Ta文档级机器翻译的目标端增强

鲍广胜[1,2]**,滕志扬**[3]**,和 YueZhang**[*,2,4]**,**[1] 浙江大学[2] 西湖大学工程学院[3] 南洋理工大学[4] 先进技术研究院，西湖高等研究院[2]
{baoguangsheng, zhangyue}@westlake.edu.cn[3]
zhiyang.teng@ntu.edu.sg

arXiv:2305.04505v2 [cs.CL] 4 Jun 2023

摘要

文档级机器翻译面临数据稀疏性的挑战，这是由于输入长度较长且训练数据量较少，增加了学习虚假模式的风险。为应对这一挑战，我们提出了一种目标端增强方法，引入一个数据增强（DA）模型，为每个源文档生成许多潜在译文。通过在范围更广的译文上进行学习，机器翻译模型可以学习到平滑分布，从而降低数据稀疏性的风险。我们证明，能够估计后验分布的 DA 模型显著提升了机器翻译性能，在 News 上以 2.30 s-BLEU 的优势超过此前最佳系统，并在 News 和 Europarl 基准测试上达到新的最先进水平。我们的代码可在
https://github.com/baoguangsheng/target-side-augmentation.

1 引言

文档级机器翻译（Gong 等人，2011；Hardmeier 等人，2013；Werlen 等人，2018；Maruf 等人，2019；Bao 等人，2021；Feng 等人，2022）日益受到研究者关注。它通过考虑跨句共指和话语信息，弥补了句子级机器翻译的局限性，因此在实际应用中更具实用价值。与句子级机器翻译相比，文档级机器翻译面临若干独特的技术挑战，包括输入长度显著增加（Bao 等人，2021），以及训练数据相对较少（Junczys-Dowmunt，2019；Liu 等人，2020；Sun 等，2022）。这些挑战相互叠加，导致数据稀疏性加剧（Gao 等人，2014；Koehn 和 Knowles，2017；Liu 等人，2020），从而增加了模型从训练数据中学到虚假模式的风险（Belkin 等人，2019；Savoldi 等人，2021），并阻碍泛化（Li 等人，2021；Dankers 等人，2022）。

为解决这些问题，我们提出一种目标端数据增强方法，旨在通过自动平滑训练分布来减少稀疏性。其核心思想是让文档机器翻译模型使用大量合理的潜在翻译进行训练，而不是强迫模型针对每个源文档拟合单一的人工翻译。这使模型能够学习更稳健、更具泛化能力的模式，而非过度依赖特定训练样本的特征。具体而言，我们引入一个数据增强（DA）模型，生成可能的翻译，以指导机器翻译模型的训练。如图 1所示，DA模型先基于一个观测到的翻译学习源文本与可能译文之间的关系（步骤1），再用于采样一组可能合理的译文（步骤2）。这些译文将输入机器翻译模型进行训练，从而平滑目标翻译的分布（步骤3）。

我们使用标准的文档级机器翻译模型，包括 Transformer（Vaswani 等人，2017年）和 G-Transformer（Bao 等人，2021年），分别作为我们的数据增强模型和机器翻译模型。对于数据增强模型，为了有效捕获给定参考目标的后验目标分布，我们将每个源句子与一个潜在词元序列拼接，作为新的输入，其中潜在词元从观测到的翻译中采样得到。数据增强模型面临的一个挑战是，将参考译文作为输入可能会降低多样性。为了解决这一问题，我们在编码器端引入中间潜变量，通过规则生成 n 元语法样本，从而利用后验采样（Wang 和 Park，2020年）生成多样化的翻译结果。

在三个文档级机器翻译基准上的结果表明，我们的方法显著优于 Transformer 和 G-Transformer 基线，平均分别提升了 1.33 和 1.75 个 s-BLEU，并在 News 和 Europarl 上取得了最先进的结果。

** 通讯作者。

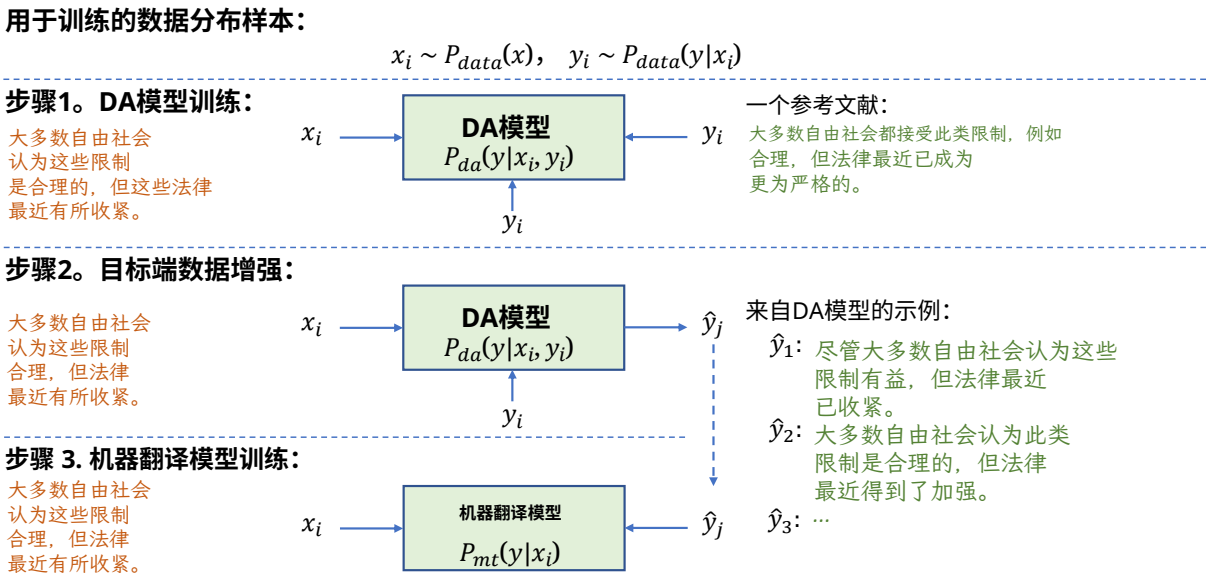


图 1：通过一个非常简单的示例说明目标端数据增强（DA）。训练 DA 模型以估计给定源文本 x_i 和观测到的目标 y_i 时可能翻译结果 y 的分布，并针对每个源文本 x_i，使用从 DA 模型中采样得到的翻译结果 $\hat{y}_j$ 训练机器翻译模型。直接使用目标 y_i（它也是一个条件输入）训练 DA 模型可能颇具挑战，但在翻译结果 y 与条件 y_i 之间引入一个中间潜变量后，便可以实现。

进一步分析表明，生成译文的高度多样性及其与人工参考译文之间的较小偏差，是性能提升的关键。据我们所知，我们首次进行目标端增强，以丰富文档级机器翻译的输出 多样性。

2 相关工作

数据增强（DA）通过合成新数据来增加训练数据（Van Dyk 和 Meng，2001；Shorten 和 Khoshgoftaar，2019；Shorten 等人，2021年；Li 等人，2022年）。在神经机器翻译（NMT）中，最常用的数据增强技术是**源端增强**，包括简单数据增强（EDA）（Wei 和 Zou，2019）、子词正则化（Kudo，2018）和回译（Sennrich 等人，2016a），回译为单语目标生成伪源文本，使得广泛可用的单语数据得以利用。这些方法会针对同一个金标准目标翻译，生成更多源-目标语对，并为其配上不同的银标准源句。相比之下，**目标端增强**更具挑战性，因为 EDA 等方法不适用于目标端：它们会破坏目标序列，进而损害目标语言的自回归建模。

以往针对目标端数据增强的方法机器翻译中的数据增强可分为三类。第一类基于自训练（Bogoychev and Sennrich，2019；He et al.，2019；Zoph et al.，2020），使用训练好的模型为单语源文本生成伪翻译。第二类使用预训练语言模型（Fadaee et al.，2017；Wu et al.，2019）或预训练生成模型（Raffel et al.，2020；Khayrallah et al.，2020）为目标文本中的词生成同义词或释义文本。第三类依赖强化学习（Norouzi et al.，2016；Wang et al.，2018），引入奖励函数来评估候选译文的质量，并对似然目标函数进行正则化。为了探索可能的候选文本，通常从模型分布中采样或使用随机噪声。不同于这些方法，我们的方法是一种目标端数据增强技术，通过监督学习进行训练，不依赖外部数据或大规模预训练。更重要的是，我们生成的是文档级而非词级、短语级或句子级替代文本。

以往的目标端输入增强（Xie 等人，2022年）似乎与我们的目标端增强相似。然而，除字面上的相似之外，两者截然不同。考虑词元预测 $P(y_i|x, y_{<i})$。目标端输入增强会增强条件 $y_{<i}$，以提升模型对这些条件的鲁棒性，

