语言语码转换方法，即用其他语言的单词替换源语言中的原始单词。在任务导向的对话任务中，随机替换是生成增强数据的一种有效方法。Peng等人[21]增强输入对话通过替代槽值来获得更多的口语理解组合。在插槽填充中，Louvan等[11]根据插槽标签进行插槽替换。Song等[73]通过复制用户话语来扩充对话状态跟踪的训练数据，并将相应的真实槽值替换为生成的随机字符串。

2.2.4 Mixup

Mixup的思想最早是由Zhang等人在图像领域提出的[74]。受此启发，Guo等人[24]提出了两种用于句子分类的Mixup变体。第一个算法wordMixup在词嵌入空间进行样本插值，第二个算法senMixup对句子编码器的隐藏状态进行插值。通过wordMixup和senMixup插值得到的新样本，它们的常见插值标签如下:

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

$$\tilde{B}_k^{ij} = \lambda f(B^i)_k + (1-\lambda) f(B^j)_k$$

$$\tilde{y}^{ij} = \lambda y^i + (1-\lambda) y^i$$

其中$B_i^t, B_j^t \in R^{N \times d}$ 表示原两个句子中的第t个单词，$f(B^i)$，$f(B^j)$ 表示隐藏层句子表示。
y^i、y^j 为对应的原始标签。
近年来，Mixup在许多工作中得到了广泛的应用。给定原始样本，Cheng等[25]首先按照[75]构建他们的对抗样本，然后应用两种Mixup策略:前者在对抗样本之间进行插值，后者在对应的两个原始样本之间进行插值。同样，Sun等[76]，Bari等[77]，Si等[78]都采用这种Mixup方法进行文本分类。Sun等人[76]提出将Mixup与基于 Transformer的预训练架构相结合的Mixup- transformer。他们在文本分类数据集上测试它的性能。Chen等人[79]将Mixup引入NER，提出了Intra-LADA和InterLADA。

- Mixup引入连续噪声而不是离散噪声，可以在不同标签之间产生增强数据。
- 与上述基于噪声的方法相比，该方法的可解释性较差，难度较大。

Noising
Advantage(s):
1. Noising-based methods improve model robustness.
2. Easy to use (in most cases).
1. Distorted syntax and semantics.
2. Limited diversity for every single method.

2.3 Sampling-based Methods

基于抽样的方法通过建模原始数据分布，在其中抽取新的数据点。与基于Paraphrasing的模型相似，它们也涉及规则和训练模型来生成增强数据。不同的是，基于抽样的方法是特定于任务的，需要任务信息，如标签和数据格式。这种方法不仅保证了增强数据的有效性，而且增加了多样性。这种方法基于人工启发式和训练模型更多的满足下游任务的需求，并可根据具体任

