

文章编号: 1003-0077(2023)11-0001-14

## 非自回归神经机器翻译综述

曹 航, 胡 驰, 肖 桐, 王成龙, 朱靖波

(东北大学 自然语言处理实验室, 辽宁 沈阳 110819)

**摘 要:** 当前的神经机器翻译系统大多采用自回归的方式进行解码, 这种串行解码的方式导致解码效率低下。与之相比, 非自回归的方式通过并行解码显著提高了推断速度, 受到研究人员的广泛关注。然而, 由于缺乏目标序列内词语间的依赖关系, 非自回归方法在性能上还存在较大差异。近些年, 有很多工作研究如何缩小非自回归机器翻译(NART)和自回归机器翻译(ART)之间的翻译质量差距, 但是目前缺少对现有方法和研究趋势的梳理。该文不仅从捕获依赖关系的角度对 NART 方法进行了详细分类和总结, 而且对 NART 研究面临的挑战进行了展望与分析, 并归纳整理了相关的论文, 还进一步根据方法、发表会议和任务等进行了分类。

**关键词:** 自然语言处理; 非自回归; 机器翻译

**中图分类号:** TP391

**文献标识码:** A

## A Survey of Non-Autoregressive Neural Machine Translation

CAO Hang, HU Chi, XIAO Tong, WANG Chenglong, ZHU Jingbo

(NLP Laboratory, Northeastern University, Shenyang, Liaoning 110819, China)

**Abstract:** Most of the current machine translation systems adopt the autoregressive method for decoding, which leads to low inference efficiency. The non-autoregressive method significantly improves the inference speed through parallel decoding, attracting increasing research interest. We conduct a systematic survey for recent efforts to narrow the translation quality gap between Non-Autoregressive Machine Translation (NART) and Autoregressive Machine Translation (ART). We categorize NART methods by the way to capture the dependencies of target sequences. We also discuss the challenges of NART research.

**Keywords:** natural language processing; non-autoregressive method; machine translation

## 0 引言

机器翻译是自然语言处理中的重要分支, 其目标是把一种自然语言(源语言)自动翻译成另一种自然语言(目标语言)。随着深度学习技术的发展, 神经机器翻译逐渐成为了主流范式, 它通过基于编码器-解码器的框架实现了语言之间的转化, 在性能上表现出了显著的优越性。大多数神经机器翻译模型采用自回归神经机器翻译(Autoregressive Neural Machine Translation, ART)模型从左到右逐词生成译文, 这使得解码的速度比较慢, 尽管有一些工作提出的模型如 CNN<sup>[1]</sup>、Transformer<sup>[2]</sup>等加快了训

练阶段的速度, 但是自回归的特性仍然限制了模型在推断阶段的并行化。为了缓解这个问题, Gu 等人<sup>[3]</sup>提出了 Non-Autoregressive Transformer (NAT) 的概念, 用非自回归神经机器翻译(Non-Autoregressive Neural Machine Translation, NART)模型并行生成完整目标语言, 并明显加快了推断速度。然而, 由于 NART 没有考虑目标语言句子中各个词汇之间的依赖关系, 导致在翻译质量方面与 ART 相比有很大差距。

近年来, 很多研究人员被 NART 的潜力吸引, 提出了一系列的工作<sup>[4-7]</sup>提高 NART 的生成质量。但目前仍然缺少对 NART 现有方法的全面梳理和总结。在我们的调查中, 一部分工作只对 NART 的

收稿日期: 2022-07-14 定稿日期: 2022-09-15

**基金项目:** 国家自然科学基金(61876035, 61732005); 科技部科技创新 2030-“新一代人工智能”重大项目(2020AAA0107904); 云南省科技厅科技计划项目(202002AD080001, 202103AA080015); 中央高校基本科研业务专项资金资助

概念进行了介绍<sup>[8]</sup>。而与本文工作相似的现有综述<sup>[9]</sup>把 NART 流程模块化,根据每个模块进行介绍,并对非自回归在其他任务中的应用进行了扩展。但是他们缺少对不同方法优缺点的分析。而我们主要从捕获目标序列之间依赖关系的角度对 NART 方法进行了分类,侧重于对非自回归机器翻译方法的介绍,总结了相应方法的优缺点,并增加了最新的相关论文。同时,为了帮助研究人员快速查阅资料,了解该领域的研究进展,本文整合了收集到的非自回归相关论文,并按照时间、方法以及会议进行了仔细的标注<sup>①</sup>。

本文的组织结构如下:第 1 节简要介绍 ART 和 NART 的背景;第 2 节总结分析 NART 的研究进展;第 3 节展望 NART 未来的研究方向;第 4 节总结全文。

## 1 背景

本节主要介绍了 ART 和 NART 的概念,分析了它们的特点和问题,并对提升 NART 翻译质量所面临的挑战进行了阐述。

### 1.1 ART 的概念及其问题分析

ART 指的是按照从左到右的顺序逐字生成译文,如图 1(a)所示。具体地,给定一个输入的源语言序列  $X = \{x_1, \dots, x_m\}$ ,使用 ART 模型从左到右地生成输出译文序列  $Y = \{y_1, \dots, y_n\}$  的概率表示如式(1)所示。

$$P(y | X) = \prod_{i=1}^n P(y_i | y_{<i}, X) \quad (1)$$

其中,  $y_i$  表示当前的输出译文单词,  $y_{<i} = \{y_1, \dots, y_{i-1}\}$  是之前输出的所有译文。这个模型就表示当前时刻的输出依赖于前面所有时刻的输出以及输入的源语言句子。

尽管这种自回归的解码方式在很多任务上取得了成功,但存在一些结构上的问题:

(1) 存在错误累积的特性,即前面产生的错误会随着句子生成往后传递,翻译质量常常随着句子的长度增加而下降。

(2) ART 模型的自回归性使得只能串行生成译文,导致解码效率较低。

### 1.2 NART 的概念及其问题分析

NART 移除了翻译的自回归性,对每个目标词

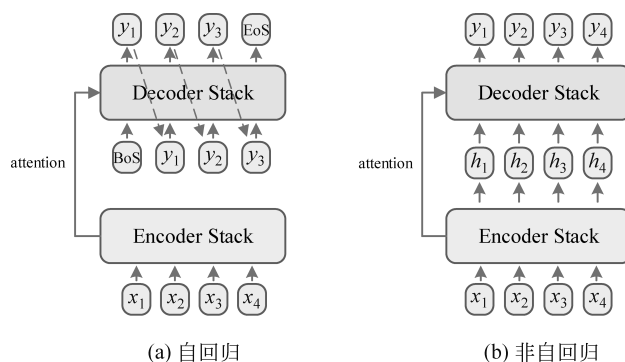


图 1 ART 和 NART 的解码对比图

的生成进行独立建模,不再依赖之前的译文输出,解决了 ART 模型结构存在的问题,如图 1(b)所示。对 NART 模型的译文概率建模如式(2)所示。

$$P(y | X) = \prod_{i=1}^n P(y_i | X) \quad (2)$$

其中,  $y_i$  表示当前的输出译文单词。显然,在 NART 模型中当前位置的输出只依赖于输入的源语言句子,与其他位置的输出无关。

这种非自回归的解码方式使得翻译的速度成倍提升,但是由于假设独立性过强,忽视掉了目标句子间的依赖关系,导致生成的句子重复或者缺失,使得翻译的质量大幅下降。

### 1.3 NART 模型面临的挑战

根据以上概念,NART 模型只能依靠源语端的信息来生成目标语,这大幅增加了模型的训练和推断难度。所以与 ART 模型相比,NART 模型仍然面临以下的主要挑战。

(1) 如何确定解码端的输入。ART 模型的解码端输入来自于上一步解码出的结果,然而 NART 模型缺少了这样的依赖性,所以在 NART 模型中的一个关键问题是如何确定解码端的输入。