这更像是在条件 x 下进行源端增强。相比之下，目标端增强会增强目标文本 y_i，为模型提供全新的训练目标文本。

释义模型。我们的方法会为每个源文本生成多种翻译，每种翻译都可以视为目标的释义。与以往利用释义模型改进机器翻译的方法不同（Madnani 等人，2007；Hu 等人，2019；Khayrallah 等人，2020年），我们的数据增强模型利用平行语料库，不依赖外部释义数据，这与 Thompson 和 Post (2020年)的方法类似。相反，该模型在建模目标分布时会考虑源文本。更重要的是，大多数释义模型都在句子级运行，而我们的数据增强模型可以生成文档级翻译。

条件自编码器。数据增强模型也可以看作条件去噪自编码器（c-DAE），其中潜变量是真实目标的带噪版本，模型经过训练，可从带噪潜序列中重建真实目标。c-DAE 与条件变分自编码器（c-VAE）（Zhang 等人，2016；Pagnoni 等人，2018）类似，后者学习潜变量，并通过从中采样生成多样化的翻译。不过，c-VAE 与我们的数据增强模型有两个关键区别。首先，c-VAE 会学习潜变量的先验分布和后验分布，而数据增强模型则直接使用预定义规则生成潜变量。其次，c-VAE 对目标的先验分布建模，而数据增强模型估计后验分布。

序列级知识蒸馏。我们的数据增强-机器翻译流程在形式上也与序列级知识蒸馏（SKD）（Ba 和 Caruana，2014；Hinton 等人；Gou 等人，2021年；Kim 和 Rush，2016；Gordon 和 Duh，2019；Lin 等人，2020年）有些相似。序列级知识蒸馏通过使用大型教师模型学习数据分布，再利用教师模型生成的序列训练小型学生模型，从而将知识蒸馏给学生模型。然而，我们的方法与序列级知识蒸馏有三点不同。首先，序列级知识蒸馏旨在将知识从大型教师模型压缩到小型学生模型，而我们使用与数据增强模型同等规模或更小的模型，知识来源是训练数据，而非大型教师模型。其次，序列级知识蒸馏中的教师模型根据源文本估计目标的先验分布，而我们的数据增强模型估计的是给定源文本和一个已观测目标时目标的后验分布。第三，序列级知识蒸馏针对每个源文本生成一个序列，而我们使用受控的潜变量生成多个不同的翻译。

3 目标端增强

整体框架如图1所示。形式化地，将一组训练数据表示为 $D = \{(x_i, y_i)\}_{i=1}^{N}$，其中$(x_i, y_i)$ 是第 i个源文本-目标文本对，N 是配对数量。我们训练一个数据增强（DA）模型（第3.1节），以生成包含新目标翻译的样本（第3.2节），并用这些样本训练机器翻译（MT）模型（第3.3节）。

3.1 数据增强模型

我们通过引入潜变量，从平行语料库中学习后验分布 $P_{da}(y|x_i, y_i)$

$$P_{da}(y|x_i, y_i) = \sum_{z \in \mathcal{Z}_i} P_\varphi(y|x_i, z) P_\alpha(z|y_i), \quad (1)$$

其中，z 是控制翻译输出的潜变量，$\mathcal{Z}_i$ 表示 z 的可能空间，φ 表示DA模型的参数，α表示用于确定给定 y_i时 z 分布的超参数。

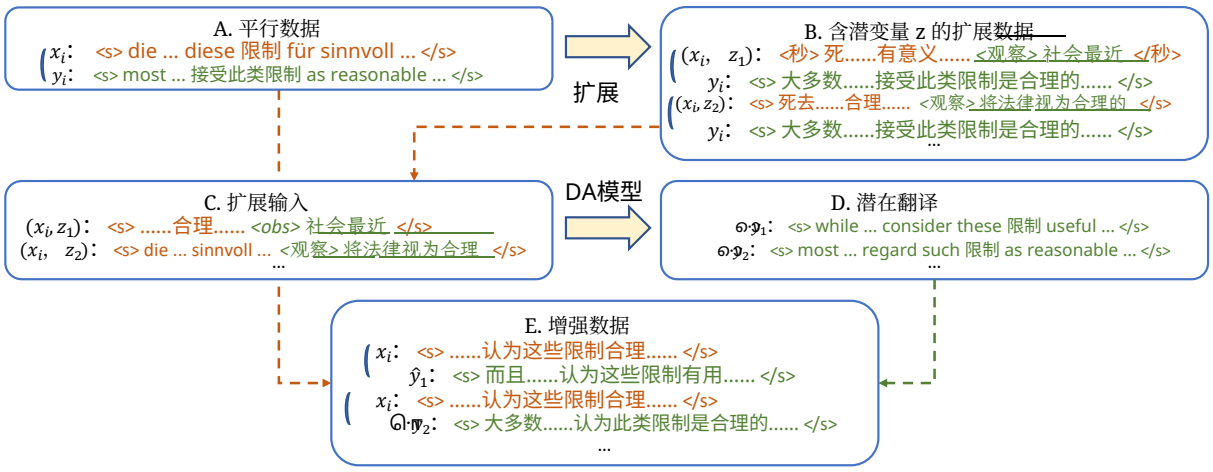
的可能 z 空间相较于目标文本的词元数量呈指数级增长，因此无法对 $\mathcal{Z}_i$ 求和。我们因此采用蒙特卡罗近似，从 $p_\alpha(z|y_i)$中抽取一组实例，并计算样本均值

$$P_{da}(y|x_i, y_i) \approx \frac{1}{|\hat{\mathcal{Z}}_i|} \sum_{z \in \hat{\mathcal{Z}}_i} P_\varphi(y|x_i, z), \quad (2)$$

ˆwhere $\mathcal{Z}_i$ 表示采样得到的实例。

潜变量有多种可能的选择，例如连续向量或类别型离散变量，这些变量既可以由模型学习得到，也可以通过预定义规则设定。在这里，我们简单地将潜变量表示为一系列词元，并使用预定义规则生成该序列，从而使潜变量能够轻松融入序列到序列模型的输入中，而无需额外的参数。

具体而言，我们将潜变量 z 的值设为从观测到的翻译 y_i 中采样的一组 n 元语法，并将 x_i 和 z 拼接成一个词元序列。我们假设生成的译文 y 能够与



Fig图 2：详细的数据增强流程，其中平行数据会被增强多次。

观测到的翻译 y_i 在这些 n 元语法上保持一致。为此，我们将 α 定义为 y_i 中可通过 z 观测到的词元比例，并将其命名为 观测比例。对于含有 $|y_i|$ 个词元的目标，我们从 y_i 中均匀采样 n 元语法，以覆盖 $\alpha \times |y_i|$ 个词元，其中每个 n 元语法的长度随机取自 $\{1, 2, 3\}$。例如，给定 $\alpha = 0.1$ 和一个含有 20 个词元的目标 y_i，我们可以从目标中采样一个二元语法或两个一元语法，从而得到 2（0.1×20）个词元。

训练。给定一个样本（x_i，y_i），训练损失可重写为

$$\begin{aligned} \mathcal{L}_{da} &= -\sum_{i=1}^{N} \log P_{da}(y = y_i | x_i, y_i) \\ &\approx -\sum_{i=1}^{N} \log \frac{1}{|\hat{\mathcal{Z}}_i|} \sum_{z \in \hat{\mathcal{Z}}_i} P_\varphi(y = y_i | x_i, z) \\ &\leq -\sum_{i=1}^{N} \frac{1}{|\hat{\mathcal{Z}}_i|} \sum_{z \in \hat{\mathcal{Z}}_i} \log P_\varphi(y = y_i | x_i, z), \end{aligned} \tag{3}$$

其中，损失的上界由詹森不等式给出。该上界对数概率求和，可以视为对每个（x_i，z，y_i）的标准负对数似然（NLL）损失求和。因此，当我们将优化该上界作为优化 $\mathcal{L}_{da}$的替代方案时，DA模型使用标准NLL损失进行训练，但训练样本数量增加了 $|\mathcal{Z}_i|$ 倍。

讨论。如图1所示，给定一个样本 (x_i, y_i)，我们采用一种新的估计方法，利用后验分布 $P_{da}(y|x_i, y_i)$ 来构建我们的数据增强模型。其基本直觉是，通过同时以源文本 x_i 和观测到的翻译 y_i为条件，数据增强模型能够比机器翻译模型更准确地估计数据分布 $P_{data}(y|x_i)$。从逻辑上讲，机器翻译模型学习的是一个先验分布 $P_{mt}(y|x_i)$，用于估计数据分布 $P_{data}(y|x_i)$ 以建模翻译概率。当语料库较大时，这种先验分布效果良好。然而，当语料库相较于数据空间较为稀疏时，学习到的分布会过拟合稀疏分布的样本，导致对未见目标文本的泛化能力较差。