务需求进行设计。因此，这种方法通常比前两类更加灵活和困难。

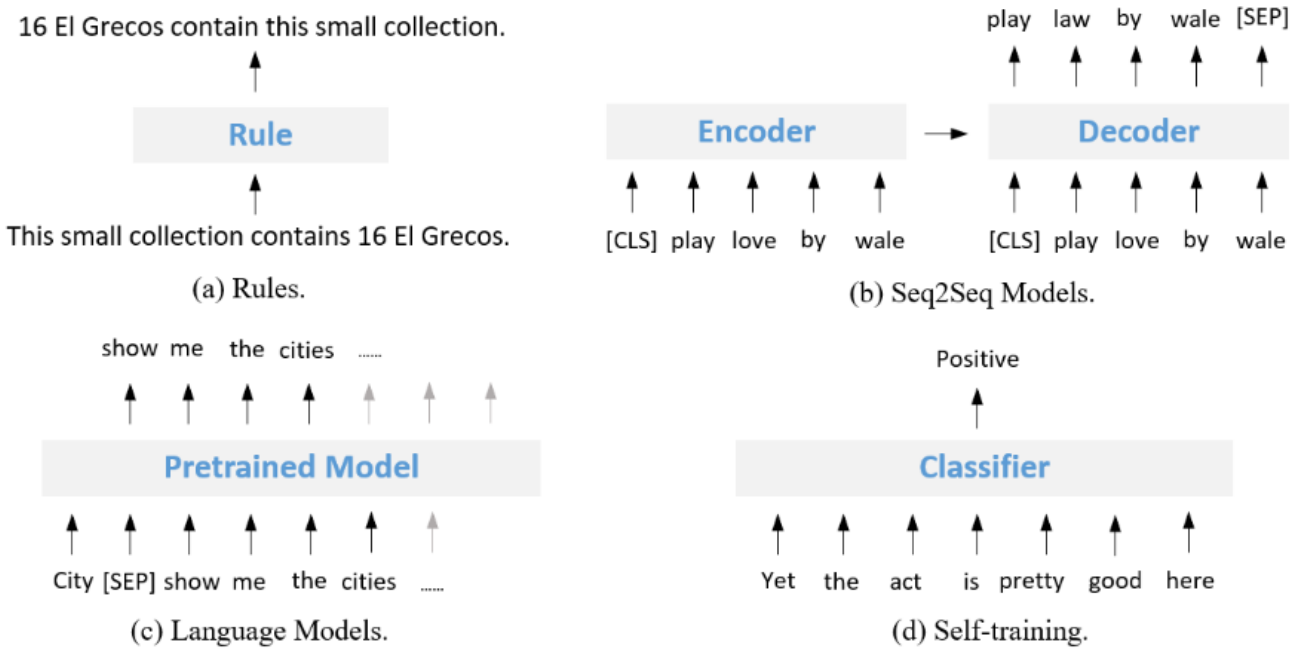


Figure 11: Sampling-based models.

2.3.1 Rules

该方法利用一些规则直接生成新的增强数据。为了保证增强数据的有效性，有时需要对自然语言和相应的标签进行启发式处理。模型结构如图11(a)所示。与上述基于规则的复述方法不同，该方法构造有效但不保证与原始数据相似(甚至会有不同的标签)。

Min等[26]交换资源句的主语和宾语，将谓语动词转换为被动形式。例如，例如“This small collection contains 16 El Grecos."。变成“16 El Grecos contain this small collection.”。新样本的标签是由规则决定的。Liu等人将数据增强方法应用于解决数学应用题(MWPs)的任务中。他们过滤掉了一些不相关的数字。然后基于双重检查的思想使用一些规则构建新的数据，例如，重用描述*time = distance/speed*的原始数据，构建描述*distance = time × speed*的扩充数据。该方法的输出方程是可以正确计算的。给定每段视频提供10个问答对的音视频场景感知对话（Audio-Video Scene-Aware Dialogue）训练集，牟某等[80]将前n对随机抽取作为对话历史，将第n + 1个问题作为需要回答的问题。在自然语言推理中，Kang等[28]利用外部资源如PPDB和人工启发式来构建新的句子。然后根据规则将新句与原句组合成扩充数据对，如A + B, B + C，则A + C。Kober等[68]用形容词-名词(AN)和名-名词(NN)复合词定义了一些规则来构建正负两对。例如，给定< car, car >，他们把< fastcar, car >作为一个积极的样本，< fastcar, redcar >作为负样本。Shakeel等人[81]通过反身性、对称性和及物性引申三种特性，构建了复述注释和非复述注释。Yin等[82]使用对称一致性和传递一致性两种规

则，以及逻辑引导的DA方法生成DA样本。

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

2.3.2. Seq2Seq Models

一些方法使用非预训练模型来生成增强数据。这种方法通常会引入反向翻译(BT)的思想，即训练一个目标到源的Seq2Seq模型，利用该模型从目标句子生成源句子，即构造伪平行句对[13]。这样的Seq2Seq模型学习目标和源分布之间的内部映射，如图11(b)所示。这与基于复述的模型生成方法不同，paraphrasing方法的扩充数据与原始数据具有相似的语义。
Sennrich等[84]利用已有的平行语料库训练一种英汉NMT模型，并利用目标英语单语语料库通过上述的英汉模型生成汉语语料库。Kang等人[28]为每个标签(蕴涵、矛盾和中性)训练Seq2Seq模型，然后使用给定句子的Seq2Seq模型生成新数据。Chen等[85]采用transformer架构，将“rewrite utterance → request utterance”映射视为机器翻译过程。此外，他们还利用 策略梯度技术来实施Seq2Seq生成的优化过程，以实现可控回报。Zhang等[13]使用Transformer作为编码器，将语法改错的知识迁移到风格转换任务上。Raille等人[29]创建了Edit-transformer，一个基于transformer的跨域模型。Yoo等人[86]提出了一种新的VAE模型，用于输出话语的语义槽序列和意图标签。

Seq2Seq Models
Advantage(s):
1. Strong diversity.
2. Strong application.
Limitation(s):
1. Require training data.
2. High training difficulty.

