基于词频效应控制的神经机器翻译用词多样性增强方法

史学文[1], 鉴萍[2]*, 唐翼琨[2], 黄河燕[2]

[1]东北财经大学 数据科学与人工智能学院，中国 辽宁 大连，116025
[2]北京理工大学 计算机学院，中国 北京，100081
polarlion@qq.com {pjian, tangyk, hhy63}@bit.edu.cn

摘要

通过最大似然估计优化的神经机器翻译（NMT）容易出现不可最大化的标记或低频词精度差等问题，这会导致生成的翻译缺乏词级别的多样性。词频在训练数据上的不均衡分布是造成上述现象的原因之一。本文旨在通过限制词频对 NMT 解码时估计概率的影响来缓解上述问题。具体地，我们采用了基于因果推断理论的半同胞回归去噪框架，并结合本文提出的自适应去噪系数来控制词频对模型估计概率的影响，以获得更准确的模型估计概率，并丰富 NMT 译文用词的多样性。本文的实验在四个代表不同资源规模的翻译任务上进行，分别是维吾尔语-汉语、汉语-英语、英语-德语和英语-法语。实验结果表明，本文所提出的方法在提升 NMT 译文词级别多样性的同时，不会损害译文的质量。另外，本文提出的方法还具有模型无关、可解释性强等优点。

关键词： 神经机器翻译 ； 译文多样性 ； 因果推断

Improving Word-level Diversity in Neural Machine Translation by Controlling the Effects of Word Frequency

Xuewen Shi[1], Ping Jian[2]*, Yi-Kun Tang[2], Heyan Huang[2]

[1]School of Data Science and Artificial Intelligence, Dongbei University of Finance and Economics, Dalian, Liaoning, China, 116025
[2]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China, 100081
polarlion@qq.com {pjian, tangyk, hhy63}@bit.edu.cn

Abstract

Neural machine translation (NMT) optimized by maximum likelihood estimation is prone to problems such as unargmaxable tokens or poor accuracy of low-frequency words, which leads to the lack of word-level diversity in the generated translations. The unbalanced distribution of word frequency on the training data is one of the reasons for the above phenomenon. This paper aims to alleviate the above problems by limiting the impact of word frequency on the estimated probability when decoding NMT . Specifically, we adopt a denoising framework of Half-Sibling Regression based on causal inference theory, combined with the adaptive denoising coefficient proposed in this paper to control the effect of word frequency on estimated probability, in order to obtain more accurate model estimated probability, and enrich the diversity of the words used in NMT translations. The experiments in this paper are carried out on four translation tasks representing different resource scales: Uyghur-Chinese, Chinese-English, English-German and English-French. In addition, the proposed method is model-agnostic and interpretable.

Keywords: Neural machine translation , Translation diversity , Causal inference

* 通讯作者：鉴萍；Corresponding author: Ping Jian

源语言	警方 现 追缉 涉案 的 另 一 名 男子 。
参考译文	Police are still hunting for the other man involved in the case .
	$\log f_{req}(\text{“hunting”}) = -4.85$
NMT译文	Police are looking for the other man in connection with the case .
	$\log f_{req}(\text{“looking”}) = -3.97$

图 1: NMT生成译文与训练数据中的原译文对比

1 引言

近年来，端到端的神经机器翻译（Neural Machine Translation，NMT）(Sutskever et al., 2014; Bahdanau et al., 2015) 在机器翻译领域取得了令人瞩目的成就，在某些特定的翻译任务上，机器译文已经接近人类译文的水准 (Wu et al., 2016; Vaswani et al., 2017; Hassan et al., 2018)。NMT 模型通常构建在编码器-解码器 (Cho et al., 2014) 架构上，其中，编码器的作用是将源语言序列 $\mathbf{x} = \{x_1, \ldots, x_{T_x}\}$ 转换成一组隐状态表示 $\mathbf{h} = \{h_1, \ldots, h_{T_x}\}$，解码器则被用来对如公式 (1) 所示的译文 $\mathbf{y} = \{y_1, \ldots, y_{T_\mathbf{y}}\}$ 的翻译概率建模：

$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T_y} p(y_t|\mathbf{y}_{<t}, \mathbf{h}) 。 \tag{1}$$

经典的 NMT 方法 (Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) 通常利用最大似然估计（maximum likelihood estimation，MLE）对 NMT 模型进行优化，训练的损失函数 $\mathcal{L}_{nmt}$ 通常采用负对数似然的形式：

$$\mathcal{L}_{nmt} = -\log p(y_t|\mathbf{y}_{<t}, \mathbf{x}, \theta), \tag{2}$$