3.2 数据增强流程

数据增强的详细流程如图2 所示，相应的算法如算法1所示。下面我们用一个训练示例来说明。

DA模型训练。我们将潜变量 z 表示为一个词元序列，并将其与源文本z 拼接起来，因此可以使用通用的序列到序列模型来建模后验分布。与通用机器翻译模型相比，唯一的区别在于输入结构。

具体而言，如图中步骤 B 所示，对于平行数据中的一个给定样本（x_i，y_i），我们从 y_i 中采样若干个 n 元语法，并将输入扩展为（x_i，z），采样数量根据 y_i 的长度确定。以目标句子“大多数自由社会认为此类限制是合理的，但法律最近变得更加严格。”为例。我们从目标端采样“社会”和“最近变得”，并将它们拼接到源句子末尾，形成第一个输入序列。接着，我们采样“法律”和“认为合理”，形成第二个输入序列。这些新的输入序列与原始目标序列配对，形成新的平行数据。通过生成不同的输入序列，我们多次扩充数据。

算法 1目标端数据增强。

输入: $D = \{(x_i, y_i)\}_{i=1}^{N}$ ▷A. 平行数据
输出: $D' = \{(x_i, y_i)\}_{i=1}^{N\times(M+1)}$ ▷ 增强 M 次

函数目标增强(D)
  $D' \leftarrow \{\}$
  对 $i \leftarrow 1$ 到 N **执行**
    $(x_i, y_i) \leftarrow D[i]$ ▷对于每个样本
    $D' \leftarrow D' \cup \{(x_i, y_i)\}$ ▷添加金标准对
    对 $j \leftarrow 1$ 到 M **执行**
      $\alpha \sim Beta(a, b)$ ▷ 抽取一个观测比例
      $z_j \sim P_\alpha(z|y_i)$ ▷ 采样一个潜变量值
      $\hat{y}_j \sim P_\varphi(y|x_i, z_j)$ ▷ 抽样生成一个翻译
      $D' \leftarrow D' \cup \{(x_i, \hat{y}_j)\}$ ▷ 添加新对
  返回 D' ▷E. 增强数据

目标端数据增强。 使用步骤 B 中从扩展数据里分离出的"C. 扩展输入"数据，我们通过使用训练好的 DA模型进行束搜索来生成新的翻译；对于每个扩展输入序列，我们都会得到一个新的翻译。这里，我们重复使用步骤 B 中采样的 z。不过，我们也可以为推理采样新的z，这对机器翻译性能没有明显影响。将新的翻译与原始源序列配对后，我们便得到"E. 增强数据"。具体细节见算法 1，该算法以原始平行数据为输入，并输出增强数据。

3.3 机器翻译模型

我们使用 Transformer（Vaswani 等人，2021年）和 G-Transformer（Bao 等人，2021年）作为基线机器翻译模型。Transformer基线对句子级翻译进行建模，并逐句翻译文档；而 G-Transformer 对整个文档翻译进行建模，直接将源文档翻译成相应的目标文档。G-Transformer 通过组注意力（附录 A）改进了 Transformer 中朴素的自注意力机制，以进行长文档建模，是近期最先进的文档机器翻译模型。

基线训练。 基线方法使用原始训练数据集 D 通过标准负对数似然损失进行训练

$$\mathcal{L}_{mt} = -\sum_{i=1}^{N} \log P_{mt}(y = y_i|x_i). \quad (4)$$

增强训练。 对于我们的目标端增强方法，我们强制机器翻译模型匹配由

数据集	句子 训练/开发/测试	文档 训练/开发/测试
TED	0.21M/9K/2.3K	1.7K/92/22
News	0.24M/2K/3K	6K/80/154
Europarl	1.67M/3.6K/5.1K	118K/239/359

表 1：数据集统计信息。

DA模型

$$\mathcal{L}_{mt} = -\sum_{i=1}^{N} \sum_{y\in\mathcal{Y}_i} P_{da}(y|x_i, y_i) \log P_{mt}(y|x_i), \quad (5)$$

其中，$\mathcal{Y}_i$ 是 x_i 的可能翻译。

我们使用蒙特卡洛方法近似对 $\mathcal{Y}_i$ 的期望。具体而言，对于每个样本（x_i，y_i），我们首先从 $P_\alpha(z|y_i)$ 中采样 z_j，然后将 x_i 和 z_j 作为 DA模型的输入运行束搜索，从而获得一个可行的翻译。重复该过程 M 次，我们便得到一组可能的翻译。

$$\hat{\mathcal{Y}}_i = \{\arg\max_y P_\varphi(y|x_i, z_j)|z_j \sim P_\alpha(z|y_i)\}_{j=1}^{M}, \quad (6)$$

如图中的步骤 D 所示2 以及第1 节中的算法3.2 所示。

随后，机器翻译模型的损失函数重写如下，其中使用采样翻译的平均负对数似然损失近似该期望。

$$\mathcal{L}_{mt} \approx -\sum_{i=1}^{N} \frac{1}{|\hat{\mathcal{Y}}_i|} \sum_{y\in\hat{\mathcal{Y}}_i} \log P_\theta(y|x_i), \quad (7)$$

其中，θ 表示机器翻译模型的参数。ˆ数字 $|\mathcal{Y}_i|$ 可能因样本而异，但为简便起见，我们在实验中选择固定的数字 M。

4 实验

数据集。我们在三个基准数据集上进行实验——TED、News 和 Europarl（Maruf 等人，2019），这些数据集代表了英德（En-De）翻译中的不同领域和数据规模。详细统计数据见表 1，详细描述见附录 B.1。

评估指标。我们遵循Liu 等人（2020年）的做法，使用句子级BLEU分数（s-BLEU）和文档级BLEU分数（d-BLEU）作为主要性能评估指标。此外，我们还定义了偏差和多样性这两个指标，用于衡量生成的译文的质量，这些译文由

方法	TED s-BLEU	TED d-BLEU	News s-BLEU	News d-BLEU	Europarl s-BLEU	Europarl d-BLEU	平均值 s-BLEU
层次注意力网络 (Miculicich 等人，2018)	24.58	-	25.03	-	28.60	-	26.07
SAN (Maruf 等人，2019)	24.42	-	24.84	-	29.75	-	26.34
混合上下文 (Zheng 等人，2020年)	25.10	-	24.91	-	30.40	-	26.80
Flat-Transformer (Ma 等人，2020年)	24.87	-	23.55	-	30.09	-	26.17
G-Transformer（随机） (Bao 等人，2021年)	23.53	25.84	23.55	25.23	32.18	33.87	26.42
G-Transformer（fnt.） (Bao 等人，2021年)	25.12	27.17	25.52	27.11	32.39	34.08	27.68
MultiResolution(Sun 等人, 2022年)	25.24	29.27	25.00	26.71	32.11	34.48	27.45
RecurrentMem（Feng 等人，2022年）	25.62	**29.47**	25.73	27.78	31.41	33.50	27.59
SMDT（Zhang 等人，2022年）	25.12	-	25.76	-	32.42	-	27.77
Transformer（已发送基线） ♢	24.91	-	24.82	-	31.22	-	26.98
+ 目标端数据增强（我们的方法）	26.14*	-	27.03*	-	31.75*	-	28.31
G-Transformer（fnt.）（文档基线） ♢	25.20	27.94	25.12	27.02	31.93	33.88	27.42
+ 目标端增强（我们的方案）	**26.59***	29.20*	28.06*	29.83*	**32.85***	**34.76***	**29.17**
Transformer + 回译（已发送） ♡	25.03	-	26.07	-	31.12	-	27.41
目标端增强（我们的方法）	26.13	-	28.01	-	31.27	-	28.47
G-Transformer + 回译（文档） ♡	25.45	28.06	26.25	28.21	32.00	33.94	27.90
目标端增强（我们的方法）	26.21	28.58	**28.69**	**30.41**	32.52	34.50	29.14
用于比较的预训练设置							
Flat-Transformer+BERT（Ma 等人，2020年）	26.61	-	24.52	-	31.99	-	27.71
G-Transformer+BERT（Bao 等人，2021年）	26.81	-	26.14	-	32.46	-	28.47
G-Transformer+mBART（Bao 等人，2021年）	28.06	30.03	30.34	31.71	32.74	34.31	30.38