2.3.2. Language Models

近年来，预训练语言模型得到了广泛的应用，并被证明了预训练模型中包含知识。因此，它们自然地被用作增强工具，如图11(c)所示。
Tavor等人[30]提出了一种名为LAMBDA的数据增强方法。他们使用在训练集上进行了微调的GPT-2生成带标注的扩充句。然后用分类器对扩充句进行过滤以保证数据质量。Kumar等[31]采用了类似的方法，没有分类器进行过滤。
一些工作采用掩码语言模型来获取增强数据。Ng等人[32]使用掩码语言模型构造了一个 破坏

模型（corruption model）和一个重建模型（reconstruction model）。给定输入的数据点，它们最初生成的数据与破坏模型的原始数据流形相距甚远。然后利用重构模型将数据点拉回原始数据流形中作为最终的扩充数据。
一些作品采用自回归模型来获得增强数据。Peng等[21]使用预训练的SC-GPT和SC-GPT-NLU分别生成话语和对话行为。对结果进行筛选，以确保数据质量。Abonizio等人[87]对原始句子微调DistilBERT，生成伪造句子。特别是，GPT-2是一种用于生成增强数据的流行模型。Quteineh等人[34]使用标签条件化的GPT-2生成增强数据。Tarján等人[89]使用GPT-2生成增强数据，并将其重新分词为统计派生的子词，以避免在富形态语言中出现词汇爆炸。Zhang et al.[44]利用GPT-2在极端多标签分类中生成了大量多样化的增强数据。

Language Models
Advantage(s):
1. Strong application.
Limitation(s):
1. Require training data.

2.3.4 Self-training

在某些情况下，未标记的原始数据很容易获得。因此，将这些数据转换为有效数据将大大增加数据量，如图11(d)所示。
一些方法在黄金数据集上训练模型，以预测未标记数据的标签。Thakur等人首先对黄金数据进行BERT微调，然后使用微调后的BERT对未标记的句子对进行标记。这样的增强数据，以及黄金数据，被用来一起训练SBERT。Miao等[90]进一步将数据蒸馏引入到自我训练过程中。他们通过迭代更新教师模型来输出未标记数据的标签。Yang等[91]在问答任务中也采用了类似的自我训练方法;采用交叉注意力为基础的教师模型来确定每一对QA的标签。Du等人[35]引入了SentAugment数据增强方法，该方法从标注数据中计算特定于任务的query Embeddings，从而从从网络上抓取的数十亿个未标记句子中检索句子。
有些方法直接从其他任务中迁移已有的模型来生成伪平行语料库。Montella等人利用Wikipedia来利用大量的句子。然后他们使用Stanford OpenIE 包提取维基百科句子的三元组。例如，“Barack Obama was born in Hawaii.”，斯坦福OpenIE返回的三元组为< *BarackObama*; *was*; *born* > 和 < *BarackObama*; *wasbornin*; *Hawaii* >;这样的映射被翻转为RDF-to-text任务的扩充数据。Aleksandr等[62]采用了类似的方法。由于BERT在对象属性(OP)关系预测和对象可视性(OA)关系预测方面做得很好，Zhao等人[92]直接使用微调后的BERT来预测OP和OA样本的标签。

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

2.4 Analysis

如表1所示，我们从各个方面对上述DA方法进行了比较。

Table 1: Comparing a selection of DA methods by various aspects. *Learnable* denotes whether the methods involve model training; *online* and *offline* denote online learning and offline learning, respectively. *Ext.Know* refers 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
	Embeddings	-	✓	-	-	t	w, p
	MLMs	-	-	✓	-	t	w
	Rules	-	✓	-	-	t	w, p, s
	MT	-	-	-	-	t	s
	Seq2Seq	offline	-	-	✓	t	s
Noising	Swapping	-	-	-	-	t	w, p, s
	Deletion	-	-	-	-	t	w, p, s
	Insertion	-	✓	-	-	t	w, p, s
	Substitution	-	✓	-	-	t	w, p, s
	Mixup	online	-	-	✓	e, l	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