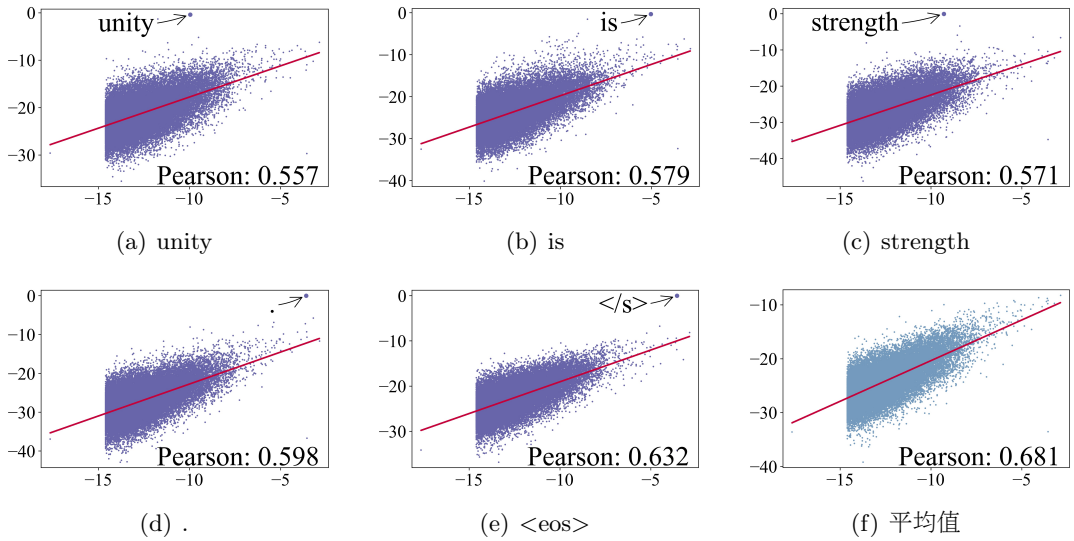
其中 θ 表示 NMT 模型的自由参数集合。

在机器翻译训练数据中，单词的词频分布是不平衡的，因此，经过 MLE 训练的 NMT 模型在解码阶段倾向于生成更高频的单词，而不是最适合的单词。例如，我们利用一个训练完成的汉语-英语 NMT 模型去重新翻译该模型所使用的训练集中的句子：“警方 现 追缉 涉案 的 另 一 名 男子 。”，图 1 给出了模型生成的译文与训练集中原始译文的对比。如图 1 所示，NMT 模型生成的译文与训练集的参考译文拥有相似的句法结构，但是对于源语言单词“追缉”，模型生成的译文使用了训练集中词频更高的单词“looking”而不是“hunting”。NMT 模型倾向于选择更高频单词的现象可能会引起“不可最大化的标记（unargmaxable tokens）”(Demeter et al., 2020; Grivas et al., 2022) 和“低频词准确率低”(Koehn and Knowles, 2017; Ott et al., 2018) 等问题。

针对上述问题，已有的工作主要分为以下两种方向：（1）在训练 NMT 时引入自适应的损失函数 (Lin et al., 2017; Gu et al., 2020; Xu et al., 2021; Zhang et al., 2022)；（2）尽可能消除词表示中与词频有关的部分信息 (Gong et al., 2018; Yang and Liu, 2020; Liu et al., 2020)。这两类工作的核心思想是尽量消除或缓解训练集中词分布不平衡对于训练损失信号或词向量分布的影响。上述方法必须作用于 NMT 的训练过程中，无法对已经存在的优化好的模型使用。

在本文，我们提出了一种基于词频效应控制的 NMT 译文用词多样性增强方法。该方法引入了半同胞回归 (Schölkopf et al., 2016) 去噪框架，结合本文提出的自适应降噪系数，通过调整目标语言单词的词频信息在 NMT 解码时的影响，以缓解 NMT 解码时倾向于选择高频词的问题，从而增强 NMT 译文的多样性。本文在 4 个不同语言对的翻译任务上进行了实验以验证提出的方法的有效性，分别是维吾尔语到汉语翻译（维-汉），汉语到英语翻译（汉-英），英语到德语翻译（英-德）以及英语到法语翻译（英-法）。上述四种翻译任务分别代表了四种任务类型：低资源翻译（维-汉），中等资源翻译（汉-英，英-德）以及丰富资源翻译（英-法）。实验结果表明，本文提出的方法在不损害译文质量的前提下，增强了 NMT 译文词级别多样性。

第二十二届中国计算语言学大会论文集，第64页-第77页，哈尔滨，中国，2023年8月3日至5日。



(a) unity (b) is (c) strength

(d) . (e) <eos> (f) 平均值

图 2: 词频分布与模型估计概率的关系

2 观察与讨论

2.1 模型估计概率与词频分布的相关性

观察发现，NMT 模型在解码时输出的概率分布（记作 O）通常与训练数据中目标语言词汇频率分布（记作 F）有很高的正相关性。例如，图 2(a) ∼ 图 2(e) 展示了汉-英翻译任务中，将源语言句子“团结 就 是 力量 。”翻译成目标语言句子“unity is strength .”时的每个解码步骤中，NMT 模型估计的目标语言词汇概率分布与目标语言词频分布的关系；而图 2(f) 展示了解码时模型输出概率的平均值与词频分布的关系。图中 x 轴表示目标语言单词频率的对数（即 $\log F$），y 轴表示对应单词的解码输出概率的对数（即 $\log O$）；图中直线表示以 x 轴为输入数据，y 轴为输出数据拟合的线性回归的曲线；“Pearson”表示 x 轴数据与 y 轴数据的 Pearson 相关系数，正值表示正相关，反之表示负相关。此外，图 2(a) ∼ 图 2(e) 标注了对应解码步骤模型估计概率排名最高的单词的位置。如图 2 所示，从线性回归的曲线图像和 Pearson 相关系数的数值（均大于 0.5）可以看出：在该翻译案例解码过程中，对于大多数目标语言单词，其在训练数据的词频与 NMT 模型解码时估计的概率值具有较高的正相关性。

2.2 相关性在不同概率区间的差异

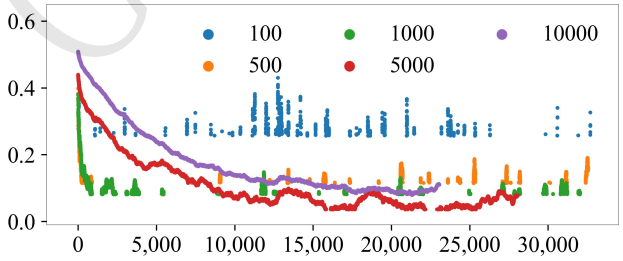


图 3: 不同概率区间下词频与模型生成概率的 Pearson 相关性对比

O 与 F 的相关性在不同的估计概率区间上的表现是有差异的。据观察，在每个解码步骤，排位较高的估计概率通常与对应的词频分布有较高的相关性。为了说明这一现象，我们选择图 2(c) 的解码步骤作为示例，在图 3 展示了在不同模型生成概率区间的 F 与 O 的 Pearson 相关系数。图 3 中横坐标表示窗口中单词的起始编号，其中单词编号越小表示对应单词的模型估计概率越大，编号为 1 的单词为训练数据中词频最高的单词；纵坐标表示区间内单词对应的 F 和 O 的 Pearson 相关系数。表中不同曲线表示不同的窗口大小，以窗口“500”对应

