arXiv:2301.02275v2 [cs.CL] 8 Sep 2023

语言作为潜在序列：用于半监督释义生成的深度潜变量模型 *

Jialin Yu、Alexandra I. Cristea、Anoushka Harit、Zhongtian Sun、Olanrewaju Tahir Aduragba、Lei Shi、Noura Al Moubayed计算机科学系 杜伦大学 英国

摘要

本文探讨用于半监督释义生成的深度潜变量模型，其中，未标注数据缺失的目标文本对被建模为潜在释义序列。我们提出了一种名为变分序列自动编码重建 **(VSAR)**的新型无监督模型，该模型根据观测到的文本执行潜在序列推断。为了利用文本对中的信息，我们还提出了一种名为双向学习**(DDL)**的新型监督模型，旨在与我们提出的 VSAR 模型相结合。将 VSAR 与 DDL 结合（**DDL+VSAR**）使我们能够开展半监督学习。然而，组合模型存在冷启动问题。为进一步解决这一问题，我们提出了一种改进的权重初始化方案，并据此设计了一种名为知识强化学习 **(KRL)** 的新型两阶段训练方案。实证评估结果表明，在完整数据上，组合模型的性能可与最先进的监督基线方法相媲美。此外，在仅有部分带标签文本对可用的情况下，我们的组合模型始终显著优于强监督模型基线（**DDL)**（$p < .05$；Wilcoxon检验）。我们的代码已公开发布于 https://github.com/jialin-yu/latent-sequence-paraphrase。

关键词 释义生成 · 半监督学习 · 深度潜变量模型

1 引言

释义生成是自然语言处理（NLP）中的一个重要问题，在许多 NLP 应用中都很有用，例如问答 [1], 信息检索 [2], 信息抽取 [3] 和自动摘要 [4]。自然语言本身十分复杂，同一底层语义内容 [5, 6]可能以多种不同的表层形式表达。因此，将释义生成模型作为组件集成到现实世界的 NLP 系统中，为最终用户的输入提供稳健的响应，至关重要。传统的释义生成解决方案通常基于规则 [7, 8], ，利用 WordNet [9],等词汇资源寻找替换词。近期的趋势则是采用神经网络模型 [10, 11, 12, 13], ，这类模型通常基于序列到序列学习范式 [14]。

由于采用了复杂的架构和精巧的条件机制，这些模型在释义生成方面取得了显著成功，例如软注意力、硬注意力和自注意力。然而，这类模型的发展主要依赖于大规模带标注的数据对。与此不同，本文探讨了半监督学习场景，其中只有一部分标签可用。由于获得高质量人工标注需要投入时间和金钱，这种半监督学习设置适用于工业场景 [15, 16], ，且极具实用价值。半监督学习模型通常由两个部分组成：无监督学习模型和监督学习模型。

因此，对于无监督学习部分，我们提出了一种新颖的深度生成模型，其灵感来自经典的变分自编码器（VAE）[17, 18, 19], ，并针对语言序列建模引入了额外的结构假设，

*已被 *AI Open* 期刊接收：*https:// www. sciencedirect. com/*《科学》*/ article/ pii/ S2666651023000025*

该模型名为变分序列自动编码重建 （**VSAR**）。进一步来说，传统的变分自编码器通常将数据表示嵌入固定的潜在空间，其总体目标是进行降维 [20]。在这里，我们转而将潜变量视为长度各异的离散语言序列。这一假设促使模型训练过程采用更多结构信息，并且还能增强模型可解释性，因为语言天然以离散变量的形式得以保留 [21]。继承近期成功引入离散潜变量以改进释义生成的先验工作 [22, 6], ，我们提出了 VSAR 模型，旨在包含一种表达能力更强的离散潜变量形式，如图 1 所示。

此外，对于监督学习部分，受双向学习 [23, 24, 25, 26], 的启发，我们提出了一种名为双向学习（**DDL**）的新型监督模型，该模型可以与我们提出的 VAE 模型结合，并共享 VSAR 模型学习循环的一部分。结合无监督模型和监督模型，可以利用 VAE 在未标注数据中边缘化潜变量的能力，实现半监督学习。