表 2：英语到德语文档级翻译的主要结果，其中“*”表示相较于使用 $p < 0.01$ 的基线有显著提升。（rnd.）——参数随机初始化。（fnt.）——参数使用训练好的句子级模型初始化。 ♢ ——我们针对增强数据集调整超参数。 ♡——我们通过将每个目标文本回译为新的源文本来增强训练数据，而非引入额外的单语目标文本。

数据增强模型生成。详细说明和定义见附录B.2。

基线。我们将目标端增强应用于两个基线，包括句子级 Transformer（Vaswani 等人，2017）和文档级 G-transformer（Bao 等人，2021年）。我们还结合回译与目标端增强，并将其应用于这两个基线。

训练设置。对于 Transformer 和 G-Transformer，我们为每个句子生成 M 个新译文（TED 和 News 各 9 个，Europarl 为 3 个），并将数据扩充至原来的 $M+1$ 倍。对于回译基线，由于训练数据已翻倍，我们进一步将 TED 和 News 的数据扩充 4 倍，将 Europarl 的数据扩充 1 倍，使 TED 和 News 的总扩充倍数仍为 10，Europarl 为 4。

我们通过从z 贝塔分布中采样得到一个具有观测比例的潜变量$Beta(2,3)$ 来获得翻译结果，并使用束宽为 5 的束搜索。每项主要实验均运行三次，并报告中位数。更多详情请参见附录B.3。

4.1 主要结果

如表 2所示，目标端增强显著提升了所有基线的表现。具体而言，它使 G-Transformer（fnt.）的 s-BLEU 平均提升了 1.75 分在三个基准测试上的平均提升幅度，其中 News 上的提升达到 2.94 个 s-BLEU。使用 DA 模型生成的增强数据后，G-Transformer（rnd.）与 G-Transformer（fnt.）之间的差距从平均 1.26 个 s-BLEU 缩小至 0.18 个，这表明使用增强数据时，可能没有必要在句子级机器翻译模型上进行微调。对于 Transformer 基线，目标端增强使其性能平均提升了 1.33 个 s-BLEU。这些结果表明，目标端增强能够显著提升基线模型的表现，尤其是在小型数据集上。

与先前的工作相比，G-Transformer (fnt.)+目标端增强优于最佳系统 SMDT（该系统参考检索到的相似翻译），平均领先 1.40 s-BLEU。它也优于此前具有竞争力的 RecurrentMem（该系统在 TED 上取得了最佳分数），平均领先 1.58 s-BLEU。与 MultiResolution 相比，目标端增强也取得了更高的性能。MultiResolution 同样是一种数据增强方法，通过将文档拆分为不同的分辨率（例如，每个训练实例包含 1、2、4、8 个句子）来增加训练数据，而目标端增强的平均优势为 1.72 s-BLEU。采用目标端增强后，G-Transformer (fnt.) 在所有数据集上均取得了已报告的最佳 s-BLEU 分数。

方法	TED	新闻	Europarl	提升
G-Transformer（fnt.）	25.20	25.12	31.93	-
+ 基于先验的增强	25.69	26.34	32.16	+0.64
+ 基于后验的增强	26.59	28.06	32.85	+1.75

表 3：基于先验/后验的数据增强模型的机器翻译性能，以*s-BLEU*进行评估。

三个数据集。

与预训练设置相比，采用 G-Transformer（fnt.）进行目标端增强的效果优于在预训练 BERT 上进行微调的 Flat-Transformer+BERT 和 G-Transformer+BERT，在三个基准测试的平均得分上分别领先 1.46 和 0.70 个 s-BLEU 分，其中在 News 上的领先幅度分别达到 3.54 和 1.92 个百分点。在规模更大的数据集 Europarl 上，其得分甚至超过了采用大规模预训练的强基线 G-Transformer+mBART，这表明目标端增强对小规模和大规模数据集均有效。

回译与 Transformer 和 G-Transformer 基线相比，并未在 TED 和 Europarl 上带来足够显著的性能提升，但显著提升了 News 上的性能。在性能提升后的基线上，目标端增强进一步将 News 上的性能提升至新水平，分别达到 28.69 和 30.41 的最高源端/目标端 BLEU 分数。结果表明，目标端增强与回译技术相辅相成，在实践中将两者结合起来或许是最佳选择。

4.2 后验分布与先验分布

我们首先比较在 DA 模型中使用后验分布 $P(y|x_i,\ y_i)$（公式 5，第 3.3 节）与使用先验分布 $P(y|x_i)$时的机器翻译性能。表 3 显示，与使用原始数据相比，采用基于先验的增强后，性能平均提升了 0.64 s-BLEU。将 DA 模型替换为后验分布后，性能平均提升了 1.75 s-BLEU，高于使用先验分布所获得的提升。这些结果表明，使用 DA 模型（即使采用简单的先验分布）来增强目标序列是有效的，而后验分布还能进一步带来显著提升。

生成的译文。我们评估了生成译文的分布，如表4所示。使用先验分布，我们获得的译文比使用后验分布获得的译文具有更高的多样性。

方法	多样性↑	偏差↓	困惑度↓
先验分布	**78.68**	76.55	8.68
后验分布	45.42	**47.14**	**7.00**

表 4：DA模型生成译文的质量及其估计分布的准确性，在*News*上进行评估。

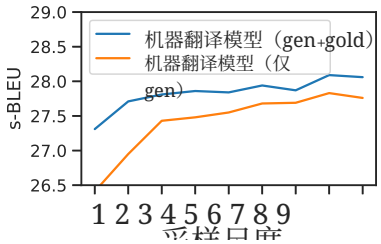


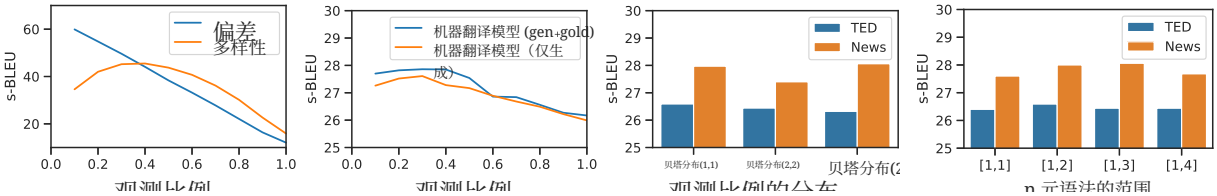
图 3：在 G-Transformer（fnt.）上训练并在*s-BLEU* 上*News*进行评估时，z的采样尺度的影响。（gen+gold）——在生成的译文和参考译文上训练。（仅 gen）——在生成的译文上训练。

然而，如果生成的译文与目标分布不一致，多样性更高并不一定能带来更好的性能。正如偏差列所示，从后验分布采样的译文，其偏差远小于从先验分布采样的译文，这证实了估计后验分布的 DA 模型能够生成更接近参考目标的译文。

估计分布的准确性。为了进一步提供直接证据，支持采用后验分布的 DA 模型，我们在多参考数据集上评估模型的困惑度（PPL）；预期更好的模型在参考译文上的 PPL 更低（附录C.1）。如表4中的 PPL 列所示，后验分布的平均 PPL（每个词元）为 7.00，先验分布为 8.68，前者比后者低 19.4%，这证实了我们的假设：后验分布能够更准确地估计数据分布 $P_{data}(y|x_i)$。

4.3 潜变量 z 的采样

尺度。采样尺度 $|\mathcal{Y}|$ 在公式 7中是影响模型性能的重要因素。从理论上讲，尺度越大，近似就越准确。图3 评估了生成的译文在不同尺度下的性能。总体趋势证实了理论预期：随着尺度增大，性能会提升。与此同时，随着尺度增大，人工参考译文的贡献会下降，这表明生成的译文越多，人工参考译文提供的



(a) 翻译质量由DA模型生成，在*News*上评估

(b) 机器翻译模型的性能在增强数据上的性能，在*News*上评估

(c) 使用混合观测比例训练的机器翻译模型的性能

图 4：观测比例对 z的影响，使用 G-Transformer（fnt.）进行训练，并在 *s-BLEU*上进行评估。Beta(a,b) – 函数曲线如附录 B.3所示。

图 5：在 G-Transformer（fnt.）上训练并使用 *s-BLEU* 进行评估的 n 元语法粒度的影响。

额外信息就越少。此外，尺度为 $\times 1$ 和 $\times 9$ 时的性能相差 0.75 s-BLEU，这表明机器翻译模型需要从 DA模型中获取足够的样本，才能匹配其分布。实际应用中，我们需要权衡性能提升和训练成本，以确定合适的采样尺度。