第二十二届中国计算语言学大会论文集，第64页-第77页，哈尔滨，中国，2023年8月3日至5日。

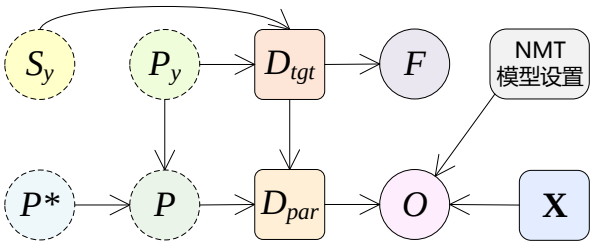


图 4: 词频与模型输出关系的因果有向图

的曲线为例，横坐标 0 对应的数据表示单词编号 1∼500 区间，而横坐标 10000 表示单词编号 10001 ∼ 10500 区间。在图 3 中，我们展示了 5 种不同的窗口大小对应的 Pearson 相关系数变化曲线，没有数据的部分表示该区间对应的 Pearson 相关系数未通过显著性检验（ p 值大于 0.01）。从图 3 可以看出，多数情况下，窗口起始位置越靠前，其对应的 Pearson 相关系数越大，即该区间内单词的 F 和 O 相关性越强。

2.3 讨论：模型估计概率与词频分布的因果关系

本小节将讨论章节 2.1 所述现象的可能原因。在解码步骤 t，给定源语言序列 $\mathbf{x}$ 和之前生成的译文序列 $\mathbf{y}_{<t}$，则目标语言单词 y_t 的条件概率分布 $P = p(y_t|\mathbf{y}_{<t}, \mathbf{x})$。$P$ 可以视为模型估计的概率分布（记作 O ）希望近似的理想值。根据贝叶斯法则，$p(y_t|\mathbf{y}_{<t}, \mathbf{x})$ 可以展开为：

$$P = p(y_t|\mathbf{y}_{<t}, \mathbf{x}) = \frac{p(y_t)p(\mathbf{y}_{<t}, \mathbf{x}|y_t)}{p(\mathbf{y}_{<t}, \mathbf{x})}。 \tag{3}$$

由于 $p(\mathbf{y}_{<t}, \mathbf{x})$ 对于任意 y_t 是确定值，由此可得到：

$$P = p(y_t|\mathbf{y}_{<t}, \mathbf{x}) \propto p(y_t)p(\mathbf{y}_{<t}, \mathbf{x}|y_t)。 \tag{4}$$

在实践中，先验概率分布 $p(y_t)$ 通常以训练数据中的词频分布 F 作为近似，这样，由公式 (3) 可以得出，条件概率分布 P 与 F 相关，该结论证明了章节 2.1 中所展示的现象的合理性。公式 (3) 表明 F 与 P 在理论上存在相关关系，尽管如此，我们仍认为在实践中 F 与 O 的相关系数要高出理论上的合理数值，也因此造成了如图 1 所呈现的问题。我们假设在训练过程中词频分布的偏差被转移到 NMT 模型中，在 NMT 的解码过程，该偏差会作为噪声干扰模型输出。

可观测性	变量名	说明
不可观测	D_{tgt}	NMT 模型训练数据中的目标语言数据
	S_y	构造 D_{tgt} 时使用的采样方法
	P_y	先验概率分布 $p(y_t)$
	P	理想的翻译概率（即 $p(y_t|\mathbf{y}_{<t}, \mathbf{x})$）
	P^*	影响 P 的其他因素，例如公式 (3) 中的 $p(\mathbf{y}_{<t}, \mathbf{x}|y_t)$
	D_{par}	平行数据，即 NMT 模型的训练数据，同时受到 D_{tgt} 和 P 的影响，这里不假设 D_{tgt} 是由源语言数据翻译而成的，例如先有目标语言数据，再经人工翻译成源语言数据，或源语言与目标语言数据均由第三语言经人工翻译而来
可观测	$\mathbf{x}$	源语言
	F	D_{tgt} 中的单词频率分布
	O	由 NMT 模型生成的估计概率，$\mathbf{x}$ 表示 NMT 模型的源语言输入
	模型设置	NMT 的模型架构、参数规模、训练算法等可能影响 NMT 性能的模型设置相关的因素