因此，本文的主要原创性贡献包括：

- 首次提出使用深度离散潜变量模型进行释义生成的半监督学习研究；
- 介绍两种新型模型：*VSAR*（无监督）和 *DDL*（监督），二者可以结合用于半监督学习；
- 提出一种新型训练方案——知识强化学习（*KRL*），以解决结合后的半监督模型（DDL+VSAR）中的冷启动问题；
- 使用组合模型在完整数据集上研究半监督学习场景，并通过实验证明我们的模型取得了具有竞争力的最先进结果；
- 研究仅使用部分已标注数据的半监督场景（即纳入未标注数据时），并证明我们的模型显著优于非常强大的监督基线。

2 相关工作

2.1 释义生成

释义表达了潜在语义内容的表层形式 [6] ，并体现了语言多样性的精髓 [27]。早期自动生成释义的工作通常基于规则 [7, 8], ，但近年来的趋势转向神经网络解决方案 [28, 21, 10, 12, 13, 22, 6, 29, 30]。当前释义生成研究主要集中于监督方法，这类方法需要大量源和目标对。在本研究中，我们转而探索一种半监督释义生成方法：只观察部分源和目标对，同时存在大量未标注源文本。我们假设，每个缺失的目标文本都可以视为深度生成模型中的一个潜变量。因此，对于无监督数据，每个缺失的释义输出都被建模为一个潜变量。与将句子语义表示为稠密高维向量的标准方法相比，这一假设促使模型学习更具结构化的表示。因此，我们提出的模型可视为一种深度潜在结构模型[31]。本文提出了两个用于释义生成的模型，并将它们结合起来：一个用于无监督学习，另一个用于监督学习。我们的组合模型拓展了 [32, 21]中的思路；与 [21], 相比，我们的模型反而采用了更自然的语言结构（有序序列，而非无序词袋）；与 [32], 相比，我们的模型采用自注意力机制而非卷积操作，并且能够对不同长度的潜在序列进行建模，而不是局限于固定长度。

此外，我们的组合模型与引入离散潜变量的先前研究相关 [22, 6], ，并且它采用了一种可以说表达能力更强的潜变量，即语言输出。

2.2 文本深度潜变量模型

深度潜变量模型已用于文本建模研究 [33, 34]。最常见且应用最广泛的潜变量模型是采用高斯先验 [35], 的标准变分自编码器模型，但它会遭遇后验坍塌 [36, 37]。已有多项研究致力于解决这一问题 [38, 39, 40]。具体而言，β-VAE [38]引入了惩罚项，以直观地平衡变分自编码器重构与先验正则化，并被用作我们的基线方法之一。

尽管许多研究聚焦于连续潜变量模型 [33, 34], ，但文本天然以离散形式呈现，可能无法用连续潜变量很好地表示。早期关于离散深度潜变量

模型 [32, 41] 采用了REINFORCE算法 [19, 42]；然而，该算法的方差非常高。随着统计松弛技术近期取得进展，研究者使用了Gumbel技巧 [43, 44] ，在文本潜变量模型中对离散结构建模 [45, 21, 46, 47]。我们的工作将结合子集采样的Gumbel技巧 [48] 用于自然语言生成任务，并且首次将离散语言序列作为释义任务的潜变量进行研究。我们提出的模型与 [32, 46]密切相关；然而，我们在释义生成任务的半监督设置下研究这一问题。此外，我们提出了一种新颖的推理算法（即知识强化学习（KRL）方案），以帮助深度生成模型的学习，并在完整数据和不完整（部分）数据设置下都取得了有竞争力的性能。在文本深度潜变量模型方面，近期人们对学习离散潜在结构 [49]的兴趣日益浓厚。本文将潜变量建模为语言序列，为这一研究领域作出了贡献。此外，近期一些研究致力于将潜变量模型与扩散模型相结合，以在文本到音频生成任务中实现最先进的性能 [50, 51] 。