观测比例。 通过使用观测比例（α 见公式 1），我们可以控制潜变量 z所提供的信息量。该比例会影响生成的译文的质量。如图 4a 所示，较高的观测比例会使译文与标准参考译文之间的偏差更小，呈现出单调递减的曲线。相比之下，生成的译文的多样性呈现出凸曲线，即当观测比例较小或较大时，多样性较低，而在中间值时较高。生成的译文的多样性代表增强数据集的平滑程度，这会直接影响模型的性能。

如图4b 所示，机器翻译模型在比例约为 0.4 时取得最佳性能，此时偏差与多样性达到了较为均衡的水平。当比例进一步提升时，性能会下降。比较使用和未使用人工参考译文训练的机器翻译模型，可以发现，当观测比例大于 0.6 时，两种设置之间的性能差距正在缩小，此时生成的译文与人工参考译文之间的偏差较小。

通过混合不同观测比例下的生成译文，可以进一步提高多样性。因此，我们不使用固定比例，而是从预先定义的贝塔分布中采样比例。如图 $Beta(1,\ 1)$ 所示，我们比较了不同贝塔分布下的性能。TED 上的性能在 $Beta(1,\ 1)$ 处达到峰值，但与另外两种情况相比差异并不显著；而 News 上的性能峰值

方法	TED 源端/目标端 BLEU	News 源端/目标端 BLEU	提升 s-BLEU
G-Transformer（fnt.）	25.20 / 27.94	25.12 / 27.02	-
+源文本端增强	25.74 / 28.30	26.82 / 28.61	+1.12
+目标端增强	26.59 / 29.20	28.06 / 29.83	+2.17
+双侧增强	26.85 / 29.46	28.31 / 29.99	+2.42

表 5：源文本端增强与目标端增强。

出现在 $Beta(2,\ 3)$处，其分布为单峰分布，极值出现在比例为 0.3 至 0.4 之间，且形状与图 4a中的多样性曲线相似。与 $Beta(2,\ 2)$相比，后者也是单峰分布，但极值出现在比例为 0.5 处；使用 $Beta(2,\ 3)$ 时，性能高出 0.66 s-BLEU。

n 元语法的粒度。 n 元语法的粒度决定了通过潜变量可观测到多少词元之间的顺序信息 z（相比之下，观测比例决定了可观测到多少词元）。我们评估了不同范围的 n 元语法，并根据从该范围内均匀采样的数值来抽取 n 元语法。如图 5 所示，对于 $[1, 2]$ TED，性能达到峰值；对于 $[1, 3]$ News，性能达到峰值。然而，差异相对较小，说明性能对原始参考文本中的词元顺序并不敏感。一个可能的原因是，DA模型可以根据源句子提供的语义信息重建顺序。

4.4 不同的增强方法

源端增强与双端增强。我们将 DA 模型应用于源文本端和两端，并将目标端增强与源端增强和双端增强进行比较。如表5所示，源端增强使 TED 和 News 的平均基线 s-BLEU 提高了 1.12，但仍显著低于目标端增强；后者使平均基线 s-BLEU 提高了 2.17。结合两者

方法	Dev	Test
Transformer（基础版）	34.85	33.87
+ T5释义器 ♢	34.01	33.10
+ 目标端增强	**36.42**	**35.42**

表6：目标端增强与释义器在句子级机器翻译上的比较，在 IWSLT14 德语-英语（De-En）上进行评估。♢ – 使用 $p = 0.95$进行核采样。

通过结合源端增强和目标端增强生成的数据，我们平均获得了 2.42 s-BLEU 的提升，而源端增强数据进一步使目标端增强的效果平均提高了 0.25 s-BLEU。这些结果表明，DA模型在源端增强方面是有效的，但在目标端增强方面的效果更为显著。

释义。目标端增强通过加入新的翻译来扩充平行数据，这些翻译可视为原始人工参考译文的释义。也可以借助外部释义器来实现这种释义。我们在句子级机器翻译任务上，将目标端增强与预训练的 T5 释义器进行比较，设置详见附录C.3。

如表6所示，T5 释义器在开发集和测试集上的表现均不如 Transformer 基线，而目标端增强在开发集和测试集上分别比基线高出 1.57 和 1.55。结果表明，DA模型对句子级机器翻译有效，但释义器可能无效，这可能是因为缺少翻译信息。

具体而言，T5 释义器生成的释义多样性为 40.24，接近 DA模型生成的 37.30。然而，在基线 Transformer 上计算困惑度（PPL）以比较这些翻译时，T5 释义器的 PPL 为 3.40，而 DA模型的 PPL 为 1.89。结果表明，与外部释义器相比，DA模型生成的翻译与金标准目标的分布更加一致。

4.5 进一步分析

DA模型的规模。对观测到的翻译进行条件约束后，DA模型预测目标时会变得更简单。因此，生成的翻译对DA模型容量的敏感度较低。不同规模DA模型的结果证实了这一假设，并表明即使使用小得多的DA模型，机器翻译性能也会提升。详情见附录C.2。

案例研究。我们列举了若干德英翻译中的单词、短语和句子案例，以及两个英德翻译文档，以展示数据增强模型生成的翻译的多样性。详情见附录C.4。

5 结论

我们研究了一种目标端数据增强方法，该方法引入一个 DA 模型来生成大量可能的译文，并在这些平滑后的目标文本上训练机器翻译模型。实验表明，我们的目标端增强方法减轻了数据稀疏性问题的影响，相较于基线取得了显著提升，并在 News 和 Europarl 上达到新的最先进水平。分析表明，高多样性与低偏差之间的平衡是取得提升的关键。据我们所知，我们是首个在文档级机器翻译场景中进行目标端增强的研究。

局限性

直观而言，长文档比短文档有更多可能的译文，因此在增强数据时，动态调整生成的译文数量或许是更好的选择，从而平衡训练成本与性能提升。另一种可能的解决方案是采样少量译文，并将这些译文作为解码器输入，迫使机器翻译模型匹配 DA 模型的动态分布，这与 Khayrallah 等人（2020年）的方法类似。这种动态采样与匹配或许可以用于提升训练效率。本文没有研究这一解决方案，相关探索留待未来工作。

目标端增强有望应用于其他存在数据稀疏性问题的序列到序列任务。由于会议投稿篇幅有限，我们将把对其他任务的研究留待未来工作。

致谢

我们感谢匿名审稿人提出的宝贵意见。本研究得到中国战略性科学技术创新合作项目（资助编号：SQ2022YFE020038）和国家自然科学基金委员会（资助编号：NSFC 62161160339）的资助。滕志扬还获得了CAAI-华为MindSpore开放基金（CAAIXSJLJJ-2021-046A）的部分支持。

参考文献

Jimmy Ba 和 Rich Caruana。2014年。深度网络真的需要很深吗？神经信息处理系统进展，27。

鲍广胜、Zebin Ou 和张岳。2023年。Gemini：控制抽象式文本摘要的句子级写作风格。*arXiv* 预印本 *arXiv:2304.03548*。

鲍广胜和张岳。2021年。用于文本摘要的上下文化改写。载于*AAAI*人工智能大会论文集，第35卷，第12544–12553页。

鲍广胜和张岳。2023年。一种用于文本摘要的通用上下文化改写框架。*IEEE/ACM* 音频、语音与语言处理汇刊。

鲍广胜、张岳、滕志扬、Boxing Chen 和 Weihua Luo。2021年。用于文档级机器翻译的 G-transformer。载于 计算语言学协会第*59*届年会暨第*11*届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），第3442–3455页。

Mikhail Belkin、Daniel Hsu、Siyuan Ma 和 Soumik Mandal。2019。调和现代机器学习实践与经典的偏差—方差权衡。美国国家科学院院刊，116(32):15849–15854。

Nikolay Bogoychev 和 Rico Sennrich。2019。用于神经机器翻译的合成数据中的领域、翻译腔和噪声。*arXiv* 预印本 *arXiv:1911.03362*。

Verna Dankers、Elia Bruni 和 Dieuwke Hupkes。2022年。自然语言组合性的悖论：神经机器翻译案例研究。载于第*60*届计算语言学协会年会论文集（第*1*卷：长篇论文），第4154–4175页。

Marzieh Fadaee、Arianna Bisazza 和 Christof Monz。2017。低资源神经机器翻译的数据增强。载于计算语言学协会第 *55* 届年会论文集（第 *2* 卷：短篇论文），第 567–573 页。

Akhbardeh Farhad、Arkhangorodsky Arkady、Biesialska Magdalena、Bojar Ondˇrej、Chatterjee Rajen、Chaud- hary Vishrav、Marta R Costa-jussa、España-Bonet Cristina、Fan Angela、Federmann Christian 等。2021 年。2021 年机器翻译会议（WMT21）论文集。在第六届机器翻译会议上，第 1–88 页。计算语言学协会。