表 1: 因果有向图中各变量的物理意义

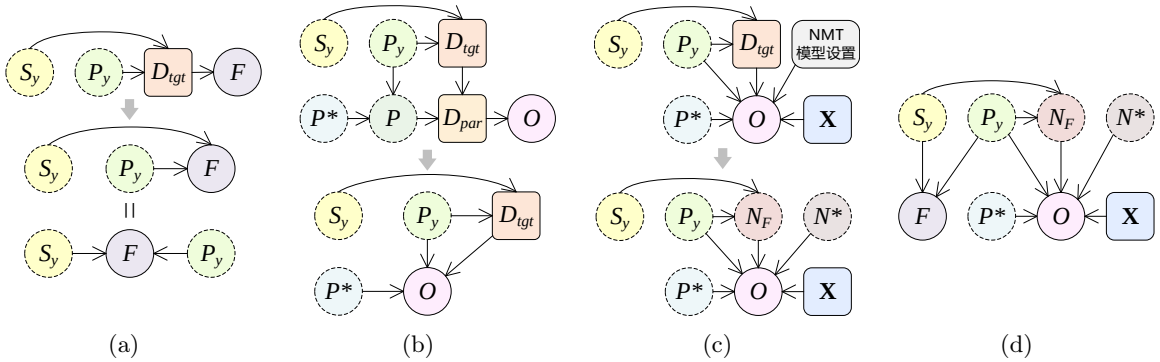


图 5: 对图 4 因果有向图的分解和简化过程

如图 4 所示，本文引入了一个因果有向无环图（Directed Acyclic Graph，DAG）说明上述假设，图 4 中各变量说明如表 1 所示。在图 4 中，S_y，P_y，P^*，P 均为不可观测的变量，无法获取其精确值。由于训练数据的源语言部分不在本文的研究范围，因此组成 D_{par} 的源语言数据、源语言分布等源语言相关因素均未在图中体现。在图 4 中，D_{par} 所示的双语平行数据，即 NMT 模型的训练数据，同时受到 D_{tgt} 和 P 影响。我们不假设 D_{tgt} 是由源语言数据翻译而成的，事实上，在语料构建过程中确实存在符合该情景的情况，例如：（1）先有目标语言数据，再经人工翻译成源语言数据，或（2）源语言与目标语言数据均由第三种语言翻译而来。

关于 F 与 O 的因果 DAG 的简化：由于本文主要关注词频分布 F 与模型输出 O 之间的关系，而图 4 所示关系涉及的变量过多，增加了研究的复杂度，因此我们在不破坏如图所示的因果关系的情况下，对图 4 进行进一步地简化。图 5(a) ∼ 图 5(c) 展示了具体的简化过程：

(a) 图 4 中 $P_y \to D_{tgt} \to F$ 和 $S_y \to D_{tgt}$ 两条路径上不存在其它因变量，因此可以合并为 $S_y \to F \leftarrow P_y$，如图 5(a) 所示；

(b) 同理，路径 $P^* \to P \to D_{par} \to O$ 和 $D_{tgt} \to D_{par} \to O$ 也可以合并，这样则可省略中间节点 P 和 D_{tgt}，如图 5(b) 所示；

(c) 根据观测到的现象，已知在 D_{tgt} 和 “模型设置” 的共同作用下得到的 O 并不准确，于是我们将 D_{tgt} 和 “模型设置” 对 O 起负面作用的因素拆分为两个不可观测变量 N_F 和 N^*，其中 N_F 表示与词频分布相关的噪声，N^* 表示其他噪声，而 D_{tgt} 和 “模型设置” 产生积极作用的部分则已经包含在路径 $P_y \to O$ 中，上述简化过程如图 5(c) 所示。

综合上述简化过程，图 4 经上述过程简化后的因果 DAG 如图 5(d) 所示，图中可观测的变量为 F、O 和 $\mathbf{x}$，其余均为不可观测的变量。由于 $\mathbf{x}$、N^* 与文本的研究内容无关，因此我们将二者合并为无关变量的集合，记作 U，这样得到最终的简化后的因果模型如图 6 所示。

3 方法

本文认为 NMT 模型解码过程中各步骤的估计概率 O 与训练数据的词频分布 F 之间的相关关系超出了合理的范围，且由此造成了 NMT 译文单词多样性下降的问题。我们假设存在与

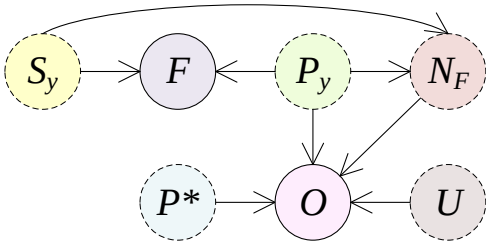


图 6: 简化后的词频与模型估计概率的因果关系图

算法 1 基于 HSR 的模型估计概率降噪方法

输入: 目标语言词汇表：V_y；NMT 模型估计的目标语言单词概率分布 O，其中 $|O| = |V_y|$；目标语言单词词频分布 F，其中 $|F| = |V_y|$；调节参数 $\alpha \in [0, 1]$。

1: 利用回归模型 $R(\cdot)$ 估计 $\mathrm{E}[O|F]$，以 F 为自变量数据，O 为因变量数据，根据最小均方误差找到回归函数参数 θ_R 的最优解：

$$\theta_R^* = \arg\min_{\theta_R} |R(F; \theta_R) - O|^2 。 \tag{7}$$

2: 从模型输出 O 中消除目标语言词频分布 F 的部分影响:

$$O' \leftarrow O - \alpha R^*(F; \theta_R^*), \tag{8}$$

其中 $R^*(\cdot)$ 表示最优化后的回归模型，θ_R^*为回归模型最优化参数。

输出: 降噪后的目标语言单词的概率分布 O'。

目标语言词频相关的噪声 N_F，该噪声造成了上述现象。由图 4 和图 6 所展示的因果关系可知，F 与 N_F 共享 S_y 和 P_y 两个原因变量，根据结构因果模型（structural causal model）中分叉结构的性质 (Pearl et al., 2016) 可知:

$$p(N_F|F, S_y, P_y) \neq p(N_F|S_y, P_y), \tag{5}$$

即 N_F 与 F 相互关联（$N_F \perp\!\!\!\perp F$）。基于 N_F 与 F 的关联性，本文试图通过控制 F 对 O 的作用，以减弱 N_F 对 O 造成的负面影响。

本文采用半同胞回归（Half-Sibling Regression，HSR）(Schölkopf et al., 2016) 的方法控制噪声 N_F 对 O 的影响。根据 HSR (Schölkopf et al., 2016)，假设存在不可观测变量 U_1 和 U_2 同时作用于可观测变量 O_1，若存在可观测变量 O_2 满足 $O_2 \not\perp\!\!\!\perp U_2$，则可通过控制 O_2 在 O_1 中产生的效应 $\mathrm{E}[O_1|O_2]$ 来控制 U_2 对O_1 的影响。在本文提出的场景中，如图 6 所示，由于 N 与 F 相关联（即 $F \not\perp\!\!\!\perp N$），通过控制 F 在 O 产生的效应，则可实现缓解噪声 N_F 对 O 的影响:

$$O' \leftarrow O - \alpha \mathrm{E}[O|F], \tag{6}$$

其中 O' 表示降噪后的模型估计概率，$\mathrm{E}[O|F]$ 表示 F 在 O 上产生的效应，在实践中，$\mathrm{E}[O|F]$ 通过回归模型估计。$\alpha \in [0, 1)$ 表示降噪系数。由图 6 可知，F 包含了 P_y 的信息，为防止执行公式 (6) 的操作时过度破坏 P_y 对 O 的影响，因此本文引入降噪系数 α 以控制降噪操作的力度。