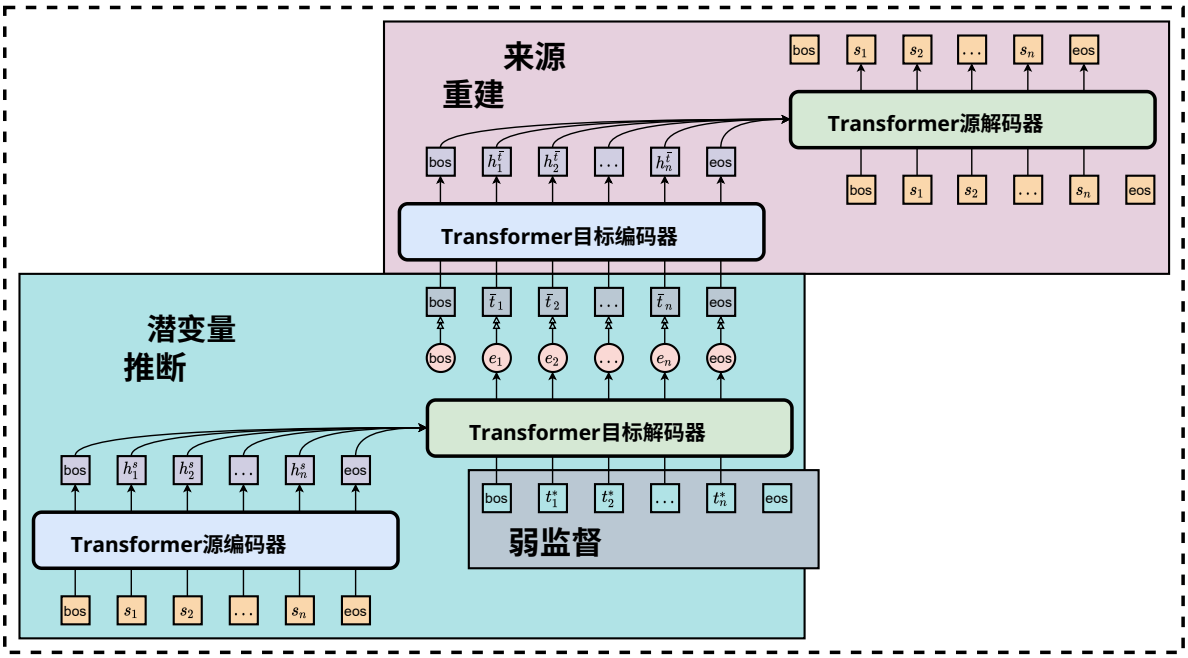

图 1: **变分序列自动编码重构模型**。

3 变分序列自动编码重构（VSAR）

本节介绍 VSAR 模型（图 1）[2]。该模型由四个独立的神经网络模型组成：源编码器、目标解码器、目标编码器和源解码器。在无监督学习设置下，我们只能观察到源文本 $\boldsymbol{s}$ ，而没有目标文本 $\boldsymbol{t}$。我们将对完全观测到的源文本 $\boldsymbol{s}$进行建模的问题，重新表述为对部分观测到的平行源文本 $\boldsymbol{s}$及其关联的潜在目标对 $\boldsymbol{t}$进行建模。我们采用贝叶斯推理，根据公式 7 对联合概率分布 $p_{\boldsymbol{\theta}}(\boldsymbol{s}, \bar{\boldsymbol{t}})$中的潜在目标字符串 $\bar{\boldsymbol{t}}$ 进行边缘化，如图 1 所示。

因此，在 VSAR 模型中，参数化为 $q_{\boldsymbol{\phi}}(\bar{\boldsymbol{t}}|\boldsymbol{s})$ 的**潜在推断**网络接收源文本 $\boldsymbol{s}$ 并生成一个潜在目标样本 $\bar{\boldsymbol{t}}$。参数化为 $p_{\boldsymbol{\theta}}(\boldsymbol{s}|\bar{\boldsymbol{t}})$ 的**源重建**网络根据潜在目标样本 $\boldsymbol{t}$重建观测到的源文本 $\boldsymbol{s}$ 。作为先验分布，我们在未标注源文本语料库上预训练一个语言模型，以近似先验分布 $p(\bar{\boldsymbol{t}})$[3]。引入先验是为了进行正则化[32, 46], ，以确保生成的样本更可能是“合理”的自然语言文本。