(2) 如何确定目标序列的长度。ART 模型中可以根据输出的截止符(<EOS>)来动态决定译文的长度,但是 NART 模型并行地生成所有译文,所以就需要提前知道将要生成的目标句子长度。

(3) 如何解决多峰问题。现有的 NART 模型与 ART 模型相比,翻译质量还有很大差距的一个主要原因是存在多峰问题。所谓的多峰问题是在给定原文时,用于参考的目标词的概率具有多峰性,更通俗的来讲就是一个源语对应多个意思相同的目标

① 地址为: <https://github.com/LittleRooki/NAT-survey>

语。一个简单汉译英的例子,源语“干得好!”可以翻译成“Good job!”或者“Well done!”,ART 模型因为是顺序生成,所以很容易获取目标语之间的依赖关系,得到正确译文。但是由于 NART 模型的条件独立假设,第一个位置生成的词并不会影响第二个位置生成词的概率,这就会使得模型生成“Good done !”或者“Well job !”这样的错误译文。

## 2 方法分类

自从 NART 的概念被提出后,各种 NART 模型层出不穷,它们的研究趋势是在保持速度优势的同时获得与 ART 模型相当甚至更好的翻译质量。

提升 NART 翻译质量的方法主要包括:①改善训练数据的方法;②基于迭代的方法;③引入隐变量的方法;④改进训练目标的方法;⑤用自回归模型辅助非自回归模型的方法;⑥其他改进方法。它们从不同的角度来学习目标序列中缺失的依赖关系。具体地,改善训练数据的方法从知识蒸馏的角度减少训练数据的复杂度,基于迭代的方法采用多次迭代生成的方式优化输出,引入隐变量的方法通过隐变量来建模目标序列之间的关系,改进训练目标的方法是去找到一个更适合的损失函数,用自回归模型辅助的方法的主要思路是转移自回归机器翻译的知识,其他改进方法从半自回归、训练策略以及其他方面来增强 NART 性能,具体如图 2 所示。

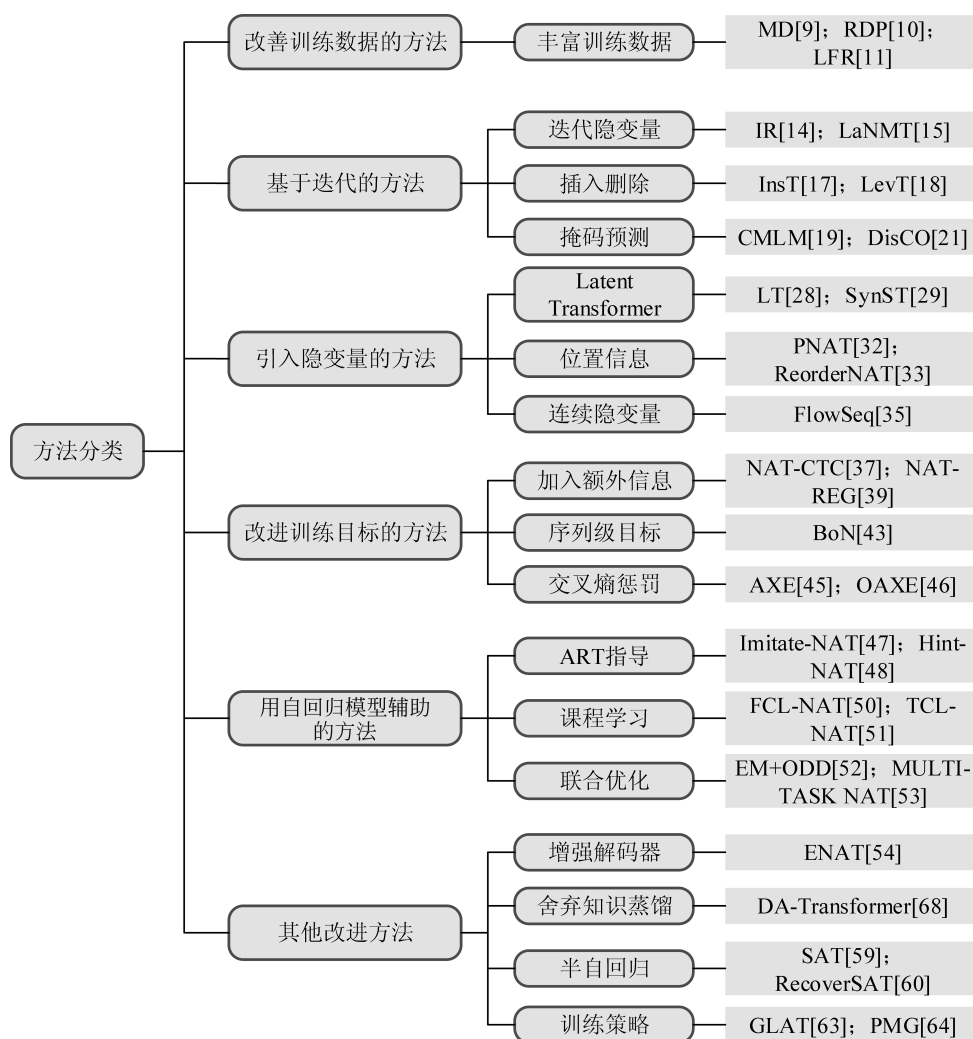


图 2 NART 模型的方法框架图

接下来将详细介绍上述方法并进一步分析和探讨当前的研究和技术进展。

### 2.1 改善训练数据的方法

提升 NART 模型翻译质量最直接的方法就是

使用序列级的知识蒸馏<sup>[10]</sup>。具体是使用原始数据  $(X, Y)$  训练一个 ART 模型,再通过训练后的 ART 模型对原始数据的源语  $X$  重新生成目标语  $Y'$ ,最后将原始数据的目标语替换成 ART 模型的输出得到新的训练数据  $(X, Y')$ 。这种方法的好处是减少

了训练数据中的噪声和不确定性,并有效缓解了译文的多峰问题,从而达到降低 NART 模型学习难度的目的。

**知识蒸馏的作用** 为了更深层次地探索序列级知识蒸馏在 NART 中发挥的作用,Zhou 等人<sup>[11]</sup>首先可视化不同模型的合成数据集,证明了知识蒸馏可以减少数据集中的模式,并且表明 ART 模型生成的数据更易于 NART 模型学习。他们还通过复杂度和忠实度两种指标来衡量训练集,并以此分析训练数据的复杂性和 NART 模型性能之间的相关性。实验表明,虽然知识蒸馏提高了 NART 模型的性能,但并不是 ART 模型越大、性能越好就越有益于 NART 模型,而是需要让数据的复杂性和 NART 模型的容量相匹配才可以实现最佳性能。Xu 等人<sup>[12]</sup>认为不同类型的复杂性具有不同的影响,他们发现知识蒸馏降低了词汇多样性和重排序复杂性,这可以帮助 NART 模型更好地学习源语和目标语之间的对齐。

**丰富训练数据** 在知识蒸馏的启发下,有很多工作从训练数据的层面对 NART 进行改进。Zhou 等人<sup>[13]</sup>增加了一个单语语料库来让 ART 模型生成额外的双语数据,用于 NART 模型的训练,并在实验中证明了这可以帮助 NART 更好地泛化。Ding 等人<sup>[14]</sup>从词汇选择的角度探索,发现知识蒸馏会传播低频词汇选择错误,因此他们提出用 KL 散度把原始数据先验集成到 NART 训练中来恢复低频词的有用信息。Ding 等人<sup>[15]</sup>对他们上一个方法做出了进一步改进,通过原始数据预训练和反向蒸馏的方式充分利用平行语料来恢复低频词汇。Shao 等人<sup>[16]</sup>认为只学习单一的参考译文会限制 NART 模型的潜力,所以提出了具有可选择参考译文的多样化蒸馏。他们利用随机种子来让模型生成多样化的