上述方法通过对 NMT 模型解码时输出的估计概率进行降噪处理来缓解词频分布的偏置对于 NMT 模型预测精度的影响，无需更改模型设置或对 NMT 的训练进行干预，是完全模型无关（model agnostic）的方法。算法 1 展示了上述方法的操作过程。

3.1 分区间回归与自适应的降噪系数

在算法 1 中，线性回归模型在 F 与 O 的全集上优化得到（记作 SR，Set-based Regression），降噪系数 α 采用固定的常数（记作 CDR，Constant Denoising Ratio），上述 SR+CDR 的方法假设：对于不同频率的目标单词 y_i，其对应的模型估计概率 o_i 受到的与词频 f_i 相关的影响函数是一致的。然而，根据章节 2.2 中所讨论的现象，不同的目标语言词频分布 F 的区间与对应的估计概率 O 的区间的相关关系是有差异的，因此，利用 F 与 O 的全集优化得到的线性回归模型可能无法拟合不同词频区间上的真实情况，这样估计得到的 $\mathrm{E}[O|F]$ 误差较大。为缓解该问题，本文引入了分区间回归和自适应的降噪系数，以优化对降噪项（即 $\alpha\mathrm{E}[O|F]$ ）的估计，二者的具体操作方法如下：

（1）分区间回归（记作 PR，Partition-based Regression），将回归模型的训练数据 $< F, O >$ 切分成 N_R 组不同的区间 $\{< F_1, O_1 >, \ldots, < F_{N_R}, O_{N_R} >\}$，并以此分别计算得到 N_R 个回归模型 $\{R_1, \ldots, R_{N_R}\}$，则公式 (8) 将改写为：

$$O'_i \leftarrow O_i - \alpha R_i^*(F_i; \theta_{R_i}^*), \quad i \in \{1, \ldots, N_R\}。 \tag{9}$$

上述区间的划分由人工凭借经验完成，本文首先将词汇表按词频从大到小排序，采用固定的步长划分区间，其中步长等于 4,000，即编号 1 ∼ 4,000 的单词对应的词频和估计概率作为第一组数据 $< F_1, O_1 >$，号 4,001 ∼ 8,000 的单词对应的词频和估计概率作为第二组数据 $< F_2, O_2 >$，以此类推；

（2）自适应的降噪系数（记作 SADR，Self-Adaptive Denoising Ratio），即对不同的训练数据对 $< f_i, o_i > \in < F, O >$ 使用不同的降噪系数 α_i：

$$o_i' \leftarrow o_i - \alpha_i R^*(f_i; \theta_R^*), \quad i \in \{1, \ldots, V_y\}。 \tag{10}$$

本文 α 通过对 F 取自然对数的相反数再经过最大-最小值缩放（Max-Min Scaling）后得到：

$$\alpha_i = \frac{-\log f_i - \min(-\log F)}{\max(-\log F) - \min(-\log F)}, \quad i \in \{1, \ldots, V_y\}, \tag{11}$$

其中 max(·) 和 min(·) 分别表示取集合中最大值和最小值的函数。上述降噪系数的计算方法的物理意义为：对于词频较高的单词取较小的降噪系数，降噪力度较小，反之对于低频词则取较大的降噪系数，降噪力度较大。

本文的实验部分（章节 4 ）将对比两种回归模型的优化方式（ SR 和 PR ）和两种降噪系数设置方法（ CDR 和 SADR ）共 4 种组合方式：（1）SR+CDR，基于全集的回归模型结合固定的降噪系数；（2）PR+CDR，分区间回归模型结合固定的降噪系数；（3）SR+SADR，基于全集的回归模型结合自适应的降噪系数；（4）PR+SADR，分区间回归模型结合自适应的降噪系数。在实验中，SR 设置为 0.01。

4 实验

4.1 实验数据

为验证本文提出的方法，我们在如下 4 组翻译任务上进行实验：维吾尔语到汉语翻译（维-汉），汉语到英语翻译（汉-英），英语到德语翻译（英-德）以及英语到法语翻译（英-法）。上述语料的具体信息如下：

维-汉：训练集数据、测试集数据以及验证集数据均来自于 CCMT2019 机器翻译评测维吾尔语到汉语新闻翻译任务 (Yang et al., 2019)。维吾尔语端除了词例化 (Koehn et al., 2007) 和词拆分 (Kudo and Richardson, 2018) 外，并未使用其他分词工具。

汉-英：训练集数据提取自多组 LDC 语料[1]. 本文采用 NIST'02 测试数据作为汉-英翻译的验证集数据，NIST'03∼NIST'06 数据作为汉-英翻译的测试集数据.