受多任务学习中自然语言生成的**参数共享** 所带来的益处启发 [52, 53, 54, 55],，我们共享源编码器和目标编码器的模型参数，记作 $\boldsymbol{f_{encode}}$；同样，我们共享

[2]为清晰起见，这里省略语言模型先验和弱监督解码。[3]我们利用释义生成任务中的语言学知识，在该任务中，释义文本字符串可以被视为其自身的释义。

源解码器和目标解码器的模型参数，记作 $\boldsymbol{f_{decode}}$。在以下各节中，我们分别用$\boldsymbol{f_{encode}}$ 和 $\boldsymbol{f_{decode}}$ 表示 VSAR 模型中的所有编码器和解码器。

3.1 弱监督

在 VSAR 模型中，我们通过实证发现，潜在序列 $\boldsymbol{t}$ 的质量非常不稳定，尤其是在训练初期。为了解决这一问题，受弱监督理念的启发 [56, 57], ，我们提出在整个训练过程中使用伪标签来引导 VSAR。在每个模型使用反向传播算法进行前向传播之前，我们首先根据当前模型参数（来自上一次迭代），为未观测到的潜在目标样本 $\boldsymbol{t}$ 中的每个词元分配伪标签。伪标签会从计算图中分离，因此在弱监督过程中不会更新梯度。伪标签可以视为“教师强制训练” [58]的弱监督信号。

编码器模型将源字符串 $\boldsymbol{s} = (s_1, \ldots\ldots, s_n)$作为输入，并生成对应的上下文向量$\boldsymbol{h^s} = (h_1^s, \ldots\ldots, h_n^s)$:

$$\boldsymbol{h^s} = \boldsymbol{f_{encode}}(\boldsymbol{s}) \tag{1}$$

我们采用贪心解码方案来分配伪目标标签 $\boldsymbol{t^*}$，并假设良好的释义应与原句长度相近 [59, 60]；即 $\boldsymbol{t^*} = (t_1^*, \ldots\ldots, t_n^*)$。令 t_i^*为伪目标序列中的第 i^{th} 个词；我们以自回归方式构造该序列:

$$t_i^* = \boldsymbol{f_{decode}}(\boldsymbol{h^s}; t_{1:i-1}^*) \tag{2}$$

3.2 目标推断

分配伪目标标签 $\boldsymbol{t^*}$ 后，我们使用潜变量推断网络执行潜变量推断。由于源字符串 $\boldsymbol{s}$ 保持不变，我们沿用弱监督部分中的上下文向量 $\boldsymbol{h^s}$ 。令 $\bar{t}_j$ 为潜在样本中的 j^{th} 个词， e_j 为目标解码器模型的相应输出。我们使用上下文向量 $\boldsymbol{h_s}$ 和伪标签中的全部 $t_{1:j-1}^*$ 个词构造潜在样本 $\boldsymbol{t}$：

$$\begin{aligned} e_j &= \boldsymbol{f_{decode}}(\boldsymbol{h_s}; t_{1:j-1}^*) \\ \bar{t}_j &\sim \textit{Gumbel}\text{-TOP}k(e_j, \tau) \end{aligned} \tag{3}$$

此处， t_i 通过 Gumbel 技巧 [43, 44] 抽取，温度 τ 作为额外的超参数，用于控制样本的概率分布。温度较高时 τ，抽样等价于从均匀分布中抽取；温度较低时 τ，抽样等价于从类别分布中抽取。由于词汇表规模庞大，从该单纯形中抽样可能很困难，因此我们进一步采用 TOP-k 子集采样技术 [48] 来提高采样效率。

我们探讨文献中常用的两种不同方案：（1）采用固定温度 τ 0.1，如[61]所示；（2）逐渐将温度 τ 从高温 10 退火至低温 0.01，如[62]所示。我们的实证结果表明，在训练期间对温度 τ 进行退火可显著提升效果（$p < .05$；Wilcoxon检验），因此最终结果均采用这一方案。我们使用 10 作为 k值，这一设置参考了 [21]。