译文,得到具有多个参考译文的数据集,让 NART 模型从多个参考译文中选择性地学习,从而提高了模型性能的上限。Huang 等人<sup>[17]</sup>通过数据集的条件总相关性来衡量目标语词汇之间丢失的依赖信息,并采用一个统一的视角揭示 NART 实际是最大化代理分布的可能性,从而减少 NART 学习中的信息丢失。

## 2.2 基于迭代的方法

随着 NART 模型的不断更新,一类基于迭代的解码方法进入了研究人员的视野。该类方法的主要思路是,先用非自回归的方法解码出比较差的初始译文,然后再把初始译文作为解码器的输入得到第二次的译文,通过多次迭代让模型对之前的错误进行修改,来达到提高翻译质量的目的。为了更清晰地了解基于迭代方法的脉络,我们进一步根据迭代的操作方式把该类方法划分为了四种。

**迭代隐变量** Lee 等人<sup>[4]</sup>开创性地提出了基于迭代的 NART 模型,如图 3 所示。该工作把隐变量模型解释为一个迭代精化的过程,首先 Decoder1 复制源语作为第一次迭代的输入(类似于 Gu 等人<sup>[3]</sup>的工作),然后再将生成的译文的词嵌入向量和最终激活向量序列在对应位置相加作为 Decoder2 的输入,最后重复多次迭代,直到满足停止标准。Shu 等人<sup>[5]</sup>以 Lee 等人<sup>[4]</sup>的研究为基础,提出使用连续的隐变量来捕捉目标句子中的依赖关系,并在隐变量上加入 Delta 后验来增加目标词预测的确定性,最后通过更新隐变量上的 Delta 后验来完成迭代精化。Lee 等人<sup>[18]</sup>优化了他们的上一个模型<sup>[5]</sup>,在连续空间中使用估计的梯度替代 Delta 推理完成隐变量的迭代更新。

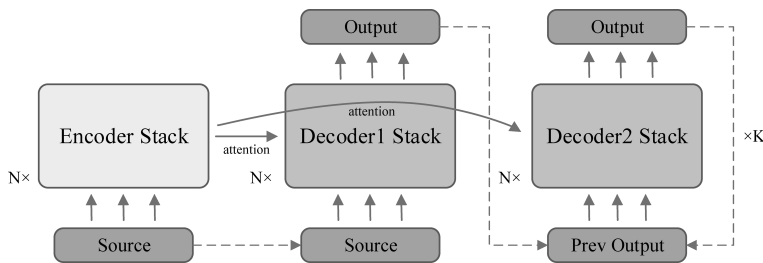


图 3 Lee 等人文中的模型结构图

**插入删除** Stern 等人<sup>[19]</sup>提出了基于插入操作(Insertion)的迭代方法,他们将需要插入的目标词和插入的位置联合建模作为一次插入操作,通过对

初始为空的输出序列进行重复插入操作,使得输出序列可以动态调整,从而免去了提前预测目标序列长度的步骤。为了实现这种模型,Stern 等人把左

右相邻的向量拼接起来得到对应的槽表示,并将所有的槽表示作为解码器的输出。Gu 等人<sup>[20]</sup>使用插入操作(Insertion)、删除操作(Deletion)两个基本操作来进行解码。插入操作是在单词之间插入占位符,然后用预测的实际字符来替换占位符。而相比于 Stern 等人<sup>[19]</sup>的工作,增加的删除操作使得模型可以修改之前插入的错误单词。

**掩码预测** Ghazvininejad 等人<sup>[6]</sup>参考了 Devlin 等人<sup>[21]</sup>的研究,把条件掩码语言模型(Conditional Masked Language Models, CMLM)和并行解码相结合。该工作在解码时分为掩码和预测两步,首先掩码掉整个句子,再并行预测所有单词,随着迭代的增加而线性减少掩码低置信度单词的数量,让模型可以在双向上下文中选择生成更好的译文。Kasai 等人<sup>[7]</sup>对掩码预测(Mask-Predict)方法<sup>[6]</sup>进行改进,提出了简单优先(Easy-First)推理方法,该方法会把所有生成的译文单词按照概率降序排列,根据概率更大的单词来对所有单词进行修改,减少了所需要的迭代次数。Ghazvininejad 等人<sup>[22]</sup>为掩码预测的解码方法设计了更加贴合的训练方法,它模仿解码的过程,对目标序列随机掩码,把预测出来的目标序列再次掩码作为训练数据的一部分。Guo 等人<sup>[23]</sup>通过实验发现 NART 模型中的编码器对翻译质量的影响更大,因此他们用 BERT<sup>[21]</sup>中的策略对编码器的输入做掩码操作,再让编码器使用掩码之后的输入编码完整的输入信息,使得对编码器的训练更加严格。这种方法在推断算法和 Ghazvininejad 等人<sup>[6]</sup>相同的情况下,获得了更好的效果。Ding 等人<sup>[24]</sup>在迭代 NART 模型中增加了一个局部交叉注意力模块,对于目标词是在注意力权重最高的源词周围以一个窗口的范围做局部交叉注意力,然后通过一个门控机制把它和原始交叉注意力结合起来。Norouzi 等人<sup>[25]</sup>提出了适用于迭代模型的蒸馏算法,让学生模型在一次解码过程中学习到教师模型(预训练的迭代 NART 模型)的多个迭代行为,在保留迭代 NART 模型翻译质量的同时还减少了迭代的次数。Geng 等人<sup>[26]</sup>针对 CMLM 算法识别错误词的问题,提出了重写机制来明确学习重写错误的翻译片段。重写包括定位和修改两部分,首先利用定位器模块定位错误词汇,再对其进行掩码和修改,明确的错误分类提升了解码速度。

**优缺点分析** 该方法的优点是,通过迭代的方式对译文重复修改,大幅地提高了非自回归翻译的质量,其中的一些方法逼近甚至超过了自回归的解

码方法。缺点是,这种质量提升让 NART 模型牺牲了应有的速度优势,因为使用足够深的编码器和单层解码器的 ART 模型可以在保证译文质量不下降的情况下拥有与它相当的速度<sup>[27]</sup>。

### 2.3 引入隐变量的方法

引入隐变量的方法是当前 NART 的一种主流方法,它通过引入一个隐变量来建模 NART 模型中目标句子间缺失的条件依赖关系,然后再根据隐变量并行地生成译文,具体如式(3)所示。

$$P(y | X) = \int_z P(z | X) \prod_{i=1}^M P(y_i | z, X) dz \quad (3)$$

其中隐变量  $z$  可以理解作为一种额外信息。

Gu 等人<sup>[3]</sup>首次将非自回归模型应用在机器翻译中,并提出引入隐变量的方法来缓解多峰问题,如图 4 所示。Gu 等人把繁衍率(Fertility)建模为隐变量,然后把源语单词按照 Fertility Predictor 生成的对应整数值进行复制,作为解码端的输入,而目标序列的长度就是繁衍率生成的所有整数之和,不需要额外的方法来建模。为了进一步解决训练数据中的不确定问题,他们采用了序列级知识蒸馏<sup>[9]</sup>的方法,用自回归模型的输出替换训练数据中的目标语句,减少了训练数据中的噪声,使得非自回归模型能够更容易去学习。在解码时,使用 Noisy Parallel Decoding(NPD)算法,其做法是采样多个繁衍率序列并对每个序列生成的译文打分。除此之外,我们把使用隐变量的方法总结为以下的类别。

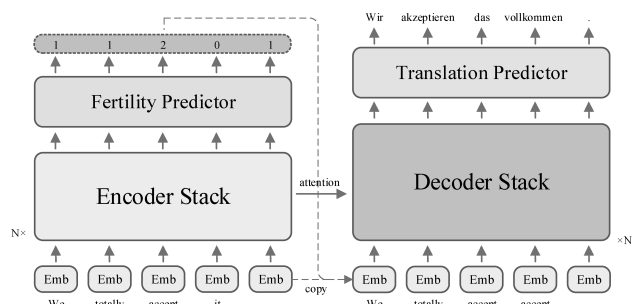


图 4 Gu 等人文中的模型结构图

**采用端到端的隐变量框架** Kaiser<sup>[28]</sup>等人提出了端到端的 Latent Transformer(LT)模型,该工作首先把目标序列自编码为一个短的离散隐变量序列,然后在推断时根据源语自回归地生成离散隐变量序列,最后再从该序列并行生成译文。Akoury 等人<sup>[29]</sup>从 LT 模型出发,对隐变量空间进行句法信息监督学习。他们在遵循 LT 模型两阶段解码设置

