# 一种面向非自回归神经机器翻译的多模错误恢复学习机制

本文基于 ACL 2020 论文《Learning to Recover from Multi-Modality Errors for Non-Autoregressive Neural Machine Translation》撰写，论文作者为腾讯微信 AI 团队。

## 导语

非自回归神经机器翻译（non-autoregressive neural machine translation，简称非自回归翻译）是近年来兴起的一种新的机器翻译方法。为了提升翻译速度，其抛弃了目标语言词间的依赖关系，并行生成所有目标词，显著提升了翻译速度。但由于同一段原文经常有多种可行的潜在译文，使得非自回归翻译模型在解码时会出现不同位置的目标语言词是从不同的潜在译文中选取的问题，被称为多峰问题（multi-modality problem）。该问题通常表现为重复翻译和漏译现象，对非自回归翻译模型的翻译质量具有不利影响。为了缓解该问题，本文提出了一种新的模型 RecoverSAT。该模型该译文拆成多个片段并逐段生成，在每个片段内部采用自回归生成方式，而段间则采用非自回归方式。进一步的，通过引入动态停止机制和段删除机制，该方法可以有效缓解重复翻译和漏译的问题。实验结果显示，该方法在获得与传统神经机器翻译模型可比效果的同时，可以获得约 4 倍的加速。论文的预印版和代码均已开放。

## 背景

自回归神经机器翻译（autoregressive neural machine translation，AT，简称自回归翻译）因其领先的翻译质量而获得广泛关注与应用。但此类模型在生成每个目标语言词时（即解码）都依赖于在其之前已经生成的词，因此译文生成过程是串行的，使解码速度成为这类模型的一个重要瓶颈。非自回归神经机器翻译（non-autoregressive neural machine translation，简称非自回归翻译）抛弃了目标语言词间的依赖关系，将逐词生成变为所有词并行生成，显著提升了翻译速度，但也相伴产生了多峰问题（multi-modlity problem），通常表现为重复翻译和漏译现象。图 1 展示了多峰问题的一个示例：

| | |
|---|---|
| 源语言句子 | 今天有很多农民在做这件事 |
| 可行译文 1 | There are lots of farmers doing this today |
| 可行译文 2 | There are a lot of farmers doing this today |
| 生成译文 1 | There are lots of of farmers doing this today |
| 生成译文 2 | There are a lot farmers doing this today |

图 1：多峰问题示例

对于源语言片段“很多农民”，有多种可行的翻译方式，如“lots of farmers”和“a lot of farmers”，由于非自回归翻译模型抛弃了目标语言词间的依赖关系，导致其生成的译文中，不同位置的词可能源于不同的可行译文。如对于生成译文 1，“lots of”源于可行译文 1 而“of farmers”源于可行译文 2。同样的现象在生成译文 2 中也可以观察到。显然多峰问题对于翻译质量具有不可忽视的负面影响。

在已有的非自回归翻译工作中，一类工作采用多轮解码（iterative decoding）的方法缓解该问题，即每次解码时都将源语言句子和上一轮生成的译文同时作为新一轮解码的输入，具有能够修复生成错误的优点。但该类方法通常需要迭代多次以获得较好的翻译质量，对翻译速度有显著影响。另一类方法则通过在模型中额外引入以自回归方式工作的部件的方式注入目标语言端的依赖关系。但这类方法无法对已经发生的生成错误进行修正。

为了缓解多峰问题，本文提出了一种新的半自回归（semi-autoregressive）翻译模型 RecoverSAT。该模型具有三个方面的优点：(1) 为了提升翻译速度，该方法将译文拆分成多个片段，并使片段可以并行生成；（2）为了更好的捕捉目标语言端依赖关系，在生成每个词时，其不仅依赖于所在片段内已经被生成的词，还依赖于其他片段内已经被生成的词。一方面，我们观察到重复翻译多发现在短片段内，因此在每个片段内部采用自回归方式生成每个词有助于减少重复翻译现象。另一方面，将其他片段已被生成的词纳入考虑，有助于模型判断其他片段是基于何种可行译文选词并相应对本段内将要生成的词进行调整，从而有助于缓解漏译现象。综合以上两方面，我们的模型考虑了更多目标语言端信息，从而能够更好得缓解多峰问题。（3）为了使模型具备从已发生的重复翻译错误中进行恢复的能力，我们提出了一种段删除机制，使得模型在发现某段翻译的内容已在其他段中被翻译时，可以动态将该段删除，从而可以从错误中恢复。