- 我们很容易发现，除了Seq2Seq和Mixup之外，几乎所有基于复述和噪声的方法都是不可习得的。然而，除了基于规则的方法外，大多数基于抽样的方法都是可学习的。可学习的方法通常比不可学习的方法更复杂，因此基于抽样的方法产生的数据比前两者更多样化和流畅。
- 在所有Learnable方法中，Mixup是唯一的在线方法。也就是说，增强数据的生成过程是独立于下游任务模型训练的。因此，Mixup是唯一一个从增强数据输出跨标签和离散嵌入的方法。
- 通过比较Learnable列和资源列，我们可以看到，大多数非可学习方法都需要外部知识资源，这超出了原始数据集和任务定义的范围。常用的资源包括语义词典(如WordNet和PPDB)、手工资源(如[7]中的拼写错误词典)以及人工启发式(如[26]和[28]中的启发式)。
- 结合前三列，我们可以看到，除了外部资源外，预训练或非预训练模型被广泛用作DA方法。这是因为预训练模型和训练对象中的知识在指导增强数据生成时起着类似于外部资源的作用。
- 对比Learnable一栏和与Task-related一栏，我们可以发现在复述和噪声这两类方法中，几乎所有的方法都与任务无关。他们可以只给出原始数据，而没有标签或任务定

义，从而生成增强数据。但是所有的抽样方法都是与任务相关的，因为它们采用启发式和模型训练来满足特定任务的需要。

- 比较level列和与Task-related列，我们可以看到它们是相关的。基于复述的方法位于文本级别。基于噪声的方法也是如此，除了Mixup，因为它对嵌入和标签都进行了更改。所有基于采样的方法都在文本和标签级别，因为标签也在增强过程中被考虑和构建。
- 对比Learnable列和Level列，我们发现几乎所有的非可学习方法都可以用于词级和短语级的数据增强，而所有可学习方法只能用于句子级的数据挖掘。虽然可学习的方法可以生成高质量的增强句，但遗憾的是，它们对文档的处理能力较弱，不能用于文档增强。因此，文献扩充仍然依赖于简单的非学习方法，这也是我们在研究中观察到的现状。

3. 策略和技巧（Strategies and Tricks）

上文介绍了三种DA方法，包括复述法、噪声法和抽样法，并分析了它们在各种NLP任务中的应用。在实际应用中，影响DA方法效果的因素很多。在这一章中，我们介绍了这些因素，以启发我们的读者选择和构建合适的DA方法。

3.1 Method Stacking

第2章中的方法不是 必须单独应用的。它们可以组合使用以获得更好的性能。常见的组合包括:

The Same Type of Methods

一些 工作结合不同的复述方法，获得了不同的复述，增加了扩充数据的丰富性。例如，Liu等人[49]同时使用词典和语义嵌入，Jiao等人[9]同时使用语义嵌入和 MLMs。对于基于噪声的方法，通常像[21]一样将之前的不可学方法结合使用。这是因为这些方法简单、有效、互补。一些方法也采用不同的噪声来源或复述，如[10]和[23]。不同资源的组合也可以提高模型的鲁棒性。

Unsupervised Methods

在某些场景下，简单且任务无关的无监督DA方法可以满足要求。自然地，它们被组合在一起并被广泛使用。Wei等人[93]提出了一种称为EDA的DA工具包，包括同义词替换、随机插入、随机交换和随机删除。EDA非常受欢迎，并用于许多任务([60,61])。Xie等人的UDA算法包括基于反向翻译和无监督噪声的方法;它也用于许多任务，如[42]。

Multi-granularity（多粒度）

有的工作在不同的层次上采用相同的方法，通过不同粒度的变化来丰富增强数据，提高模型的鲁棒性。例如，Wang等人[8]利用Word2Vec训练词嵌入和框架嵌入;Guo等人[24]在单词和句子层面应用Mixup, Yu等人[22]在单词和句子层面使用一系列基于噪声的方法。

3.2 Optimization

数据增强方法的优化过程直接影响到增强数据的质量。我们从增强数据的使用、超参数、训练策略和训练对象四个角度对其进行了介绍。

3.2.1 The Use of Augmented Data

增强数据的使用方式直接影响最终的效果。从数据质量的角度看，如果增强数据的质量不高，可以用增强数据对模型进行预训练;否则，它可以直接用于训练模型。从数据量来看，如果增