的情况下,把短语块序列作为离散隐变量序列。Bao 等人<sup>[30]</sup>受到 Kaiser 等人<sup>[28]</sup>和 Akoury 等人<sup>[29]</sup>工作的启发,在 LT 模型框架下,把目标词隐含的分类信息建模为隐变量,从而弥补了缺失的依赖关系并提高了模型容量。Roy 等人<sup>[31]</sup>基于向量量化变分自编码器(VQ-VAE)提出学习离散隐变量的训练技术,使用最近邻搜索来训练模型并利用其与期望最大化算法的联系进一步提升性能,实现了与 ART 模型相近的效果。

**建模位置信息** Bao 等人<sup>[32]</sup>认为位置信息在 NART 中至关重要,他们将位置信息建模为隐变量,在机器翻译任务上取得了明显的效果提升。Ran 等人<sup>[33]</sup>把重排序信息建模为隐变量,在编码器和解码器之间增加了一个重排序模块,重排序模块的作用是提供重排序信息,帮助解码器捕获目标序列之间的依赖关系。Song 等人<sup>[34]</sup>显式建模单词对齐信息为隐变量,利用对齐信息降低目标分布的模态。为了引导解码器专注于简单的一对一翻译,他们通过 Aligner 模块来估计对齐并生成解码器的输入。

**建模连续隐变量** 除了基于离散变量的工作,建模连续隐变量也是一个主要的探索方向。Ma 等人<sup>[35]</sup>提出了基于流的模型,用生成流(Generative Flow)<sup>[36]</sup>对隐变量的条件密度建模,同时这个框架也可以更好地建模复杂分布。具体来说,他们从当前的后验中抽取隐变量  $z$ ,再把  $z$  和源编码一起作为解码器的输入。与之前方法不同的是,长度预测是通过分类器在一个范围内预测源语和目标语的长度差异。

**优缺点分析** 该方法的优点是,通过引入额外的信息进行建模,模型能够更好地捕获依赖关系,减小了译文的搜索空间。该方法的缺点是,对隐变量的选取以及建模并不是一项简单的工作,而且建模隐变量会增加额外的计算代价并降低模型的翻译速度。

## 2.4 改进训练目标的方法

训练目标通常不会增加解码的次数,所以也是完全 NART 探索的主要方向。之前的 NART 模型大多数都是沿用 ART 模型的训练目标使用交叉熵进行训练,但是对于缺乏自回归性的 NART 模型来说,要求预测的每个词都与目标词位置对应的交叉熵并不是最合适的,因此有很多方法开始尝试改进训练目标来提高模型性能。

**加入额外信息** 现有的一些工作通过增加模块、正则化项和语法信息辅助 NART 模型进行训练

和解码,其本质上可以看作是利用额外的信息来指导 NART 模型学习。Libovicky 和 Helcl<sup>[37]</sup>将翻译看作是一个序列标注问题,考虑到目标语可能比源语更长,他们将输入编码后的隐藏状态进行线性投影,再把投影后的每个状态按照投影的倍数分割成多个向量,最后采用 CTC 损失训练模型。Saharia 等人<sup>[38]</sup>进一步地提出了 Imputer 模型让预测序列与目标序列构建映射对齐,在解码时是对不同迭代间的依赖关系进行建模。Wang 等人<sup>[39]</sup>将 NART 的质量下降归结于低质量的解码器隐藏状态,在解码器的训练过程中加入了两个正则化项。相似度正则化(Similarity Regularization)让相邻隐藏状态的相似度和对应相邻目标词的相似度对齐,而重构正则化(Reconstruction Regularization)让隐藏状态重新生成源语,有效提高了解码器隐藏表示,从而缓解了重复翻译和翻译不全的问题。Sun 等人<sup>[40]</sup>的想法和 CTC 的工作<sup>[37]</sup>类似,也是将机器翻译视为序列标注问题,区别是他们在 NART 模型中加入了 CRF 模块,把简单的填充符( $\langle\text{pad}\rangle$ )和截止符( $\langle\text{eos}\rangle$ )作为解码器的输入,并使用 CRF 建模输出的多峰目标分布。Liu 等人<sup>[41]</sup>认为语法和语义结构信息对 NART 模型的影响至关重要,他们把连续文本和结构标签作为输入,根据预测的结构标签得到译文,为了捕获目标语之间的依赖信息,他们还考虑将中间解码器和目标语进行对齐。

**使用序列级目标** Shao 等人<sup>[42]</sup>把强化算法应用在 NART 模型中,用序列级目标替换传统的交叉熵进行序列级训练,并把序列信息融入 Decoder 的顶层来引导目标语句的生成。Shao 等人<sup>[43]</sup>认为词级的交叉熵损失与翻译质量相关性低,会过度纠正 NART 在对应位置生成的目标词,而不考虑全局正确性,针对这个问题,他们提出把缩小模型输出和目标句子的 Bag-of-Ngrams (BoN) 差异作为训练目标,以此来让模型捕获目标端的依赖关系。Shao 等人<sup>[44]</sup>在之前的工作基础上,应用三阶段训练策略将强化算法和 BoN 两种方法结合起来,进一步提升了 NART 模型的性能。

**减轻词顺序惩罚** Ghazvininejad 等人<sup>[45]</sup>从减轻交叉熵对词顺序错误的惩罚角度提出了对齐交叉熵损失(Aligned Cross Entropy, AXE),它采用动态规划的方式来单调对齐模型预测输出和目标句子,把惩罚集中在错误的词上而不是词的绝对顺序。Du 等人<sup>[46]</sup>进一步提出的顺序无关交叉熵(Order-Agnostic Cross Entropy, OAXE),消除了对词顺序

错误的惩罚,他们使用 Hungarian Algorithm 找到模型预测和目标句子的最佳对齐,相比于交叉熵和 AXE, OAXE 更关注词汇的匹配,与 NART 的并行解码也更兼容。不同交叉熵的对比如图 5 所示。

**优缺点分析** 该方法的优点是,对训练目标的改进只会出现在训练过程中,而不会对推断算法进

行修改,所以不会为推断增加额外的计算代价,相比于迭代的 NART 模型,这种方法在大幅提高翻译质量的同时还保留了 NART 模型的主要优势。该方法的缺点是,设计的训练目标并不是简单的替换,通常还需要借助外部工具或者高效的实现算法才能获得好的效果。

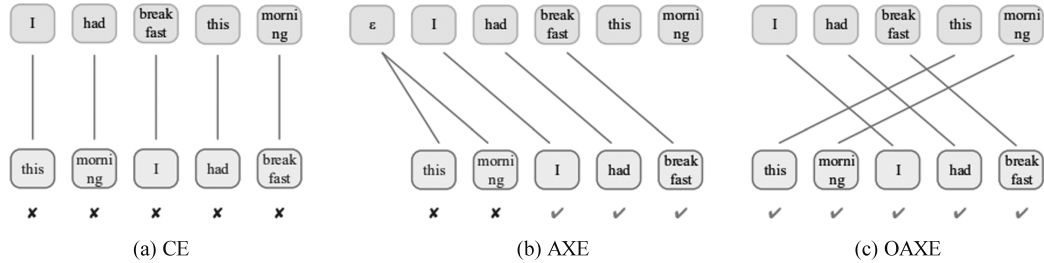


图 5 Du 等人文中的交叉熵对比图

## 2.5 用自回归模型辅助的方法

由于 ART 模型和 NART 模型的结构相似,并且 ART 模型的性能更好,所以一些工作尝试利用 ART 模型辅助训练 NART 模型,让 NART 学习 ART 模型的输出分布,以此把知识转移到 NART 中,从而在只需要解码一次的情况下提高 NART 模型的性能。