## 模型结构

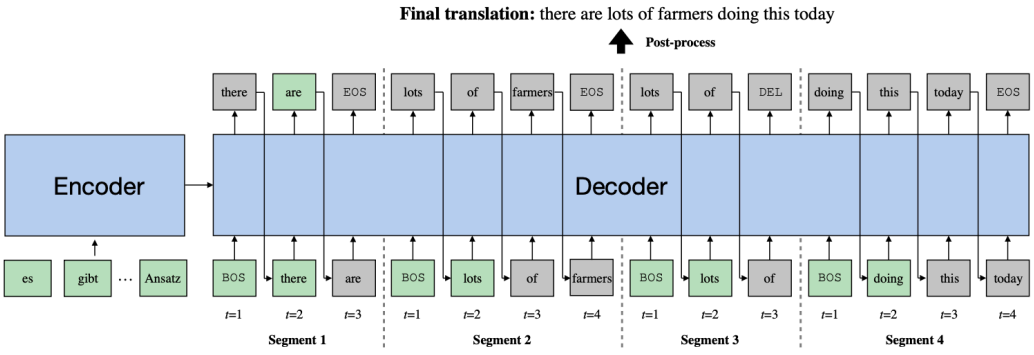


图 2：模型结构图

图 2 展示了 RecoverSAT 的模型结构。对于编码器（encoder），RecoverSAT 沿用了常用的 Transformer 结构，这里不再赘述。RecoverSAT 的主要创新点集中于解码器（decoder）。如图所示，在生成译文时，RecoverSAT 将译文拆分成多个片段，每个片段内自左向右逐词生成，而片段间

则并行生成。其中“EOS”表示对应的片段已生成完毕，而“DEL”表示相应的片段需要被整体删除。整体过程的概率公式表示如下：

$$P(y|x) = \prod_{t=1}^{L}\prod_{i=1}^{K} P(S_t^i|S_{<t}^1, \cdots, S_{<t}^K; x)$$

其中$x$表示源语言句子，$y$表示译文，$S_t^i$表示第$i$个片段的第$t$个词，$S_{<t}^i = \{S_1^i, \cdots, S_{t-1}^i\}$表示第$i$个片段的历史信息，$L$表示片段长度，$K$表示片段数目。

为了更好的缓解多峰问题，我们提出了“动态停止”和“段删除”两个机制，其工作原理如下：

动态停止机制：如上所述，与多数已有的非自回归翻译模型不同，我们并不预先指定译文的长度，而是通过让模型动态为每段生成特殊符号“EOS”的方式来让模型自行决定译文的长度。我们称之为动态停止机制。该机制可以从两个方面缓解多峰问题：

1. 对于每个片段，其开头几个词的选择可以更加灵活。以图 2 为例，对于第 2 个片段（Segment 2），如果模型将其第一个词选为“of”而不是“lots”，那么模型只需在第 1 个片段（Segment 1）的结尾处多生成出词“lots”即可避免漏译错误。同理，如果第 2 个片段的第 1 个词被选择为“are”，那么第 1 个片段只需不生成“are”即可避免重复翻译错误。
2. 如$P(y|x)$的定义所示，生成每个词时，所有片段中已经生成的词均被作为历史使用，从而使模型可以考虑更多历史信息，从而更易发现多峰问题相关错误并从中恢复。

这里需要指出的是，如何训练模型使其具备上述能力但又保持较好的解码速度是一个很大的挑战。一方面，RecoverSAT 的翻译速度与各个片段的最长长度成正比，相应的，我们需要在构造训练数据时尽量将参考译文拆分成长度相等的片段，以鼓励模型生成长度大致相等的片段；另一方面，我们又需要在训练过程中给模型提供充足的具有多峰问题相关错误的训练数据，以使模型可以学习到从相关错误中恢复的能力。为了平衡这两方面的要求，在训练过程中，我们随机决定每条译文是随机拆分成长度不等的片段还是长度相当的片段，并且随着训练的进行，逐渐增加拆分成长度相当的片段的译文比例。