英-德和英-法：英-德和英-法翻译任务的训练集数据来自于 WMT'14 (Bojar et al., 2014) 的公开数据，验证集和测试集数据分别是 newstest2013 和 newstest2014。

上述四个翻译任务分别代表低资源翻译（维-汉）、中等资源翻译（汉-英、英-德）和丰富资源翻译（英-法）等三类翻译任务。翻译任务的数据均经过 Moses (Koehn et al., 2007) 的 *tokenizer.perl* 脚本[2]进行词例化，之后由 sentence-piece (Kudo and Richardson, 2018) 工具进行词拆分以缩小词表。对于维-汉和汉-英翻译任务，汉语端文本在词例化之前先利用 LTP (Che et al., 2010) 中文分词工具进行分词。上述翻译任务数据经预处理后的统计信息如表 2 所示。

	维-汉	汉-英	英-德	英-法
平行句对的数目	0.17M	1.3M	4.5M	18M
源语言词表规模	32K	37K	37K	30K
目标语言词表规模	27K	33K	37K	30K

表 2: 各翻译任务训练数据的统计信息

[1]包括 LDC2005T10，LDC2003E14，LDC2004T08 以及LDC2002E18. 其中 LDC2003E14 是文档级别对齐语料，我们采用Champollion 句对齐工具 (Ma, 2006) 从中提取平行句对.

[2]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

4.2 模型设置与评价指标

基线模型设置：本文采用 Transformer（base） (Vaswani et al., 2017) 作为 NMT 模型的具体实现。参考 Vaswani et al. (2017)，训练完成后，我们选择在验证集上表现最佳的 5 个模型存储点并求输出层平均，得到最终的测试用模型。在测试时，NMT 的解码采用柱搜索（beam search）方法，柱搜索的宽度为 4。

机器翻译评价指标：机器翻译采用大小写不敏感的 BLEU (Papineni et al., 2002) 作为评价指标，本文实验采用 Moses (Koehn et al., 2007) 提供的 *multi-bleu.perl* 脚本[3]对译文进行打分。对于维-汉翻译，译文评价是在中文字符级别上进行的 BLEU 评分。

词级别译文多样性评价指标：本文从（1）生成译文中独立的 N 元文法（N-gram）的占比（记作 *Dist-N*）和（2）译文的词频分布（Mean 和 Median）两个角度评价机器翻译译文词级别多样性，其计算方法如下：

（1）*Dist-N*：生成译文中独立的 N 元文法在全部 N 元文法的总数目中的占比 (Li et al., 2016a)，其中 $N \in \{1, 2, 3, 4\}$，该评价指标的数值越高，表示译文多样性程度越高；

（2）Mean 和 Median：本文通过计算译文词汇的平均词频（记作 Mean）和词频的中位数（记作 Median）对译文的词频分布情况进行量化，该指标对应的数值越低，表示译文受到词频偏置的影响越小。本文所有的词频或概率值均以自然对数数值的形式进行展示。

4.3 机器翻译实验结果

本文提出的方法旨在不影响译文质量的情况下提升机器译文的词汇多样性。为验证该假设，本文在四个代表着不同资源丰富度的机器翻译任务上进行了实验，实验结果在表 3 和表 4 中展示。在表 4 中，“汉-英”列对应的数据表示将 NIST03~06 合并后再计算得到的 BLEU 值。从表 3 和表 4 可以看出，本文提出的 4 种方法生成的译文相对于基线模型在 BLEU 指标上均有所提升，说明本文提出的方法在不改变模型设置和训练方法的情况下，仅通过在解码时对模型的输出层降噪即可提升译文质量。

另一方面，从表 3 和表 4 可以看出，“+SADR”的方法通常能得到最佳的 BLEU 评分，证明了本文提出的自适应降噪系数的有效性，同时侧面反映了模型输出 O 和词频 F 之间的关联确实是与词频所在的区间相关的。

模型	NIST03	NIST04	NIST05	NIST06
Transformer	42.73	45.76	43.53	44.34
+SR+CDR	42.87	45.77	43.85	44.48
+PR+CDR	42.86	45.76	43.81	44.51
+SR+SADR	43.27	**45.79**	44.00	**44.52**
+PR+SADR	**43.37**	45.77	**44.02**	44.36

表 3: 汉-英翻译任务 4 组测试数据的译文 BLEU 对比

模型	汉-英	维-汉	英-德	英-法
Transformer	44.34	39.71	27.33	40.08
+SR+CDR	44.57	39.73	27.39	40.09
+PR+CDR	44.56	39.74	27.43	40.09
+SR+SADR	**44.73**	39.78	**27.61**	**40.17**
+PR+SADR	44.71	**39.83**	**27.61**	40.15

表 4: 不同方法在四个翻译任务上的生成的译文 BLEU 对比

[3] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

模型	*Dist*-*N*				词频分布	
	Dist-1	*Dist*-2	*Dist*-3	*Dist*-4	平均数	中位数
维-汉测试集	.1572	.6710	.9039	.9601	-4.8390	-7.6416
Transformer	.1452	.6460	.9026	.9663	-4.8259	-7.5277
+SR+CDR	.1468	.6517	.9046	.9669	-4.8371	-7.6088
+PR+CDR	.1468	.6520	.9051	.9671	-4.8380	-7.6088
+SR+SADR	.1503	.6560	.9063	**.9674**	-4.8421	-7.6278
+PR+SADR	**.1516**	**.6578**	**.9066**	.9673	**-4.8472**	**-7.6406**
汉-英测试集	.0697	.4544	.7758	.9058	-4.6298	-7.4397
Transformer	.0593	.3969	.7217	.8806	-4.4997	-7.0097
+SR+CDR	.0601	.4016	.7258	.8827	-4.5093	-7.0893
+PR+CDR	.0601	.4015	.7255	.8824	-4.5088	-7.0893
+SR+SADR	.0616	.4078	.7314	.8856	-4.5169	-7.1433
+PR+SADR	**.0621**	**.4113**	**.7350**	**.8880**	**-4.5218**	**-7.2065**
英-德测试集	.1522	.6565	.9072	.9727	-4.9994	-7.9051
Transformer	.1459	.6252	.8901	.9684	-4.9290	-7.6590
+SR+CDR	.1467	.6295	.8921	.9691	-4.9346	-7.6939
+PR+CDR	.1467	.6295	.8921	.9690	-4.9347	-7.6939
+SR+SADR	.1491	**.6365**	**.8950**	**.9697**	-4.9426	**-7.8249**
+PR+SADR	**.1492**	**.6365**	**.8950**	**.9697**	**-4.9428**	**-7.8249**
英-法测试集	.1055	.4878	.7811	.9117	-4.5886	-6.8775
Transformer	.1020	.4733	.7697	.9057	-4.5824	-6.7374
+SR+CDR	.1026	.4782	.7745	.9089	-4.5921	-6.8319
+PR+CDR	.1034	**.4822**	**.7781**	**.9108**	**-4.5988**	**-6.9009**
+SR+SADR	**.1040**	.4816	.7773	.9106	-4.5960	-6.8775
+PR+SADR	.1039	.4817	.7773	.9106	-4.5959	-6.8775