3.3 源重建

对于源重建网络，编码器模型将潜在目标序列字符串 $\bar{\boldsymbol{t}} = (\bar{t}_1, ..., \bar{t}_n)$ 作为输入，并生成其对应的上下文向量 $\boldsymbol{h^{\bar{t}}} = (h_1^{\bar{t}}, ..., h_n^{\bar{t}})$:

$$\boldsymbol{h^{\bar{t}}} = \boldsymbol{f_{encode}}(\bar{\boldsymbol{t}}) \tag{4}$$

设$\hat{s}_k$ 为训练过程中重建源字符串中的 k^{th} 词；我们通过以下方式检索重建源字符串$\hat{\boldsymbol{s}}$：

$$\hat{s}_k = \boldsymbol{f_{decode}}(\boldsymbol{h^{\bar{t}}}; s_{1:k-1}) \tag{5}$$

3.4 VSAR 的学习与推断

在 SVAR 模型中，有两组需要更新的参数： ϕ 和 $\boldsymbol{\theta}$。设 $\boldsymbol{S}$ 为源文本的观测随机变量， $\boldsymbol{T}$为目标文本的潜随机变量， N 为未标注源文本的总数。对于参数化为 $\boldsymbol{\theta}$的 SVAR 模型，我们有如下联合似然:

$$p(\boldsymbol{S}, \bar{\boldsymbol{T}}; \boldsymbol{\theta}) = \prod_{i=1}^{N} p(s_{(i)}|\bar{t}_{(i)}; \boldsymbol{\theta}) p(\bar{t}_{(i)}) \tag{6}$$

训练过程中将近似的观测数据对数边际似然 $\boldsymbol{L_1}$ 为 $\log\ p(\boldsymbol{S};\boldsymbol{\theta})$。我们采用摊销变分推断 [17, 18, 19]，并构建一个由神经网络 $\_q(\boldsymbol{T}|\boldsymbol{S};\ \boldsymbol{\phi})$ 近似的替代函数，该网络由 ϕ 参数化，用于推导联合似然的证据下界（ELBO）:

$$\begin{aligned} \boldsymbol{L_1} &= \log \sum_{\bar{\boldsymbol{T}}} p(\boldsymbol{S}, \bar{\boldsymbol{T}}; \boldsymbol{\theta}) \geq \mathcal{L}_{ELBO}(\boldsymbol{S}, \bar{\boldsymbol{T}}; \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= \sum_{i=1}^{N} \{\mathbb{E}_{q(\bar{t}|s_{(i)};\boldsymbol{\phi})}[\log p(s_{(i)}|\bar{t};\boldsymbol{\theta})] - \mathbb{D}_{KL}[q(\bar{t}|s_{(i)};\boldsymbol{\phi})||p(\bar{t})]\} \end{aligned} \tag{7}$$

变分自编码器框架中最常见的变分族依赖于重参数化技巧 [17], , 但该技巧不适用于不可微的离散潜变量。一种使用此类潜变量优化学习的方法是 REINFORCE算法 [19, 42]；然而，该算法通常存在高方差问题。本文则使用 Gumbel-Softmax [43, 44]与可微子集采样 [48], , 以无放回方式获取前-k个样本。不过，由于对独热形式向量进行采样会导致高方差，我们采用直通技巧 [63] 作为梯度的有偏估计量，以降低这种方差。

在训练过程中，优化对数似然时，我们会同时进行学习（$\boldsymbol{\theta}$）和推断（$\boldsymbol{\phi}$）。这些参数使用同一个优化器进行联合优化。由于我们的模型共享参数，实际上，我们只使用源数据更新同一组参数（由 $\boldsymbol{\theta}$ 和 $\boldsymbol{\phi}$共享）。

4 双向学习（DDL）

本节将介绍双向学习（DDL）模型，我们使用该模型进行有监督释义生成。DDL 模型由两组标准 Transformer 模型 [64], 组成，每组模型都包含两个彼此独立的神经网络——一个编码器和一个解码器。我们利用完全可观测的平行源文本 $\boldsymbol{s}$ 及其关联的目标文本对 $\boldsymbol{t}$，在两个方向上执行标准的序列到序列学习。**目标生成** 网络$p_{\boldsymbol{\theta}_{t|s}}$（$\boldsymbol{t}|\boldsymbol{s}$）以源文本 $\boldsymbol{s}$ 为输入，生成目标文本 $\boldsymbol{t}$；而**源生成** 网络 $p_{\boldsymbol{\theta}_{s|t}}$（$\boldsymbol{s}|\boldsymbol{t}$）以目标文本 $\boldsymbol{t}$ 为输入，生成源文本 $\boldsymbol{s}$。