强后的数据量远远高于原始数据，通常不会直接将它们一起用于模型训练。相反，一些常见的做法包括(1)在训练模型之前对原始数据进行过采样 (2)用增强数据对模型进行预训练，并在原始数据上对其进行微调。

3.2.2 Hyperparameters

上述方法均涉及超参数，超参数对增强效果影响较大。我们在图12中列出了一些常见的超参数:

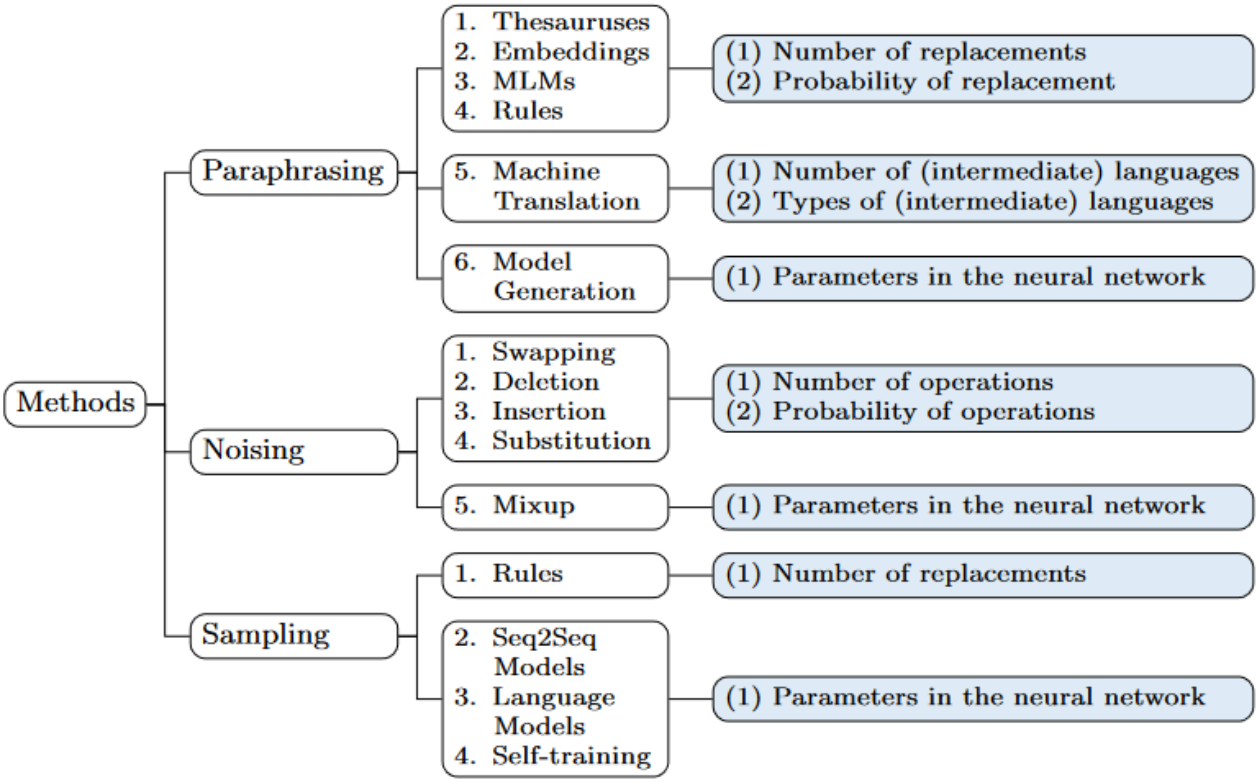


Figure 12: Hyperparameters that affect the augmentation effect in each DA method.