**ART 模型进行指导** Wei 等人<sup>[47]</sup>直接引入模仿学习来提升 NART 的性能,如图 6 所示。他们把 ART 模型生成的动作序列分布作为监督信号,对 NART 模型不同层和不同时间步的每个解码状态进行监督,逐步学习 ART 模型的行为,有效缓解了搜索空间大的问题。Li 等人<sup>[48]</sup>提出了基于提示的方法,从 ART 模型中设计隐藏状态和词对齐的两

种提示,让 NART 模型学习隐藏状态的差异和注意力分布,以此帮助 NART 模型进行训练。Tu 等人<sup>[49]</sup>将 NART 模型视为推断网络,并给出了多种方法放松 ART 模型的能量(Energy),让模型以最小化 ART 模型的能量进行训练,这种方法充分结合了自回归能量的有效性和非自回归的推理速度。

**结合课程学习** Guo 等人<sup>[50]</sup>将解码器输入和注意力掩码(Attention Mask)作为课程,通过课程学习微调的策略让模型完成从简单的 ART 模型到困难的 NART 模型的训练过程,从而将 ART 模型强大的语言建模能力转移到 NART 模型中。Liu 等人<sup>[51]</sup>提出了任务级的课程学习,把半自回归翻译(SAT)任务作为课程,利用一个超参数让模型训练由 ART 任务先转移到 SAT 任务,最后再过渡到 NART 任务,他们还尝试用一个任务窗口去减少每个阶段之间的差距,让整个过程更加平滑。

**联合优化** Sun 等人<sup>[52]</sup>认为以往只在训练数据集上做一次知识蒸馏不足以优化 NART 模型,因此他们提出期望最大化(EM)的方法,让 ART 模型学习 NART 模型的输出,再把更新后的 ART 模型的输出用来训练 NART 模型,通过这种迭代联合优化 ART 和 NART 模型。Hao 等人<sup>[53]</sup>将 ART 模型作为辅助任务增强 NART 模型,提出在多任务学习的框架下共享 ART 模型和 NART 模型的编码器而分离解码器,并在训练过程中采用动态的方式给 ART 任务分配权重。

**优缺点分析** 该方法的优点是,由 ART 模型过渡到 NART 模型降低了 NART 模型的训练难

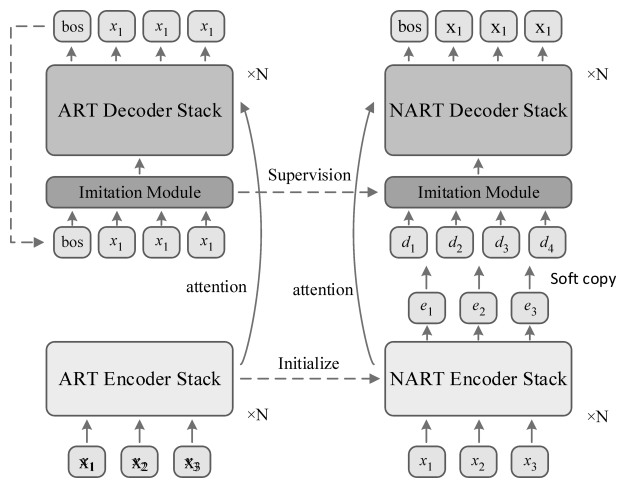


图 6 Wei 等人文中的模型结构图

度,可以有效结合 ART 模型的翻译质量和 NART 模型的推断速度。该方法的缺点是,需要训练 ART 模型作为指导,这增加了额外的训练时间,同时也让模型更为复杂。

## 2.6 其他改进方法

除了上面主流的方法之外,还有很多方法尝试从数据、训练策略和半自回归等角度提高 NART 的翻译质量,并且都取得了不错的结果。

**增强解码器** Guo 等人<sup>[54]</sup>提出了两种方法增强解码器的输入,一种是基于短语表的方法,直接将源语变成目标语作为解码器输入,但由于短语表是预训练生成的,所以无法更新。另一种是嵌入映射的方法,把源语的嵌入线性映射到目标语嵌入空间,用对抗训练的思想学习映射。Zhan 等人<sup>[55]</sup>在输出嵌入矩阵上应用注意力机制,将原始解码器输入从源语言空间转换到目标语空间,再从目标语空间重新组装解码器输入。Li 等人<sup>[56]</sup>提出词汇注意(Vocabulary Attention, VA)机制,让每个位置都关注词表的所有词来建模长期依赖关系,再利用环顾(Look-Around, LA)策略,预测左右相邻的词以生成当前位置的词,避免翻译成相同的词。Deng 等人<sup>[57]</sup>使用搜索网格替换解码器,该方法首先基于源语构建目标语候选网格,再根据编码器对网格进行评分,最后采用动态规划寻找最佳路径,在保持非自回归性的同时降低了深解码器的计算成本。Zheng 等人<sup>[58]</sup>设计了没有显式编码器和解码器区别的可逆双工 Transformer (Reversible Duplex Transformer),通过非自回归建模解决了编码器与解码器的异构,在实现可逆机器翻译的同时,也让 NART 的性能接近 ART 模型。

**采用半自回归** 半自回归是自回归和非自回归的一种过渡状态,其对速度和质量进行了权衡。Wang 等人<sup>[59]</sup>首次提出了半自回归的方法,这种方法在全局保持自回归性,而在每个时间步并行生成多个连续的词汇。Ran 等人<sup>[60]</sup>提出了 RecoverSAT 模型,与 Wang 等人<sup>[59]</sup>工作不同的是,他们让模型并行生成多个片段,而在片段内是用自回归生成每个词的,这有利于减少重复词。为了从重复翻译和翻译缺失的错误中恢复,他们在模型中引入了动态终止机制和段删除机制来确定段的长度以及删除重复的片段,有效缓解了多峰问题。Kong 等人<sup>[61]</sup>通过在每个目标解码位置使用自回归的方式预测一个词汇序列来捕获输出之间的局部依赖关系,再通过

增量对齐和合并相邻的片段得到最终译文。Deng 等人<sup>[62]</sup>对 ART 模型进行扩展,提出了马尔可夫 Transformer 模型,可以对特定的自回归上下文进行解码。同时,在给定正确的参数的情况下,可以探索不同程度的自回归依赖关系,实现翻译质量和翻译速度的平衡。

**改进训练策略** Qian 等人<sup>[63]</sup>提出了一种自适应的扫描采样策略,类似于课程学习,它在训练开始阶段采样很多词并给予参考目标词的信息去学习预测少部分未选择的词,而随着模型不断变好,采样的词逐渐减少,模型逐渐学习一次性地生成整个句子。Ding 等人<sup>[64]</sup>认为 NART 模型更容易学习细粒度的知识,因此他们将训练数据划分为单词、短语和句子三种粒度,提出从细粒度到粗粒度的训练策略,引导 NART 模型逐步学习。Xie 等人<sup>[65]</sup>提出了一致性训练的方法,强制同一个目标句子在不同掩码方式或者不同模型下对于相同掩码位置的预测一致,即掩码一致性和模型一致性,从这两个方面继续开发基于 CMLM 的 NART 模型的潜力。Zeng 等人<sup>[66]</sup>采用对齐约束训练的方式,利用约束对齐把源语端上下文翻译为目标端的上下文。从而缓解了 NART 模型不能很好处理低频约束的问题。