段删除机制：显然动态停止机制并不能消除所有多峰问题相关错误，因此我们引入段删除机制来在已经发生重复翻译错误时从错误中恢复。具体的，当模型发现某一个片段的译文对应的内容已在其他片段中被删除时，模型将生成一个特殊符号“DEL”表示该片段需要被删除。在解码过程完成后，我们将在后处理阶段将对应的片段从译文中删除，以获得最终译文。需要特别指出的是，我们观察到重复翻译现象通常在每段的前几个词处发生，因为此时模型可供参考的历史信息非常有限。因此我们采用预测整段是否删除而非每个词是否该被删除的建模方式以提升模型的简洁性。在训练过程中，我们通过人为插入一些重复片段的方式为模型提供相应的训练数据。

## 实验效果

为验证模型的有效性，我们在 WMT14 En-De、WMT16 En-Ro、IWSLT16 En-De 三个数据集、五个翻译方向上进行了实验，相应的实验结果见图 3。从实验结果中我们可以看到：（1）RecoverSAT 获得了与自回归翻译模型（Transformer）可比的翻译质量（BLEU 值）。特别的，在分段数为 2 时，翻译质量差距很小但解码速度快 2 倍以上。而在分段数为 10 时，BLEU 值下降不足 5%，而解码速度为自回归翻译模型的 4 倍。（2）RecoverSAT 在翻译质量上超过了除 CMLM 以外的所有非自回归翻译模型。而 RecoverSAT 与 CMLM 的翻译质量差别很小但 RecoverSAT 解码速度显著快于 CMLM。（3）随着分段数的增加，RecoverSAT 翻译质量有较小的下降而解码速度有显著提升，显示其具有较好的泛化性。

| Model | Iterative Decoding | WMT14 En-De | | | WMT16 En-Ro | | | IWSLT16 En-De | |
|---|---|---|---|---|---|---|---|---|---|
| | | En→ | De→ | Speedup | En→ | Ro→ | Speedup | En→ | Speedup |
| Transformer | | 27.17 | 31.95 | 1.00× | 32.86 | 32.60 | 1.00× | 31.18 | 1.00× |
| NAT-FT+NPD ($n = 100$) | | 19.17 | 23.20 | - | 29.79 | 31.44 | - | 28.16 | 2.36× |
| SynST | | 20.74 | 25.50 | 4.86× | - | - | - | 23.82 | 3.78× |
| NAT-IR ($iter = 10$) | ✓ | 21.61 | 25.48 | 2.01× | 29.32 | 30.19 | 2.15× | 27.11 | 1.55× |
| NAT-FS | | 22.27 | 27.25 | 3.75× | 30.57 | 30.83 | 3.70× | 27.78 | 3.38× |
| imitate-NAT+LPD ($n = 7$) | | 24.15 | 27.28 | - | 31.45 | 31.81 | - | 30.68 | 9.70× |
| PNAT+LPD ($n = 9$) | | 24.48 | 29.16 | - | - | - | - | - | - |
| NAT-REG+LPD ($n = 9$) | | 24.61 | 28.90 | - | - | - | - | 27.02 | - |
| LV NAR | | 25.10 | - | 6.8× | - | - | - | - | - |
| NART+LPD ($n = 9$) | | 25.20 | 29.52 | 17.8× | - | - | - | - | - |
| FlowSeq+NPD ($n = 30$) | | 25.31 | 30.68 | <1.5× | 32.20 | 32.84 | - | - | - |
| FCL-NAT+NPD ($n = 9$) | | 25.75 | 29.50 | 16.0× | - | - | - | - | - |
| ReorderNAT | | 26.51 | 31.13 | - | 31.70 | 31.99 | - | 30.26 | 5.96× |
| NART-DCRF+LPD ($n = 19$) | | 26.80 | 30.04 | 4.39× | - | - | - | - | - |
| SAT ($K = 2$) | | 26.90 | - | 1.51× | - | - | - | - | - |
| CMLM ($iter = 10$) | ✓ | 27.03 | 30.53 | <1.5× | **33.08** | **33.31** | - | - | - |
| RecoverSAT ($K = 2$) | | **27.11** | **31.67** | 2.16× | 32.92 | 33.19 | 2.02× | **30.78** | 2.06× |
| RecoverSAT ($K = 5$) | | 26.91 | 31.22 | 3.17× | 32.81 | 32.80 | 3.16× | 30.55 | 3.28× |
| RecoverSAT ($K = 10$) | | 26.32 | 30.46 | 4.31× | 32.59 | 32.29 | 4.31× | 29.90 | 4.68× |