3.2.2 Training Strategies

一些工作运用了基于基本数据增强方法的训练策略。例如，Qu等[65]将反向翻译与对抗性训练相结合。类似地，Quteineh et al.[34]将基本的预训练模型转化为优化问题，以最大化生成的输出的有用性。Hu等人[94]和Liu等人[95]使用预训练的语言模型来生成增强数据，并将这个过程迁移到强化学习。一些工作([61,96])采用生成对抗网络(Generative Adversarial Networks)的思想来生成具有挑战性的增强数据。

3.2.2 Training Objects

训练对象是模型训练的关键，特别是对于可学习的数据 增强方法。Nugent等人[64]提出了一系列softmax温度系数设置，以确保多样性，同时保留语义。Hou等[67]利用重复感知注意力和多元导向正则化生成更多样化的句子。Cheng等人[25]使用课程学习来鼓励模型关注困难的训练例子。

3.2 Filtering

有时数据增强过程中不可避免地会引入一些噪声甚至误差，因此引入了过滤机制来避免这一问题。
有些工作在初始阶段会对部分输入数据进行过滤，以避免输入不当影响增强效果。一个典型的例子是句子长度，即过滤掉太短的句子([17])。Liu等人[27]在解决数学应用题时过滤掉不相关的数字，以确保生成的数据在计算上是正确的。
此外，一些工作在最后阶段对生成的增强数据进行了筛选。这通常是通过模型实现的。例如，Zhang等人[13]使用了一个鉴别器来过滤反向翻译结果。Tavor等[30]和Peng等[21]都使用分类器对预训练模型生成的增强句进行过滤，以保证数据质量。

4. Applications on NLP Tasks（在NLP任务中的应用）

虽然近年来在自然语言处理领域出现了多种数据增强方法，但很难直接比较它们的性能。这是因为不同的任务、评估指标、数据集、模型架构和实验设置使得直接比较变得毫无意义。因此，在上述工作的基础上，我们从文本分类、文本生成、结构化预测等不同的NLP任务的角度分析数据增强方法[97]。

- 文本分类是最简单、最基本的自然语言处理问题。也就是说，对于一段文本输入，输出文本所属的类别，其中的类别是预定义的封闭集。
- 文本生成，顾名思义，就是根据输入数据生成相应的文本。最经典的例子就是机器翻译。
- 结构化预测问题通常是NLP所特有的。与文本分类不同，结构化预测问题中输出类别之间存在较强的相关性和格式要求。
  相对于其他的自然语言处理任务，DA方法在文本分类中的应用更为广泛。此外，每一种独立的DA方法都可以用于文本分类。这种应用优势在于文本分类形式简单：给定输入文本，通过标签预测直接考察模型对语义的理解。因此，对于数据增强来说，只考虑保留对分类重要的词的语义是相对简单的。
  在文本生成方面，它更倾向于基于抽样的方法，以带来更多的语义多样性。而结构化预测更喜欢基于复述的方法，因为它对数据格式很敏感。因此，对数据的有效性提出了更高的要求。
  通过比较每种DA方法，我们可以看到，简单有效的无监督方法，包括机器翻译、基于辞典的复述和随机替换，是非常受欢迎的。此外，可学习的方法，如Seq2Seq转述模型、预训练模型和自我训练，也因其多样性和有效性而获得了很多关注。
  我们还通过时间轴展示了DA方法在三种类型任务上的发展过程。从整体上看，DA在这些任务上的应用数量逐年增加。文本分类是使用DA的第一个任务，对应的论文数量也多于其他两个任务。在文本生成和结构化预测方面，DA越来越受到关注。基于复述的方法一直是一种流行的方法，近年来基于抽样的方法在文本分类和文本生成中也被证明是有效的，但人们在结构化预测任务中仍然倾向于使用基于复述和噪声的方法。

5. Related Topics

数据增强与其他学习方法有什么关系?在本节中，我们将数据增强与其他类似主题联系起来。

5.1 Pretrained Language Models

大多数预训练语言模型(PLMs)的训练都是基于自监督学习的。自监督学习主要是利用辅助任务从大规模的无监督数据中挖掘自己的监督信息，并通过这些构建的监督信息对网络进行训练，从而为下游任务学习有价值的表示。从这个角度来看，PLMs还以隐性的方式将更多的训练数据引入到下游任务中。另一方面，PLMs中一般的大规模无监督数据对于特定的任务可能是域外的。不同的是，与任务相关的数据增强方法本质上专注于特定的任务。

5.2 Contrastive Learning

对比学习是学习一个嵌入空间，在这个嵌入空间中，相似的样本离得很近，而不同的样本离得很远。它侧重于学习相似样本之间的共同特征，区分不同样本之间的差异。对比学习的第一步是利用数据增强构建具有相同标签的相似样本，第二步是随机选择实例作为负样本。因此，对比学习是数据增强的应用之一。