**不依赖知识蒸馏** 现在的 NART 模型都依赖知识蒸馏来生成训练数据,也就是需要预先训练一个 ART 模型,这使得训练过程变得复杂且影响了 NART 的实际应用。同时,让 NART 学习 ART 模型会给 NART 增加一个上限,限制了 NART 的潜力。Bao 等人<sup>[67]</sup>引入一组离散隐变量来捕获目标词汇的分类信息,并让原始目标进行隐变量建模和句子重构,这样所得到的分类信息简化了原始目标语中的多峰现象,从而不依赖于知识蒸馏进行学习。Huang 等人<sup>[68]</sup>通过有向无环解码器直接捕获多种翻译模态,避免了知识蒸馏,并将隐藏状态组装成有向无环图,沿预测路径生成译文,在不依赖知识蒸馏的情况下获得了与 ART 模型相近的性能。Huang 等人<sup>[69]</sup>通过暴露位置编码来区分相邻的词汇表示,并加入校正损失让模型纠正初始的翻译错误两种措施,使得 NART 模型可以在没有知识蒸馏的情况下充分利用原始数据。

**融合有效方法** Gu 等人<sup>[70]</sup>以减少目标之间的依赖关系作为指导,在训练数据、模型架构、训练目标和训练策略四个方面把现有的有效方法结合起来,通过利用这些技术的优势,缩小了 NART 模型和 ART 模型之间的性能差距。

现有的 NART 模型性能对比如表 1 所示。

表 1 NART 模型的性能(BLEU)对比

| 类别          | 模型             | Iteration          | Speedup | WMT14 |       | WMT16 |       |
|-------------|----------------|--------------------|---------|-------|-------|-------|-------|
|             |                |                    |         | EN-DE | DE-EN | EN-RO | RO-EN |
| 改善训练数据的方法   | MD             | 1                  | —       | 25.73 | 30.18 | 31.96 | 33.57 |
|             | RDP            | 2.5                | 3.5x    | 27.8  | —     | —     | 33.8  |
|             | LFR            | 2.5                | 3.5x    | 28.2  | —     | —     | 33.8  |
|             | DDRS           | 1                  | 5.0x    | 28.33 | 32.43 | 35.42 | 35.81 |
| 引入隐变量的方法    | NAT-FT         | 1                  | 15.6 x  | 17.69 | 21.47 | 27.29 | 29.06 |
|             | LT             | 1                  | 3.89x   | 19.80 | —     | —     | —     |
|             | SynST          | 1                  | 4.86x   | 20.74 | 25.50 | —     | —     |
|             | PNAT           | 1                  | 7.3x    | 23.05 | 27.18 | —     | —     |
|             | Flowseq        | 1                  | 1.1x    | 23.72 | 28.39 | 29.73 | 30.72 |
|             | ReorderNAT     | 1                  | 16.1x   | 22.79 | 27.28 | 29.30 | 29.50 |
|             | CNAT           | 1                  | 10.4x   | 25.56 | 29.36 | —     | —     |
| 基于迭代的方法     | IR-NAT         | 10                 | 1.5x    | 21.61 | 25.48 | 29.32 | 30.19 |
|             | InsT           | $\approx \log_2 n$ | 4.8x    | 27.41 | —     | —     | —     |
|             | CMLM           | 10                 | 1.7x    | 27.03 | 30.53 | 33.08 | 33.31 |
|             | LevT           | Adaptive           | 4.0x    | 27.27 | —     | —     | 33.26 |
|             | LaNMT          | 4                  | 5.7x    | 26.30 | —     | —     | 29.10 |
|             | DisCO          | Adaptive           | 3.5x    | 27.34 | 31.31 | 33.22 | 33.25 |
|             | SMART          | 10                 | 1.7x    | 27.65 | 31.27 | —     | —     |
|             | JM-NAT         | 10                 | 5.73x   | 27.69 | 32.24 | 33.52 | 33.72 |
|             | Imputer        | 8                  | 3.9x    | 28.20 | 31.80 | 34.40 | 34.10 |
| 改进训练目标的方法   | NAT-CTC        | 1                  | —       | 16.56 | 18.64 | 19.54 | 24.67 |
|             | NAT-REG        | 1                  | 27.6x   | 20.65 | 24.77 | —     | —     |
|             | NAT-DCRF       | 1                  | 10.4x   | 23.44 | 27.22 | —     | —     |
|             | Reinforce-NAT  | 1                  | 3.73x   | 22.27 | 27.25 | 30.57 | 30.83 |
|             | BoN            | 1                  | 10.1x   | 20.90 | 24.61 | 28.31 | 29.29 |
|             | AXE            | 1                  | 15.3x   | 23.53 | 27.90 | 30.75 | 31.54 |
|             | OAXE           | 1                  | 15.3x   | 26.10 | 30.20 | 32.40 | 33.30 |
| 用自回归模型辅助的方法 | Imitate-NAT    | 1                  | 18.6x   | 22.44 | 25.67 | 28.61 | 28.90 |
|             | Hint-NAT       | 1                  | 30.2x   | 21.11 | 25.24 | —     | —     |
|             | ENGINE         | 10                 | —       | —     | —     | —     | 34.04 |
|             | FCL-NAT        | 1                  | 28.9x   | 21.70 | 25.32 | —     | —     |
|             | TCL-NAT        | 1                  | 27.6x   | 21.94 | 25.62 | —     | —     |
|             | EM+ODD         | 1                  | 16.4x   | 24.54 | 27.93 | —     | —     |
|             | MULTI-TASK NAT | 10                 | —       | 27.98 | 31.27 | 33.80 | 33.60 |

续表

| 类别     | 模型             | Iteration | Speedup | WMT14 |       | WMT16 |       |
|--------|----------------|-----------|---------|-------|-------|-------|-------|
|        |                |           |         | EN-DE | DE-EN | EN-RO | RO-EN |
| 其他改进方法 | ENAT           | 1         | 25.3x   | 20.65 | 23.02 | 30.08 | —     |
|        | LAVA NAT       | 1         | 20.2x   | 27.94 | 31.33 | —     | 32.85 |
|        | RecoverSAT     | n/2       | 2.09x   | 27.11 | 31.67 | 32.92 | 33.19 |
|        | GLAT           | 1         | 15.3x   | 25.21 | 29.84 | 31.19 | 32.04 |
|        | PMG            | 2.5       | 3.5x    | 27.8  | —     | —     | 33.8  |
|        | MvCR-NAT       | 10        | 3.77x   | 27.39 | 31.18 | 33.38 | 33.56 |
|        | DA-Transformer | 1         | 7.0x    | 27.25 | 31.54 | —     | —     |
|        | Fully NAT      | 1         | 16.8x   | 27.20 | 31.39 | 33.71 | 34.16 |

### 3 展望与分析

#### 3.1 NART 模型解码速度的公平评估

NART 模型的主要动机之一是加速推断过程,但 Helcl 等人<sup>[71]</sup>指出现有 NART 工作的速度评估方法存在两大缺陷:一方面,大多数方法的解码速度是在不同硬件架构上测试的,这让结果无法比较;另一方面,这些方法只测试了在 GPU 上每个 Batch 用单个句子进行解码的速度,Helcl 等人通过 GPU 和 CPU 上五个场景的实验结果证明这是唯一有利于 NART 模型的设置,而随着 GPU 上 Batch 增大, NART 模型的速度优势就会不断减小,而真实场景下 ART 模型能从更大的 Batch 中受益。这也导致了在一定程度上, NART 研究工作的实验结果不容易复现,对比不充分。所以,在统一的环境下公平地评估 NART 模型的速度,增加系统的可复现性,提高 NART 模型的实际应用速度,是促进 NART 模型真实场景应用的一个重要方向。

#### 3.2 在其他任务上应用非自回归技术

非自回归在机器翻译中的成功应用,也给很多其他任务提供了思路,例如,文本到语音<sup>[72-73]</sup>、自动语音生成<sup>[74-76]</sup>和图像描述<sup>[76-77]</sup>等任务已经把非自回归的方法结合到任务中,并获得了不同程度的性能提升。从另一个角度来看,相比于 ART 模型只能获得当前词之前的历史信息, NART 模型可以同时建模上下文信息,更有利于提升生成质量。Susanto 等人<sup>[79]</sup>提出了一种基于词约束的 NART 模型,通过在推断的时候加入术语约束,把机器翻译