表 5: 不同方法生成译文的词汇多样性量化指标对比

4.4 词级别译文多样性实验结果

表 5 给出了本文提出的方法在词级别多样性评价指标下的实验结果。其中 *Dist*-*N* 表示数据中的独立 *N*-gram 的占比，是直观展示词汇多样性的指标，该指标数值越高，则表示用词越丰富；而“词频分布”则可以看出数据中的词频组成情况，其中低频词越多，可以侧面反映出词汇多样性越高。

从表 5 可以看出，在全部四个翻译任务上，本文提出的方法相对于基线模型在“*Dist*-*N*”和“词频分布”两类指标上均有所提升。在维-汉、汉-英和英-德翻译任务上，“+PR+SADR”方法在多数评价指标上效果最佳。此外“+PR”和“+SADR”方法得到的实验结果均优于采用固定降噪系数的方法（“+CDR”），说明本文提出的分区间回归和自适应的降噪系数可以很好的应对章节 2.2 所展示的问题。

4.5 不同方法解码效率对比

为展示引入本文提出的降噪方法对于 NMT 模型解码速率的影响，本节以英-德翻译任务为例，展示了不同方法下 NMT 模型的解码速率。速率的测量单位是“词/秒”，即每秒钟机器翻译模型生成词数。实验采用的软件平台是 Centos 7.5.1804+Python 3.9.2+PyTorch 1.12.1+CUDA 12.0，硬件方面，处理器为英特尔 E5-2650 v4，显卡型号为 Nvidia GTX 1080 Ti。所有实验均在硬件空闲时完成，且实验结果取 3 次重复实验的平均值。

	Transformer	+SR+CDR	+PR+CDR	+SR+SADR	+PR+SADR
解码速率（词/秒）	1366.13	1251.67	1049.10	1211.57	1047.13
解码速率比	1.00	0.92	0.77	0.89	0.77

表 6: 不同方法解码效率对比

实验结果如表 6 所示，“速度比”表示各个方法的速率与基线模型 Transformer 的解码速率的比值，以便于更好的展示解码速率之间的差异。从表 6 中可以看出，本文提出的降噪方法对 NMT 的解码速率虽有负面影响，但影响较小，具体的解码速率损失在 10% ∼ 25% 之间，另外，从表中可以看出，“+SADR” 方法的解码效率较高，且 “+SADR”在译文 BLEU 和译文词汇多样性等指标上综合表现较好，因此，从实验中得到的多个指标综合考虑，推荐使用 “+SADR” 的方法。

4.6 译文案例

本节给出了本文提出方法与基线模型的译文案例对比，如表 7 所示。该案例来自于中-英翻译任务测试数据 NIST'06，其中“源语言”表示输入的源语言句子，“参考译文1”表示数据集中四个参考译文中的一个。表 7 在各机器译文的下方给出了句子级的 BLEU 评分，并标出了译文用词的主要差异，以及差异词在训练数据中的词频的对数（$\log f_{req}(\cdot)$）。

从表 7 可以看出，本文提出的方法生成的译文与基线系统生成的译文在句子结构上是相似的，唯一区别在于对源语言单词“发表”的翻译时选词：“made” 和 “delivered”。如表 7 所示，基线系统（“Transformer”）选择了更高频的单词 “made”（做，作出）作为对“发表”的翻译，而本文提出的方法则选择了相对低频但更符合语境的单词 “delivered”（发表），而该单词也与参考译文中选择的单词一致。

方法	译文
源语言	澳大利亚 总理 霍华德 还 在 追悼 仪式 中 发表 了 讲话 。
参考译文 1	Australian Prime Minister Howard also delivered a speech at the memorial service .
Transformer	Australian Prime Minister Howard also made a speech at the memorial ceremony .
	$\log f_{req}$(“made”) = −7.11, 句子级 BLEU：59.23
+SR+CDR	Australian Prime Minister Howard also delivered a speech at the memorial ceremony .
	$\log f_{req}$(“delivered”) = −9.84, 句子级 BLEU：84.24
+PR+CDR	Australian Prime Minister Howard also delivered a speech at the memorial ceremony .
	$\log f_{req}$(“delivered”) = −9.84, 句子级 BLEU：84.24
+SR+SADR	Australian Prime Minister Howard also delivered a speech at the memorial ceremony .
	$\log f_{req}$(“delivered”) = −9.84, 句子级 BLEU：84.24
+PR+SADR	Australian Prime Minister Howard also delivered a speech at the memorial ceremony .
	$\log f_{req}$(“delivered”) = −9.84, 句子级 BLEU：84.24

表 7: 译文案例

5 相关工作

译文多样性受限是 NMT 的经典问题之一，早期对译文多样性研究主要关注句子级译文多样性，旨在最大化同一输入源文对应的众多候选译文之间的差异，并尽可能地保证译文质量 (Li et al., 2016b; Wu et al., 2020; Sun et al., 2020; Lin et al., 2022)。近年来， NMT 受训练数据偏见影响的问题逐渐受到越来越多的关注，其中代表性的问题包括 NMT 性别偏见 (Stanovsky et al., 2019; Costa-jussà et al., 2022) 和个性化翻译(Lin et al., 2021)等。关于数据偏见的重要表现形式之一是 NMT 训练数据的词频分布不平衡问题，这使得低频词翻译一直是 NMT 面临的挑战(Koehn and Knowles, 2017)。