5.3. Other Data Manipulation Methods

除了DA，还有一些其他的数据操作方法可以提高模型的泛化[117,94]。***Oversampling***通常用于数据不平衡的场景。它只是简单地从少数群体中抽取原始数据作为新样本，而不是生成增强数据。***Data cleaning*** 对原始数据再进行数据清洗，提高数据质量，降低数据噪声。它通常包括小写、词干化、词根化等。***Data weighting***在训练时，根据不同样本的重要程度，对其分配不同的权重，不产生新的数据。***Data synthesis***提供完整的人工标注示例，而不是由模型或规则生成的扩充数据。

5.4 Generative Adversarial Networks

生成对抗网络(GANs)是由Goodfellow等人首先提出的[118]。作为一种半监督方法，GANs包括生成模型，主要用于挑战GANs的鉴别器，而某些DA方法中的生成模型直接用于增强训练数据。此外，GANS的生成模型在[61,119,96,68,109,116]等场景中被用作数据增强方法，并被证明是有效的数据增强方法。

5.5 Adversarial Attacks

对抗性攻击是指生成攻击机器学习模型的对抗性例子，即导致模型出错的技术。一些工作使用DA方法，如code-switch 替换，以生成对抗的例子作为一致性正则化。

6. Challenges and Opportunities

数据增强在过去的几年里有了很大的发展，为大规模模型训练和下游任务的开发提供了很大的贡献。尽管有这个过程，仍有一些挑战需要解决。在本节中，我们将讨论其中的一些挑战和未来可能有助于推进该领域的方向。

Theoretical Narrative

现阶段，对自然语言处理中的数据增强方法还缺乏系统的探索工作和理论分析。以往的研究大多提出了新的方法或证明了DA方法对下游任务的有效性，但没有从数学的角度探究DA方法背后的原因和规律。自然语言的离散性使得理论叙述至关重要，因为叙述可以帮助我们理解DA的本质，而不局限于通过实验来确定有效性。

More Exploration on Pretrained Language Models

近年来，预训练语言模型在自然语言处理中得到了广泛的应用，它通过对大规模语料库的自监

督学习，包含了丰富的知识。有一些工作使用预先训练好的语言模型进行DA，但大多局限于[MASK]的完成，微调或者自训练后直接生成。DA在预训练语言模型的时代仍然有用吗?或者，如何进一步利用预训练模型中的信息，以更低的成本生成更多样化和高质量的数据?以上是值得考虑的方向。

More Generalized Methods for NLP

自然语言与图像或声音的最大不同在于它的表征是离散的。与此同时，NLP还包括其他模式无法实现的特定任务，如结构化预测。因此，目前还没有一种DA方法可以适用于所有的NLP任务，这与一般的图像增强的裁剪方法或音频增强的速度扰动方法不同。这意味着不同NLP任务之间的DA方法仍然存在差距。随着预训练模型的发展，这似乎有一些可能性。特别是T5[121]和GPT3[122]的提出，以及prompting learning的出现，进一步验证了自然语言中任务的形式化可以独立于传统的类别，通过统一任务定义可以得到一个更一般化的模型。

Working with Long Texts and Low Resources Languages

现有的方法在短文本和通用语言方面取得了重大进展。然而,限于模型功能,DA方法在长文本仍使用基于复述、噪音的简单方法(如表1所示)。与此同时,受限于数据资源, ,尽管他们对数据增强有更多的需求，低资源语言上的增强方法很是稀缺。显然，在这两个方向上的探索仍然有限，但它们可能是有前景的方向。

7. Conclusion

在本文中，我们对自然语言处理中的数据增强进行了全面和结构化的研究。为了检验数据增强的本质，我们根据扩充数据的多样性将数据分析方法分为三类，包括复述、噪声和抽样。这样的分类可以帮助理解和发展DA方法。我们还介绍了DA方法在NLP任务中的应用，并通过时间轴对其进行了分析。此外，我们还介绍了一些技巧和策略，供研究者和实践者参考，以获得更好的模型性能。最后，我们将DA与一些相关主题区分开来，并概述了当前的挑战和未来研究的机会。