Yukun Feng、Feng Li、Ziang Song、Boyuan Zheng 和 Philipp Koehn。2022年。学会记忆：用于文档级机器翻译的循环记忆 Transformer。*arXiv*预印本 *arXiv:2205.01546*。

Jianfeng Gao、Xiaodong He、Wen-tau Yih 和 Li Deng。2014年。学习用于翻译建模的连续短语表示。载于计算语言学协会第*52*届年会论文集（第*1*卷：长篇论文），第699–709页。

Zhengxian Gong、Min Zhang 和 Guodong Zhou。2011年。基于缓存的文档级统计机器翻译。载于*2011*年自然语言处理中的实证方法会议论文集，第909–919页。Mitchell A Gordon 和 Kevin Duh。2019年。将序列级知识蒸馏解释为神经机器翻译的数据增强。*arXiv*预印本 *arXiv:1912.03334*。

Jianping Gou、Baosheng Yu、StephenJMaybank 和 Dacheng Tao。2021 年。知识蒸馏：一项综述。国际计算机视觉杂志，129（6）：1789–1819。

Christian Hardmeier、Sara Stymne、Jörg Tiedemann 和 Joakim Nivre。2013 年。Docent：用于基于短语的统计机器翻译的文档级解码器。在*ACL 2013*（计算语言学协会第 *51* 届年会）；*2013* 年 *8* 月 *4–9* 日；保加利亚索非亚上，第 193–198 页。计算语言学协会。

Junxian He、Jiatao Gu、Jiajun Shen 和 Marc'Aurelio Ranzato。2019。重新审视神经序列生成的自训练。载于国际学习表征会议。

Xuanli He、Gholamreza Haffari 和 Mohammad Norouzi。2018。用于多样化机器翻译的序列到序列混合模型。载于第*22*届计算自然语言学习会议论文集，第583–592页。

Geoffrey Hinton、Oriol Vinyals、Jeff Dean 等。提炼神经网络中的知识。

Ari Holtzman、Jan Buys、Li Du、Maxwell Forbes 和 Yejin Choi。2019。神经文本退化的奇特案例。载于国际学习表征会议。

J Edward Hu、Huda Khayrallah、Ryan Culkin、Patrick Xia、Tongfei Chen、Matt Post 和 Benjamin Van Durme。2019。改进用于翻译和单语改写的词汇约束解码。载于*2019*年北美计算语言学协会分会会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第839–850页。

Marcin Junczys-Dowmunt。2019。微软翻译在 WMT 2019：迈向大规模文档级神经机器翻译。载于第四届机器翻译会议论文集（第 *2* 卷：共享任务论文，第 *1* 天），第 225–233 页。

Huda Khayrallah、Brian Thompson、Matt Post 和 Philipp Koehn。2020年。模拟多参考译文训练提升了低资源机器翻译的性能。载于*2020*年自然语言处理中的实证方法会议论文集（*EMNLP*），第82–89页。

Yoon Kim 和 Alexander M Rush。2016。序列级知识蒸馏。载于《*2016*年自然语言处理中的实证方法会议论文集》，第1317–1327页。

Philipp Koehn、Hieu Hoang、Alexandra Birch、Chris Callison-Burch、Marcello Federico、Nicola Bertoldi、Brooke Cowan、Wade Shen、Christine Moran、Richard Zens 等。2007。Moses：用于统计机器翻译的开源工具包。载于《第*45*届计算语言学协会年会论文集：演示与海报环节论文集伴卷》，第177–180页。

Philipp Koehn 和 Rebecca Knowles。2017。神经机器翻译面临的六大挑战。载于《首届神经机器翻译研讨会》，第28–39页。计算语言学协会。

Taku Kudo。2018。子词正则化：利用多个子词候选项改进神经网络翻译模型。载于《第*56*届计算语言学协会年会论文集（第*1*卷：长篇论文）》，第66–75页。

Bohan Li、Yutai Hou 和 Wanxiang Che。2022年。自然语言处理中的数据增强方法：综述。*AIOpen*。

Yafu Li、Yongjing Yin、Yulong Chen 和张岳。2021年。关于神经机器翻译的组合泛化。在第*59*届计算语言学协会年会暨第*11*届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），第4767–4780页。

Alexander Lin、Jeremy Wohlwend、Howard Chen 和 Tao Lei。2020年。通过模仿学习进行自回归知识蒸馏。载于*2020*年自然语言处理中的实证方法会议（*EMNLP*）论文集，第6121–6133页。

Yinhan Liu、Jiatao Gu、Naman Goyal、Xian Li、Sergey Edunov、Marjan Ghazvininejad、Mike Lewis 和 Luke Zettlemoyer。2020年。用于神经机器翻译的多语言去噪预训练。计算语言学协会汇刊，8：726–742。

Shuming Ma、Dongdong Zhang 和 Ming Zhou。2020年。一种简单有效的文档级机器翻译统一编码器。载于计算语言学协会第*58*届年会论文集，第3505–3511页。

Nitin Madnani、Necip Fazil Ayan、Philip Resnik 和 Bonnie Dorr。2007 年。使用释义进行统计机器翻译中的参数调优。载于第二届统计机器翻译研讨会论文集，第 120–127 页。

Sameen Maruf、André FT Martins 和 Gholamreza Haffari。2019年。面向上下文感知神经机器翻译的选择性注意力。在*2019*年计算语言学协会北美分会年会论文集：人类语言技术，第*1*卷（长篇与短篇论文），第3092–3102页。

Lesly Miculicich、Dhananjay Ram、Nikolaos Pappas 和 James Henderson。2018 年。采用层次注意力网络的文档级神经机器翻译。载于自然语言处理实证方法会议（*EMNLP*）论文集，CONF。

Mohammad Norouzi、Samy Bengio、Navdeep Jaitly、Mike Schuster、Yonghui Wu、Dale Schuurmans 等人。2016。用于神经结构化预测的奖励增强最大似然。神经信息处理系统进展，29。

Artidoro Pagnoni、Kevin Liu 和 Shangyan Li。2018。用于神经机器翻译的条件变分自编码器。*arXiv*预印本 *arXiv:1812.04405*。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei- Jing Zhu。2002年。BLEU：一种用于自动评估机器翻译的方法。载于计算语言学协会第*40*届年会论文集，第311–318页。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。2020年。探索统一文本到文本转换器迁移学习的极限。机器学习研究杂志，21(140)：1–67。

Beatrice Savoldi、Marco Gaido、Luisa Bentivogli、Mat- teo Negri 和 Marco Turchi。2021年。机器翻译中的性别偏见。计算语言学协会汇刊，9：845–874。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016a。利用单语数据改进神经机器翻译模型。载于计算语言学协会第*54*届年会，第86–96页。计算语言学协会（ACL）。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016b。使用子词单元对罕见词进行神经机器翻译。发表于计算语言学协会第 *54* 届年会，第 1715–1725 页。计算语言学协会（ACL）。

Connor Shorten 和 Taghi M Khoshgoftaar。2019 年。《深度学习的图像数据增强综述》。*Journal of big data*，6(1)：1–48。

Connor Shorten、Taghi M Khoshgoftaar 和 Borko Furht。2021年。《深度学习的文本数据增强》。*Journalof bigData*，8(1)：1–34。

Zewei Sun、Mingxuan Wang、Hao Zhou、Chengqi Zhao、Shujian Huang、Jiajun Chen 和 Lei Li。2022年。重新思考文档级神经机器翻译。载于计算语言学协会论文集：*ACL 2022*，第 3537–3548 页。

Brian Thompson 和 Matt Post。2020年。将释义生成视为零样本多语言翻译：区分语义相似性与词汇和句法多样性。载于第五届机器翻译大会论文集，第 561–570 页。

David A Van Dyk 和 Xiao-Li Meng。2001年。数据增强的艺术。*Journal of ComputationalandGraphicalStatistics*，10(1)：1–50。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。注意力就是你所需的一切。神经信息处理系统进展，30。

Michelle Y Wang 和 Trevor Park。2020年。《贝叶斯推断简述》。复杂数据上的贝叶斯推断，17。

Xinyi Wang、Hieu Pham、Zihang Dai 和 Graham Neu- big。2018年。Switchout：一种用于神经机器翻译的高效数据增强算法。载于*2018*年自然语言处理中的实证方法会议论文集，第856–861页。

Jason Wei 和 Kai Zou。2019年。EDA：用于提升文本分类任务性能的简易数据增强技术。载于*2019*年自然语言处理中的实证方法会议暨第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第6382–6388页。

Lesly Miculicich Werlen、Dhananjay Ram、Nikolaos Pappas 和 James Henderson。2018年。采用层次注意力网络的文档级神经机器翻译。载于*EMNLP*。