4.1 参数学习

在 DDL 模型中，有两组需要更新的参数， $\boldsymbol{\theta_{s|t}}$ 和 $\boldsymbol{\theta_{t|s}}$。令 $\boldsymbol{S}$ 为源文本的观测随机变量， $\boldsymbol{T}$为目标文本的观测随机变量， M 为带标签文本对的数量；那么，我们的 DDL 模型具有如下条件似然:

$$\begin{aligned} p(\boldsymbol{S}|\boldsymbol{T}; \boldsymbol{\theta_{s|t}}) &= \prod_{i=1}^{M} p(s_{(i)}|t_{(i)}; \boldsymbol{\theta_{s|t}}) \\ p(\boldsymbol{T}|\boldsymbol{S}; \boldsymbol{\theta_{t|s}}) &= \prod_{i=1}^{M} p(t_{(i)}|s_{(i)}; \boldsymbol{\theta_{t|s}}) \end{aligned} \tag{8}$$

The 观测数据对的对数条件似然 $\boldsymbol{L_2}$ 可以在训练过程中联合学习。 s:

$$\boldsymbol{L_2} = \sum_{i=1}^{M} (\log p(s_{(i)}|t_{(i)}; \boldsymbol{\theta_{s|t}}) + \log p(t_{(i)}|s_{(i)}; \boldsymbol{\theta_{t|s}})) \tag{9}$$

训练期间，我们同时进行双向学习（$\boldsymbol{\theta_{s|t}}$ 和 $\boldsymbol{\theta_{t|s}}$），并使用同一个优化器联合优化参数。

4.2 参数共享

再次受到多任务学习对自然语言生成的益处的启发， [52, 53, 54, 55], 我们为目标生成网络和源生成网络共享模型参数。尽管参数共享是一种非常简单的技术，但如表1和表2所示，与仅进行单向序列到序列学习的 Transformer 基线模型相比，DDL 模型显著提升了释义生成的性能（$p < .05$；Wilcoxon检验）。