转换成了填空优化的问题,实验表明在术语约束的机器翻译任务上,获得了比 ART 模型更好的性能。所以,将 NART 技术应用到契合非自回归特性的任务,或者进一步研究 NART 模型如何更好地捕获上下文信息也是一个值得探索的方向。特别是,对于需要引入上下文信息的任务, NART 模型也有很大潜力。

#### 3.3 结合预训练模型

预训练模型已经在很多工作中证实可以显著改善神经机器翻译的性能,所以也有一些研究人员探索了提高 NART 模型性能的预训练模型。Li 等人<sup>[79]</sup>提出了一个由双向编码器和双向解码器组成的多语言条件掩码模型,由于解码器具有双向结构,所以提出的模型也可以为 NART 模型提供初始化参数,带来了大幅的性能增益。Zhang 等人<sup>[81]</sup>提出了基于插入的非自回归预训练方法,使用渐进插入的目标对模型进行预训练,在受控文本生成方面实现了最先进的性能。然而,采用预训练模型来增强 NART 模型的工作极少,并且都是对非自回归训练技术的初步尝试,远不及自回归预训练技术的探索程度,所以进一步地研究 NART 和其他相关任务的非自回归预训练方法是一个值得思考的问题。

### 4 总结

本文首先系统性地把近年来非自回归机器翻译方法的研究进展进行了分类,总结分析了各类方法的优缺点,并对非自回归的未来研究方向进行了展望。最后,本文对相关论文进行了归纳整理,包括方

法分类、会议类型和发表年份。希望本文可以帮助研究人员开展与非自回归相关的工作,起到促进非自回归技术发展的作用。

## 参考文献

- [1] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[C]//Proceedings of the 34th International Conference on Machine Learning, 2017:1243-1252.
- [2] VASWANI A, SHAZEER N M, PARMAR N, et al. Attention is all you need[C]//Proceedings of the Advances in Neural Information Processing Systems, 2017:6000-6010.
- [3] GU J, BRADBURY J, XIONG C, et al. Non-autoregressive neural machine translation[C]//Proceedings of the 6th International Conference on Learning Representations, 2018.
- [4] LEE J, MANSIMOV E, CHO K. Deterministic non-autoregressive neural sequence modeling by iterative refinement[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 1173-1182.
- [5] SHU R, LEE J, NAKAYAMA H, et al. Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020: 8846-8853.
- [6] GHAZVININEJAD M, LEVY O, LIU Y, et al. Mask-predict: Parallel decoding of conditional masked language models[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2019: 6111-6120.
- [7] KASAI J, CROSS J, GHAZVININEJAD M, et al. Non-autoregressive machine translation with disentangled context transformer[C]//Proceedings of the 37th International Conference on Machine Learning, 2020: 5144-5155.
- [8] 肖桐,朱靖波.机器翻译:基础与模型[M].北京:电子工业出版社, 2021:373-374.
- [9] XIAO Y, WU L, GUO J, et al. A survey on non-autoregressive generation for neural machine translation and beyond[J]. arXiv preprint arXiv: 2204.09269, 2022.
- [10] KIM Y, RUSH A M. Sequence-level knowledge distillation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016:1317-1327.
- [11] ZHOU C, NEUBIG G, GU J. Understanding knowledge distillation in non-autoregressive machine translation[C]//Proceedings of the 8th International Conference on Learning Representations, 2020.
- [12] XU W, MA S, ZHANG D, et al. How does distilled data complexity impact the quality and confidence of non-autoregressive machine translation? [C]//Proceedings of the Findings of the Association for Computational Linguistics, 2021:4392-4400.
- [13] ZHOU J, KEUNG P. Improving non-autoregressive neural machine translation with monolingual data[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 1893-1898.
- [14] DING L, WANG L, LIU X, et al. Understanding and improving lexical choice in non-autoregressive translation[C]//Proceedings of the 9th International Conference on Learning Representations, 2021.
- [15] DING L, WANG L, LIU X, et al. Rejuvenating low-frequency words: Making the most of parallel data in non-autoregressive translation [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021: 3431-3441.
- [16] SHAO C, WU X, FENG Y. One reference is not enough: Diverse distillation with reference selection for non-autoregressive translation[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2022: 3779-3791.
- [17] HUANG F, TAO T, ZHOU H, et al. On the learning of non-autoregressive transformers[C]//Proceedings of the International Conference on Machine Learning, 2022:9356-9376.
- [18] LEE J, SHU R, CHO K. Iterative refinement in the continuous space for non-autoregressive neural machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020:1006-1015.
- [19] STERN M, CHAN W, KIROS J R, et al. Insertion transformer: Flexible sequence generation via insertion operations[C]//Proceedings of the 36th International Conference on Machine Learning, 2019: 5976-5985.
- [20] GU J, WANG C, ZHAO J, et al. Levenshtein transformer[C]//Proceedings of the Advances in Neural Information Processing Systems, 2019:11179-11189.
- [21] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4101-4110.

- tion for Computational Linguistics, 2019: 4171-4186.
- [22] GHAZVININEJAD M, LEVY O, ZETTLEMOYER L. Semi-autoregressive training improves mask-predict decoding[J]. arXiv preprint arXiv:2001.08785, 2020.
- [23] GUO J, XU L, CHEN E. Jointly masked sequence-to-sequence model for non-autoregressive neural machine translation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 376-385.
- [24] DING L, WANG L, WU D, et al. Context-aware cross-attention for non-autoregressive translation [C]//Proceedings of the 28th International Conference on Computational Linguistics, 2020:4396-4402.
- [25] NOROUZI S, HOSSEINZADEH R, PÉREZ F, et al. DiMS: Distilling multiple steps of iterative non-autoregressive transformers[J]. arXiv preprint arXiv: 2206.02999, 2022.
- [26] GENG X, FENG X, QIN B. Learning to rewrite for non-autoregressive neural machine translation [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021:3297-3308.
- [27] KASAI J, PAPPAS N, PENG H, et al. Deep encoder, shallow decoder: Reevaluating the speed-quality tradeoff in machine translation[J].arXiv preprint arXiv:2006.10369, 2020.
- [28] KAISER Ł, ROY A, VASWANI A, et al. Fast decoding in sequence models using discrete latent variables[C]//Proceedings of the 35th International Conference on Machine Learning, 2018: 2395-2404.
- [29] AKOURY N, KRISHNA K, IYYER M. Syntactically supervised transformers for faster neural machine translation[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019: 1269-1281.
- [30] BAO Y, HUANG S, XIAO T, et al. Non-autoregressive translation by learning target categorical codes [C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2021:5749-5759.
- [31] ROY A, VASWANI A, NEELAKANTAN A, et al. Theory and experiments on vector quantized autoencoders[J]. arXiv preprint arXiv:1805.11063, 2018.
- [32] BAO Y, ZHOU H, FENG J, et al. Non-autoregressive transformer by position learning[J]. arXiv preprint arXiv:1911.10677, 2019.
- [33] RAN Q, LIN Y, LI P, et al. Guiding non-autoregressive neural machine translation decoding with re-ordering information [C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021: 13727-13735.
- [34] SONG J, KIM S, YOON S. AlignNART: Non-autoregressive neural machine translation by jointly learning to estimate alignment and translate [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021:1-14.
- [35] MA X, ZHOU C, LI X, et al. FlowSeq: Non-autoregressive conditional sequence generation with generative flow[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2019: 4281-4291.
- [36] REZENDE D J, MOHAMED S. Variational inference with normalizing flows[C]//Proceedings of the 32nd International Conference on Machine Learning, 2015: 1530-1538.
- [37] LIBOVICKÝ J, HELCL J. End-to-end non-autoregressive neural machine translation with connectionist temporal classification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 3016-3021.
- [38] SAHARIA C, CHAN W, SAXENA S, et al. Non-autoregressive machine translation with latent alignments[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 1098-1108.
- [39] WANG Y, TIAN F, HE D, et al. Non-autoregressive machine translation with auxiliary regularization [C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019: 5377-5384.
- [40] SUN Z, LI Z, WANG H, et al. Fast structured decoding for sequence models[C]//Proceedings of the Advances in Neural Information Processing Systems, 2019: 3011-3020.
- [41] LIU Y, WAN Y, ZHANG J, et al. Enriching non-autoregressive transformer with syntactic and semantic structures for neural machine translation [C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021:1235-1244.
- [42] SHAO C, FENG Y, ZHANG J, et al. Retrieving sequential information for non-autoregressive neural machine translation [C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019: 3013-3024.
- [43] SHAO C, ZHANG J, FENG Y, et al. Minimizing the bag-of-ngrams difference for non-autoregressive neural machine translation [C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence,