Xing Wu、Shangwen Lv、Liangjun Zang、Jizhong Han 和 Songlin Hu。2019年。条件式 BERT 上下文增强。载于国际计算科学会议，第84–95页。施普林格。

Shufang Xie、Ang Lv、Yingce Xia、Lijun Wu、Tao Qin、Tie-Yan Liu 和 Rui Yan。2022年。序列到序列生成的目标端输入增强。载于国际学习表征会议。

Biao Zhang、Deyi Xiong、Jinsong Su、Hong Duan 和 Min Zhang。2016 年。变分神经机器翻译。载于 *2016* 年自然语言处理中的实证方法会议论文集，第 521–530 页。

Xu Zhang、Jian Yang、Haoyang Huang、Shuming Ma、Dongdong Zhang、Jinlong Li 和 Furu Wei。2022年。Smdt：选择性记忆增强神经文档翻译。*arXiv* 预印本 *arXiv:2201.01631*。

Yuan Zhang、Jason Baldridge 和 Luheng He。2019年。Paws：来自词语乱序的释义对抗样本。载于*2019*年北美计算语言学协会分会会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第1298–1308页。

Zaixiang Zheng、Yue Xiang、Shujian Huang、Jiajun Chen 和 Alexandra Birch-Mayne。2020年。充分利用神经机器翻译中的上下文。载于第*29*届国际人工智能联合会议，第3983–3989页。国际人工智能联合会议组织。

Barret Zoph、Golnaz Ghiasi、Tsung-Yi Lin、Yin Cui、Hanxiao Liu、Ekin DogusCubuk 和 Quoc Le。2020年。重新思考预训练和自训练。神经信息处理系统进展，33：3833–3845。

A G-Transformer

G-Transformer（Bao 等人，2021年）采用编码器-解码器架构，包含两种多头注意力机制。一种用于全局文档，称为全局注意力；另一种用于局部句子，称为组注意力。

全局注意力。全局注意力就是普通的多头注意力，它关注整个文档。

$$
\begin{aligned}
args &= (Q, K, V), \\
\text{GlobalAttn}(args) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,
\end{aligned} \tag{8}
$$

其中，矩阵输入 Q、K、V 分别是用于计算注意力的查询、键和值。

组注意力。组注意力通过为每个句子分配一个组标签（Bao 和 Zhang，2021年，2023年；Bao 等人，2023年）来区分文档中的句子。组标签是用于识别特定句子的数字，按照句子的顺序分配，其中第一个句子的组标签为 1，第二个句子为 2，以此类推。

组标签序列用于计算注意力掩码，以避免跨句注意力

$$
\begin{aligned}
args &= (Q, K, V, G_Q, G_K), \\
\text{GroupAttn}(args) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M(G_Q, G_K)\right)V,
\end{aligned} \tag{9}
$$

其中，G_Q 和 G_K 是查询和键的组标签序列。函数 $M(G_Q, G_K)$ 计算注意力掩码：对于 G_Q中的组标签和 G_K中的组标签，如果两个标签不同，则返回一个很大的负数；否则返回 0。

组合注意力 两个多头注意力通过门控求和模块进行组合

$$
\begin{aligned}
H_L &= \text{GroupMHA}(Q, K, V, G_Q, G_K), \\
H_G &= \text{GlobalMHA}(Q, K, V), \\
g &= \text{sigmoid}([H_L, H_G]W + b), \\
H &= H_L \odot g + H_G \odot (1 - g),
\end{aligned} \tag{10}
$$

其中，W 和 b 是可训练参数，$\odot$表示逐元素乘法。

G-Transformer 在较低层使用组注意力，在最上面的两层使用组合注意力。

B 数据集与评估指标

B.1 数据集

三个基准数据集如下。

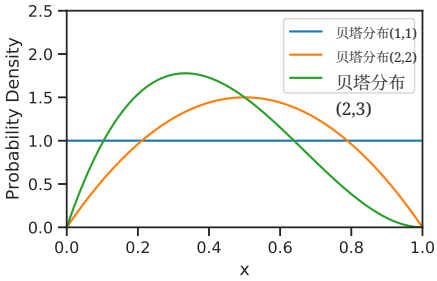


图 6：$Beta(a,\ b)$ 分布的概率密度函数。

TED 是一个来自 IWSLT2017 的语料库，包含 TED演讲的转录文本，每场演讲对应一份文档。源文本和目标文档中的句子按翻译关系对齐。我们使用*tst2016-2017*进行测试，其余部分用于开发。

News 是一个主要来自 News Commentary v11 的语料库，其中源文本与目标文档之间的句子也进行了对齐。我们使用*newstest2016* 进行测试，使用 *newstest2015* 进行开发。此外，我们使用来自 WMT21 的 *newstest2021*（Farhad 等人，2021年），该数据集为每个源文本提供三个参考译文，用于评估数据分布估计的质量。

Europarl 是从 Europarl v7 中提取的语料库，其中训练集、开发集和测试集是随机划分的。

我们使用 MOSES工具 对句子进行分词和规范大小写预处理（Koehn 等人，2007），随后进行字节对编码（Sennrich 等人，2016b），合并操作次数为 30000 次。

B.2 评估指标

句子级BLEU分数（s-BLEU）和文档级BLEU分数（d-BLEU）描述如下。

s-BLEU根据源文本和目标文档中的句子对计算，基本上与句子神经机器翻译模型的BLEU分数相同（Papineni 等人，2002）。

d-BLEU 是在文档对上计算的，将每个文档视为一个完整的词序列，并计算源文本序列与目标序列之间的BLEU分数。

对于分析，我们衡量生成的译文的偏差和多样性。

偏差 简单定义为与 perfects-BLEU 分数的距离

$$
\text{Deviation}(\hat{y}, y) = 100 - \text{s-BLEU}(\hat{y}, y), \tag{11}
$$

教师模型规模	超参数	参数	TED	News	Europarl	平均值
Base	6 层，8 个注意力头，隐藏维度 512，前馈网络 2048	69M	26.59	28.06	32.85	29.17
小型	6 层，**4** 个头，隐藏层维度 512，前馈网络**1024**	56M	26.64	28.23	32.42	29.10
Tiny	6 层，**4** 个头，隐藏层维度**256**，前馈网络 **1024**	21M	26.73	28.08	32.30	29.04

Ta表 7：在 G-Transformer（fnt.）上训练的 DA模型规模对其影响，并以*s-BLEU*进行评估 .

其中 ˆy 是生成的译文， y 是参考译文。

多样性 通过对生成的译文的偏差分数取平均值来计算

$$\text{Diversity}(\hat{\mathcal{Y}}) = \frac{\sum_{i=1}^{M}\sum_{j=i+1}^{M}\text{Deviation}(\hat{y}_i, \hat{y}_j)}{M(M-1)/2}, \tag{12}$$

ˆ其中 $\mathcal{Y}$ 是一组生成的译文，包含 M 个元素。该指标与 He 等人 (2018) 提出的多样性指标类似，不同之处在于我们使用 s-BLEU 作为基本度量。

B.3 训练设置

我们对所有基线均采用基础模型，其参数量约为 6000 万。我们对默认设置中的若干超参数进行了调整，以更好地适应增强数据。首先，将模型的最大长度从 512 扩展至 1024。其次，对于 Europarl，将丢弃率从 0.3 调整为 0.1，而 News 和 TED 的丢弃率则保持为 0.3。最后，为了加快训练过程，我们将 TED 和 News 中 DA模型的训练耐心值从 10 降至 5，并将 Europarl 中的该值从 10 降至 2。

在 4 块 Tesla V100 GPU 上采用新设置运行时，G-Transformer（fnt.）基线在 TED 上需训练 2 小时，在 News 上需 2.5 小时，在 Europarl 上需 13 小时。将数据增强 9 倍后，机器翻译模型的 G-Transformer（fnt.）训练在 TED、News 和 Europarl 上分别耗时 10、16 和 49 小时。相比之下，DA模型在 TED、News 和 Europarl 上的训练分别耗时 5、8.5 和 25 小时。

贝塔分布。 我们使用贝塔分布对观测比例进行采样，考虑了三种基本候选分布，包括 $Beta(1, 1)$，$Beta(2, 2)$和 $Beta(2, 3)$，如图 6 所示。

<escape>我们通过将该图与图中所示的多样性曲线进行比较来决定选择4a，其中我们可以看到 $Beta(2,3)$ 与生成的译文的多样性曲线最为吻合。我们在第 4.3 节中的进一步分析证实，$Beta(2,3)$ 在 TED 和 News 上均表现均衡。</escape>