针对上述问题，早期的研究工作主要关注引入更细的翻译单元或优化词汇表 (Sennrich et al., 2016; Lee et al., 2017; Kudo, 2018)，通过将单词拆分成成更小的单元从而减少低频词的个数，调整输入单元的频率分布。此外，通过消除 NMT 模型词向量中与词频相关的信息，也可以达到缓解词频分布影响的问题： Yang and Liu (2020) 采用 HSR (Schölkopf et al., 2016) 方法静态消除词向量中的词频信息，而 Gong et al. (2018)和 Liu et al. (2020) 则分别通过对抗训练和课程学习（curriculum learning）的方法在训练过程消除词向量中的词频信息。最近，一些方法根据目标语言词频 (Gu et al., 2020)和双语互信息 (Xu et al., 2021)通过自适应权重的损失函数来缓解这个问题。类似地 Zhang et al. (2022) 则提出了一种标记级对比学习方法，并引入了频率感知软权重来自适应地对比目标词的表示。

上述前人的工作需要在 NMT 的训练过程或训练之前介入，无法对已经存在的优化好的模型使用。而本文提出的方法无需修改 NMT 模型的架构和训练模式，并且是模型无关的，适用于目前已知的各类 NMT 模型。

6 结论

本文针对 NMT 训练时易受训练集词频分布偏置影响，从而导致模型输出译文用词多样性受限的问题，提出了一种基于半同胞回归的 NMT 模型估计概率 O 去噪方法，通过从 O 中消除与目标语言词频 F 相关的部分信息，从而实现缓解 O 受数据偏置影响的问题。上述方法是模型无关的，且从理论推导而来，具有完整的可解释性。在四种不同规模数据翻译任务的实验结果表明，本文提出的方法可以在不损坏译文质量的情况下，提升译文的用词多样性。

在未来的工作中，我们将继续继续分析模型估计概率与词汇分布的相关性与训练数据规模以及训练轮次之间的关系，即探索本文提出的方法在不同模型规模和训练数据规模下的适用性。另外，我们也将尝试为其他自然语言生成任务引入本文提出的方法。

致谢

感谢所有匿名审稿人的宝贵意见，由于会议论文的篇幅有限，无法将这些意见悉数吸纳到此版本的论文中，因此，未及时采纳的意见将呈现在未来版本的论文中。本文的工作得到了国家重点研发计划（批准号：2017YFB1002103）的资助。

参考文献

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. LTP: A Chinese language technology platform. In *Coling 2010: Demonstrations*, pages 13–16, Beijing, China, August. Coling 2010 Organizing Committee.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October. Association for Computational Linguistics.

Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2022. Interpreting gender bias in neural machine translation: Multilingual architecture matters. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11855–11863. AAAI Press.

David Demeter, Gregory Kimmel, and Doug Downey. 2020. Stolen probability: A structural weakness of neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2191–2197, Online, July.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

ChengYue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. FRAGE: frequency-agnostic word representation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1341–1352.

Andreas Grivas, Nikolay Bogoychev, and Adam Lopez. 2022. Low-rank softmax can have unargmaxable classes in theory but rarely in practice. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6738–6758, Dublin, Ireland, May.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. 2020. Token-level adaptive training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1035–1046, Online, November. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 10.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

Huan Lin, Liang Yao, Baosong Yang, Dayiheng Liu, Haibo Zhang, Weihua Luo, Degen Huang, and Jinsong Su. 2021. Towards user-driven neural machine translation. 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)*, pages 4008–4018, Online, August. Association for Computational Linguistics.

Huan Lin, Baosong Yang, Liang Yao, Dayiheng Liu, Haibo Zhang, Jun Xie, Min Zhang, and Jinsong Su. 2022. Bridging the gap between training and inference: Multi-candidate optimization for diverse neural machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2622–2632, Seattle, United States, July. Association for Computational Linguistics.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436, Online, July.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3953–3962. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

J. Pearl, M. Glymour, and N.P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. Wiley.

Bernhard Schölkopf, David W. Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. 2016. Modeling confounding by half-sibling regression. *Proc. Natl. Acad. Sci. USA*, 113(27):7391–7398.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July. Association for Computational Linguistics.

Zewei Sun, Shujian Huang, Hao-Ran Wei, Xinyu Dai, and Jiajun Chen. 2020. Generating diverse translation by manipulating multi-head attention. 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, New York, NY, USA, February 7-12, 2020*, pages 8976–8983. AAAI Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 3104–3112. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Xuanfu Wu, Yang Feng, and Chenze Shao. 2020. Generating diverse translation from model distribution with dropout. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1088–1097, Online, November. Association for Computational Linguistics.

Yangyifan Xu, Yijin Liu, Fandong Meng, Jiajun Zhang, Jinan Xu, and Jie Zhou. 2021. Bilingual mutual information based adaptive training for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 511–516, Online, August.

Zekun Yang and Tianlin Liu. 2020. Causally denoise word embeddings using half-sibling regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9426–9433, Apr.

Muyun Yang, Xixin Hu, Hao Xiong, Jiayi Wang, Yiliyaer Jiaermuhamaiti, Zhongjun He, Weihua Luo, and Shujian Huang. 2019. Ccmt 2019 machine translation evaluation report. In Shujian Huang and Kevin Knight, editors, *Machine Translation*, pages 105–128, Singapore. Springer Singapore.

Tong Zhang, Wei Ye, Baosong Yang, Long Zhang, Xingzhang Ren, Dayiheng Liu, Jinan Sun, Shikun Zhang, Haibo Zhang, and Wen Zhao. 2022. Frequency-aware contrastive learning for neural machine translation. pages 11712–11720.