- 2020; 198-205.
- [44] SHAO C, FENG Y, ZHANG J, et al. Sequence-level training for non-autoregressive neural machine translation[C]//Proceedings of the Computational Linguistics, 2021:891-925.
- [45] GHAZVININEJAD M, KARPUKHIN V, ZET-LEMOYER L, et al. Aligned cross entropy for non-autoregressive machine translation[C]//Proceedings of the 37th International Conference on Machine Learning, 2020; 3515-3523.
- [46] DU C, TU Z, JIANG J. Order-agnostic cross entropy for non-autoregressive machine translation[C]//Proceedings of the 38th International Conference on Machine Learning, 2021; 2849-2859.
- [47] WEI B, WANG M, ZHOU H, et al. Imitation learning for non-autoregressive neural machine translation[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019; 1304-1312.
- [48] LI Z, LIN Z, HE D, et al. Hint-based training for non-autoregressive machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2019; 5707-5712.
- [49] TU L, PANG R, WISEMAN S, et al. ENGINE: Energy-Based inference networks for non-autoregressive machine translation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020; 2819-2826.
- [50] GUO J, TAN X, XU L, et al. Fine-tuning by curriculum learning for non-autoregressive neural machine translation[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2020; 7839-7846.
- [51] LIU J, REN Y, TAN X, et al. Task-level curriculum learning for non-autoregressive neural machine translation[C]//Proceedings of the 29th International Joint Conference on Artificial Intelligence, 2020; 3861-3867.
- [52] SUN Z, YANG Y. An EM Approach to Non-autoregressive Conditional Sequence Generation[C]//Proceedings of the 37th International Conference on Machine Learning, 2020; 9249-9258.
- [53] HAO Y, HE S, JIAO W, et al. Multi-task learning with shared encoder for non-autoregressive machine translation[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2021; 3989-3996.
- [54] GUO J, TAN X, HE D, et al. Non-autoregressive neural machine translation with enhanced decoder input[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019; 3723-3730.
- [55] ZHAN J, CHEN Q, CHEN B, et al. Non-autoregressive translation with dependency-aware decoder[J]. arXiv preprint arXiv:2203.16266, 2022.
- [56] LI X, MENG Y, YUAN A, et al. LAVA NAT: A non-autoregressive translation model with look-around decoding and vocabulary attention[J]. arXiv preprint arXiv:2002.03084, 2020.
- [57] DENG Y, RUSH A M. Sequence-to-lattice models for fast translation[C]//Proceedings of the Association for Computational Linguistics, 2021; 3765-3772.
- [58] ZHENG Z, ZHOU H, HUANG S, et al. Duplex sequence-to-sequence learning for reversible machine translation[C]//Proceedings of the Advances in Neural Information Processing Systems, 2021; 21070-21084.
- [59] WANG C, ZHANG J, CHEN H. Semi-autoregressive neural machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018; 479-488.
- [60] RAN Q, LIN Y, LI P, et al. Learning to recover from multi-modality errors for non-autoregressive neural machine translation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020; 3059-3069.
- [61] KONG X, ZHANG Z, HOVY E H. Incorporating a local translation mechanism into non-autoregressive translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020; 1067-1073.
- [62] DENG Y, RUSH A M. Cascaded text generation with markov transformers[C]//Proceedings of the Advances in Neural Information Processing Systems, 2020.
- [63] QIAN L, ZHOU H, BAO Y, et al. Glancing transformer for non-autoregressive neural machine translation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021; 1993-2003.
- [64] DING L, WANG L, LIU X, et al. Progressive multi-granularity training for non-autoregressive translation[C]//Proceedings of the Association for Computational Linguistics, 2021; 2797-2803.
- [65] XIE P, LI Z, HU X. MvSR-NAT: Multi-view subset regularization for non-autoregressive machine translation[J]. arXiv preprint arXiv:2108.08447, 2021.
- [66] ZENG C, CHEN J, ZHUANG T, et al. Neighbors are not strangers: Improving non-autoregressive translation under low-frequency lexical constraints

- [J]. arXiv preprint arXiv:2204.13355, 2022.
- [67] BAO Y, ZHOU H, HUANG S, et al. GLAT: Glancing at latent variables for parallel text generation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 8398-8409.
- [68] HUANG F, ZHOU H, LIU Y, et al. Directed acyclic transformer for non-autoregressive machine translation[C]//Proceedings of the International Conference on Machine Learning, 2022:9410-9428.
- [69] HUANG X S, PEREZ F, VOLKOV M. Improving non-autoregressive translation models without distillation[C]//Proceedings of the International Conference on Learning Representations, 2022.
- [70] GU J, KONG X. Fully non-autoregressive neural machine translation: Tricks of the trade[C]//Proceedings of the Association for Computational Linguistics, 2021: 120-133.
- [71] HELCL J, HADDOW B, BIRCH A. Non-autoregressive machine translation: It's not as fast as it seems [C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2022:1780-1790.
- [72] REN Y, RUAN Y, TAN X, et al. FastSpeech: Fast, robust and controllable text to speech[C]//Proceedings of the Advances in Neural Information Processing Systems, 2019:3165-3174.
- [73] PENG K, PING W, SONG Z, et al. Non-autoregressive neural text-to-speech [C]//Proceedings of the 37th International Conference on Machine Learning, 2020: 7586-7598.
- [74] CHEN N, WATANABE S, VILLALBA J, et al. Listen and fill in the missing letters: non-autoregressive transformer for speech recognition[J]. arXiv preprint arXiv:1911.04908, 2019.
- [75] FUJITA Y, WATANABE S, OMACHI M, et al. Insertion-based modeling for end-to-end automatic speech recognition[C]//Proceedings of the 21st Annual Conference of the International Speech Communication Association, 2020: 3660-3664.
- [76] HIGUCHI Y, WATANABE S, CHEN N, et al. Mask CTC: Non-autoregressive end-to-end ASR with CTC and mask predict[C]//Proceedings of the 21st Annual Conference of the International Speech Communication Association, 2020: 3655-3659.
- [77] FEI Z. Fast image caption generation with position alignment[J]. arXiv preprint arXiv:1912.06365, 2019.
- [78] GAO J, MENG X, WANG S, et al. Masked non-autoregressive image captioning[J]. arXiv preprint arXiv:1906.00717, 2019.
- [79] SUSANTO R H, CHOLLAMPATT S, TAN L. Lexically constrained neural machine translation with levenshtein transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020:3536-3543.
- [80] LI P, LI L, ZHANG M, et al. Universal conditional masked language pre-training for neural machine translation[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2022.
- [81] ZHANG Y, WANG G, LI C, et al. Pointer: Constrained text generation via insertion-based generative pre-training[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020:8649-8670.



曹航(1999—),硕士研究生,主要研究领域为自然语言处理。

E-mail: caohang\_neu@outlook.com



肖桐(1982—),通信作者,博士,教授,主要研究领域为自然语言处理。

E-mail: xiaotong@mail.neu.edu.cn



胡驰(1997—),博士研究生,主要研究领域为神经网络结构搜索、机器翻译等。

E-mail: huchinlp@gmail.com