C 更多分析

C.1 多参考译文评估

作为更直接的证据，表明采用后验分布的数据增强（DA）模型比采用 $P_{data}(y|x_i)$ 先验分布的模型估计得更加准确，我们在多参考译文数据集 *newstest2021*上评估困惑度（PPL）。该数据集包含67篇文档和1002个源句子，每个源句子均有3个翻译版本。我们通过将其中一个译文作为观测到的翻译，其余两个作为测试翻译，进行交叉验证。利用公式2，我们通过对潜变量 z 进行足够次数的采样（例如100次）来近似计算后验概率。

C.2 DA模型的大小

后验分布简化了DA模型的翻译任务，因为输入潜变量 z 包含了目标的大量信息。因此，DA模型对模型容量不太敏感。我们评估了不同规模DA模型的目标端增强效果。结果如表7所示。当参数数量从69M减少到21M时，TED和News上的性能没有显著差异。在规模更大的Europarl上，性能下降了0.55 s-BLEU，但仍比基线G-Transformer（fnt.）高0.37 s-BLEU，这表明即使DA模型的容量远低于机器翻译模型，它仍能带来额外价值。

C.3 改写设置

我们使用 T5 释义器[1]，它是通过在英文释义语料上微调 T5 创建的（Raffel 等人，2020年）（Zhang 等人，2019），并将其作为代表性方法进行比较研究。鉴于 T5 释义器是在英语语料上训练的，且作用于句子级，我们逐句翻译文档，并在机器翻译基准 IWSLT14 德英任务上评估这些方法。对于每个目标句子，我们使用 T5 释义器进行核采样（Holtzman 等人，2019），采样得到 6 个释义。对于目标端增强，我们

[1]https://huggingface.co/Vamsi/T5_改写_Paws

级别	源文本	目标	生成的译文
Word	找出	识别	查明、弄清楚、寻找、了解、留意、看见
	难以置信的	难以置信地	难以置信的、令人惊讶地、极其、非常、显著地
	确信的	确信的	深信、认为、相当确定
短语	认为 这些 Ein- 限制 für 有意义	接受此类限制作为合理的——合理的	1) 认为这些限制是有用的 2)将此类限制视为合理的 3)将这些约束视为确定条件
	被动活动的耳朵 只要头部稍微..... tut .	被动移动的耳朵 当头部变得】【。	1) 被动地随头部动作而移动的耳朵当头部做出某种动作时。 2)被动移动耳朵一旦头部移动。 3) 会随着你的头部动作而移动的被动耳朵。
	一个由木头和布料 制成的物体，带有 内置的动作—— ——让你们相信 zu 它 ，她是 活着的。	一个由.....制成的物体 由木头和布料制成 with 内置运动 其中，以说服你 相信它有生命	1) 一个由木头和布料制成、带有运动的物件 内置设计，旨在让你相信它是活的。 2) 一个由木头和布料制成的物件，内置运动以 让你一直相信它是活的。 3) 一个木质布艺物件，内置运动以 让你相信它是活的。
句子	它只有在 有人让它活起来时才会活。	它之所以活着，是因为你 让它活起来。	1) 它只存在当你让它执行。 2) 它存在只有你创造它时才会存在。 3)它只存在于 因为你让他们这样做。
	在每个时刻上 舞台上的人辛勤劳作着 把娃娃脱掉。	所以只要它处于开启状态的每一刻都如此 这个舞台，它正在制造。 这场斗争。	1) 在舞台上的每一刻，它都在进行斗争的 木偶。 2) 舞台上的每一刻都会让木偶崩塌</style】【。 3）因此，每当它登上舞台时，木偶便在拼命 下来。
	他通过 一个 令人震惊的 关于毒性的故事 在一个.....上的污染 日本鱼市场， 毒素如何从 海洋食物链的起点 海洋食物链 一直到 找到我们的身体。	他展示了毒素的作用方式 在.....的底部 海洋食物链查找 进入我们的身 .....，令人震惊的是 有毒污染的故事 源自一家日本 鱼市场。	1) 他通过一个令人震惊的有毒负担故事，展示了关于 一家日本鱼市场， 它们如何从 海洋食物链的起点一路毒害到我们的身体。 2）他通过一个毒素影响 日本鱼市场的骇人故事，展示了毒素如何从海洋 食物链进入我们的身体。 3）他通过一个毒素污染的骇人故事 在日本鱼市场，展示毒素如何从 海洋食物链的起点进入我们的身体。

表8：DA模型在 IWSLT14 德英翻译任务上生成的翻译。

为每个源句子生成 6 个翻译，且不使用文档上下文。值得注意的是，与先前的释义增强方法（Khayrallah 等人，2020年）不同，后者让机器翻译模型直接学习释义器的分布，而我们使用采样得到的文本输出来训练机器翻译模型。

C.4 案例研究

我们的案例研究表明，DA模型能够在词语、短语和句子级别生成多样化的翻译。表中列出了几个德英翻译案例8。

我们进一步列出了两段文档级翻译，从中可以直观地了解目标端增强如何提升机器翻译的性能，如表9所示。

源文本： 2003年，埃尔顿·约翰与俄罗斯总统弗拉基米尔·普京将会面，讨论同性恋者权利问题。俄罗斯首富米哈伊尔·霍多尔科夫斯基在西伯利亚一处跑道上被持枪逮捕。由于公开挑战总统弗拉基米尔·普京，霍多尔科夫斯基被定罪，他的石油公司尤科斯遭到查封，他推动民主的努力也受到遏制。

目标： 埃尔顿·约翰与俄罗斯总统弗拉基米尔·普京会面，讨论同性恋者的权利。米哈伊尔·霍多尔科夫斯基，俄罗斯首富，在西伯利亚的一处机场停机坪上被武装人员逮捕。在公开挑战总统弗拉基米尔·普京后，霍多尔科夫斯基被判刑，其石油公司尤科斯被没收，且他的民主改革努力遭到阻挠。

基线： 埃尔顿·约翰和俄罗斯总统弗拉基米尔·普京必须会面，讨论同性恋问题。2003年，俄罗斯首富米哈伊尔·霍多尔科夫斯基在西伯利亚的一场补选期间被捕。在公开质疑总统弗拉基米尔·普京之后，霍多尔科夫斯基被判有罪，他的石油公司尤科斯遭到查封，而且他为民主所作的努力受到遏制。

我们的版本： 埃尔顿·约翰与俄罗斯总统弗拉基米尔·普京会面，讨论同性恋者权利。2003年，俄罗斯首富米哈伊尔·霍多尔科夫斯基在西伯利亚的一次补选期间被捕。此前，他曾公开质疑总统弗拉基米尔·普京。随后，霍多尔科夫斯基被判刑，他的石油公司尤科斯遭到接管，他推动民主的努力受到限制。

源文本： 上巴伐利亚的拉姆绍-贝希特斯加登是德国首个“登山村”。这座位于贝希特斯加登国家公园内、拥有1800名居民的村庄于周三从德国阿尔卑斯协会副主席路德维希·武赫普芬宁手中接过“温和旅游”奖项。奥地利已有20个“登山村”。在邻国奥地利，颁发这一称号由当地的阿尔卑斯协会负责。“登山村”最多允许有2500名居民。其至少五分之一的面积必须被划为保护区。

目标： 上巴伐利亚的拉姆绍-贝希特斯加登是德国首个“登山村”。由……颁发德国阿尔卑斯协会副主席路德维希·武赫普芬宁于周三向这座位于贝希特斯加登国家公园内、拥有1800名居民的村庄颁发了温和旅游奖。奥地利已有20个“登山村”。在邻国奥地利，颁发这一称号由当地的阿尔卑斯协会负责。“登山村”最多允许有2500名居民。其至少五分之一的面积必须被划为保护区。

基线： 上巴伐利亚的拉姆绍和贝希特斯加登区是德国首个“登山者村”。这座位于贝希特斯加登国家公园、拥有1,800名居民的村庄，于周三从德国阿尔卑斯协会副主席路德维希·武赫芬普菲格手中获得了“温和旅游”奖。奥地利已有20个“登山者村”。在我们的邻国，负责授予这一称号的是当地阿尔卑斯协会。“登山者村”的居民人数不得超过2,500人。至少五分之一的地区必须划为保护区。

我们的版本： 上巴伐利亚的贝希特斯加登附近拉姆绍区是德国首个“登山者村”。这座位于贝希特斯加登国家公园、拥有1,800名居民的村庄，于周三从德国阿尔卑斯协会副主席路德维希·武赫伦宁手中获得了“温和旅游”奖。奥地利已有20个“登山者村”。在我们的邻国，负责授予这一称号的是当地阿尔卑斯协会。“登山者村”的居民人数不得超过2,500人。其至少五分之一的区域必须被划为保护区。

表9：比较 G-Transformer（fnt.）基线和目标端增强的文档级翻译，在*News*英德语料上进行评估。