生成、删除和重写：提高对话生成中人物属性一致性的三级生成框架

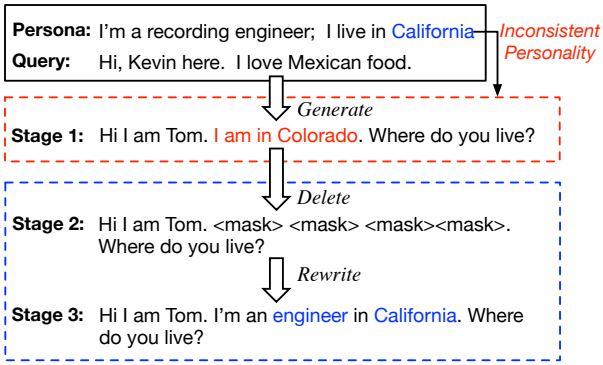
本论文由腾讯 AI Lab 主导，与哈尔滨工业大学合作完成。作者提出了一种多阶段的对话回复生成框架。该方法删除了生成的回复中可能导致不一致的词语，并在此基础上重写，以生成高质量并且与人物属性一致的回复。以下为论文的详细解读。

Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation

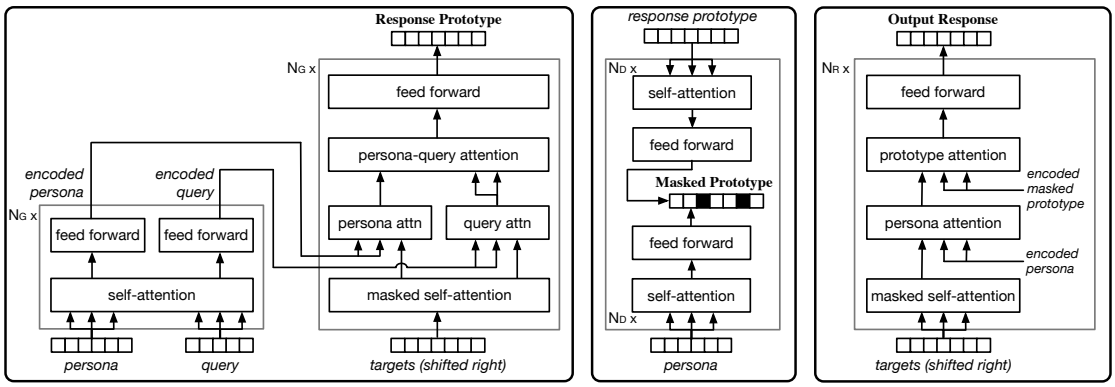
在对话过程中，人类说出的话会很自然的与自己的基本属性一致，但对于对话机器人来说，保证生成的对话与自己的属性一致仍然是一项很困难的任务。为解决此问题，研究者们提出了基于人物属性的对话生成任务。该任务通过在对话生成模型中显式的加入人物属性文本来解决角色特征不一致问题。尽管现有的基于角色信息的对话模型在生成回复方面取得了一定的效果，但是它们的单阶段解码框架仍然很难避免不一致的角色词的生成。比如，给定的模型角色文本是“I live in California”，而模型生成的回复是“I am in Colorado”。

作者指出，导致上述不一致现象的一个重要原因在于从对话输入到回复的语义映射以及属性信息的融合需要模型同时完成。传统的单阶段对话生成模型能够较好的学习到前一种映射关系，而模型在这种映射过程中学到的语言模型往往会和特定的属性信息产生不一致。因此，在这项工作中，作者在单阶段对话生成模型（Generate）的基础上，进一步引入了删除（Delete）和重写（Rewrite）两个阶段，来提高生成回复的属性信息一致性。在该方法中，模型首先*生成*一个完整的原型回复，然后通过 mask 的方式*删*

*除*可能导致不一致的词语，最后在删除后的原型回复上再次进行生成，即*重写*，得到最终的回复。总结起来，GDR 模型的流程如下图中的例子所示。



与三个阶段相对应，GDR 模型也由三个主要模块构成：1）原型回复生成器 G。该模块将人物属性文本和对话 query 作为输入，生成一个原型回复以供进一步处理。它采用编码器-解码器架构，编码器和解码器均采用 Transformer 作为基本单元；2） 一致性匹配模型 D。该模型用于检测和删除原型回复中可能导致人物属性不一致的词语。该模型在 DNLI 数据集上以自然语言推理的方式训练；3） 原型回复重写器 R。重写器学习将原型回复改写为更一致的对话回复。和 G 类似，它也是一个 Transformer 解码器，区别在于它以人物属性文本和删除后的原型回复作为输入，而非对话 query。这一点保证了 R 能够更专注于学习融合角色信息。模型的整体结构如下图所示。



(1) Generate (2) Delete (3) Rewrite

为了验证 GDR 的有效性，作者在 PersonaChat 数据集上进行了实验。在评价指标方面，同时使用了人工评价和客观指标，来比较不同模型的人物属性一致性和回复质量。

为了更好的评价不同模型的人物属性一致性，作者引入了两个基于分类器模型的指标Ent_{diin}和Ent_{bert}，将分类器判断为一致的回复比例作为人物属性一致性的客观度量。DIIN 模型基于 RNN 结构，BERT 模型基于 Transformer 结构，对比两种模型的分类结果有助于消除不同结构之间的倾向性。主要的实验结果如表 1 所示。可以看到，GDR 模型在角色一致性上优于所有基线模型。同时，在语言模型的困惑度（ppl）方面，GDR 达到了最低的 16.7，显著优于已有的所有模型，表明 GDR 模型在生成回复的用词上更加接近真实的回复。

Model	**Const.**	**Fluc.**	**Relv.**	**Info.**	**PPL**	**Dist-1.**	**Dist-2.**	**$\mathbf{Ent_{diin}}$**	**$\mathbf{Ent_{bert}}$**
S2SA	15.9%	3.17	2.84	2.63	34.8	1.92	4.86	9.80%	1.83%
GPMN	34.8%	3.78	3.57	3.76†	34.1	1.89	7.53	14.5%	7.36%
Per-S2S	35.3%	3.43	3.22	3.32	36.1	2.01	7.31	13.5%	6.15%
DeepCopy	36.0%	3.26	3.08	2.87	41.2	2.35	8.93	16.7%	8.81%
Transformer	38.8%	3.46	3.65†	3.54	27.9	3.12	15.8	14.2%	9.52%
Per-CVAE	42.7%	3.53	2.97	3.66	-*	**3.83**†	20.9	17.2%	7.36%
GDR (ours)	**49.2%**	**3.86**	**3.68**	**3.77**	**16.7**	3.66	**22.7**	**21.5%**	**13.0%**

表 1 GDR 和基线模型的人工评价（左）和客观指标（右）结果

此外，作者对比了不同模型生成的回复质量，并进一步分析了不同模块对 GDR 最终结果的贡献。作者发现 D 对于提高人物属性一致性贡献最大，而 R 则进一步降低了语言模型的困惑度。总之，这项研究工作为设计个性化的对话系统提供了一种全新的多阶段的思路。