图 3：实验效果（BLEU 值）。其中 RecoverSAT 括号中的 K 表示分段数

图 4 展示了一个实际例子上的翻译结果。在这个例子上，我们通过要求模型在特定位置生成特定词的方式（绿色标记），探测模型从多峰相关问题恢复的能力。可以看到：（1）已有的非自回归翻译模型（NAT）生成的译文中有多处重复翻译和漏译，而 RecoverSAT 生成的译文中相应现象很少。（2）RecoverSAT 具备动态决定片段长度以缓解重复翻译错误的能力（译文 B），也具备从漏译错误中恢复的能力（译文 C 和 D）。（3）RecoverSAT 具备发现和删除重复片段的能力（译文 D）。

| | | |
|---|---|---|
| Source | | die er_greif_endste Abteilung ist das Denk_mal für die Kinder , das zum Ged_enken an die 1,5 Millionen Kinder , die in den Konzent_rations_lagern und Gas_k_ammern vernichtet wurden , erbaut wurde . |
| Reference | | the most tragic section is the children's mem_orial , built in memory of 1.5 million children killed in concentration camps and gas cham_bers . |
| NAT | Translation | the most tangible department department the monument monument the children , which was built commem_commem_orate 1.5 1.5 million children were destroyed in the concentration camps and gas cham_bers . |
| RecoverSAT ($K = 10$) | Translation | **A**: [1]the EOS [2]most tangible department is the EOS [3]monument for children EOS [4]built to EOS [5]commem_orate the 1.5 EOS [6]million children destroyed EOS [7]in the concentration camps and EOS [8]in DEL [9]gas EOS [10]cham_bers . EOS |
| | Forced Translation | **B**: [1]the EOS [2]most tangible department is the EOS [3]monument for children EOS [4]built to EOS [5]commem_orate EOS [6]the 1.5 million children destroyed EOS [7]in the concentration camps and EOS [8]in DEL [9]gas EOS [10]cham_bers . EOS |
| | | **C**: [1]the EOS [2]most tangible department is the EOS [3]monument for children EOS [4]built to EOS [5]commem_orate the 1.5 million children EOS [6]destroyed EOS [7]in concentration camps and EOS [8]in DEL [9]gas EOS [10]cham_bers . EOS |
| | | **D**: [1]the EOS [2]most tangible department is the EOS [3]monument for children EOS [4]built to EOS [5]commem_orate the 1.5 million children destroyed EOS [6]in the concentration camps and EOS [7]in the DEL [8]in DEL [9]gas EOS [10]cham_bers . EOS |

图 4：实际翻译质量示例。“Forced Translation”表示我们强制要求模型生成某个特定的词（绿色标记）以观察相应现象。黄色：重复翻译；红色：漏译；灰色：模型预测需要被删除的片段

## 结语

为了缓解非自回归翻译中的多峰问题，本文提出一种新的半自回归翻译模型 RecoverSAT。该模型通过将译文拆分成多个片段、片段内用自回归方式生成而片段间用非自归方式生成的方法在获得解码速度提升的同时保持了较好的翻译质量。进一步的，该方法通过引入动态停止机制和段删除机制，可以进一步缓解多峰问题并具备一定的从多峰问题相关错误中恢复的能力。将本文中的方法与其他用于解决多峰问题的方法进行有机结合以获得更好的翻译质量和解码速度将是非常值得关注的研究方向。