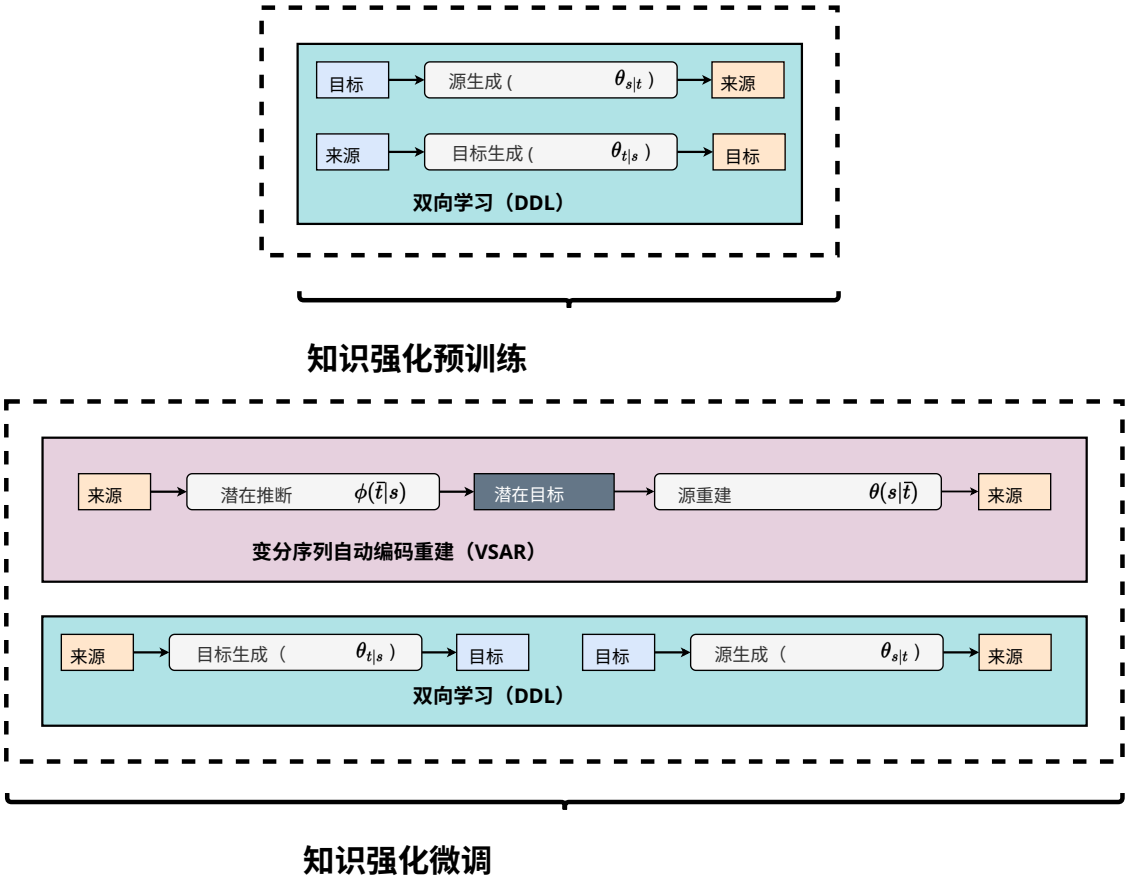

图 2：知识强化学习（**KRL**）。

5 结合 VSAR 和 DDL 进行半监督学习

本节将介绍我们的半监督学习模型（VSAR+DDL），该模型结合了前几节介绍的模型。对于半监督学习，数据的对数似然可表示如下：

$$
\begin{aligned}
&\boldsymbol{L} = \boldsymbol{L_1} + \boldsymbol{L_2} \\
&= \sum_{i=1}^{N} \{\mathbb{E}_{q(\bar{t}|s_{(i)};\boldsymbol{\phi})}[\log p(s_{(i)}|\bar{t};\boldsymbol{\theta})] - \mathbb{D}_{KL}[q(\bar{t}|s_{(i)};\boldsymbol{\phi})||p(\bar{t})]\} \\
&+ \sum_{i=1}^{M} (\log p(s_{(i)}|t_{(i)};\boldsymbol{\theta_{s|t}}) + \log p(t_{(i)}|s_{(i)};\boldsymbol{\theta_{t|s}}))
\end{aligned}
\tag{10}
$$

如公式 10 所示，对于无监督学习和监督学习，似然函数涉及 $\boldsymbol{s}$ 与 $\boldsymbol{t}$之间相同的一组条件概率。我们假设，在这两个模型之间共享参数是有益的。因此，我们共享 VSAR 和 DDL 模型的两组神经网络参数（即$q_{\boldsymbol{\phi}}(\bar{\boldsymbol{t}}|\boldsymbol{s})$ $\equiv p_{\boldsymbol{\theta}_{t|s}}(\boldsymbol{t}|\boldsymbol{s})$ 和 $p_{\boldsymbol{\theta}}(\boldsymbol{s}|\bar{\boldsymbol{t}}) \equiv p_{\boldsymbol{\theta}_{s|t}}(\boldsymbol{s}|\boldsymbol{t})$）。这使得 DDL 模型的强监督信号能够直接作用于 VSAR 模型。同时，VSAR 模型的无监督信号也有助于提升 DDL 模型的泛化能力。

5.1 知识强化学习

我们的实证实验表明，在从头开始进行半监督学习时，我们的组合模型（DDL+VSAR）在参数优化方面存在冷启动问题。我们发现，模型成功的关键在于更好地初始化模型权重。因此，我们提出了一种名为**知识强化学习**的新型训练方案（图 2），其中包括两阶段训练。第一阶段（预训练）中，我们按照算法 1 所示，使用 DDL 模型在配对训练集上进行监督学习。第二阶段（微调）中，我们使用第一阶段性能最佳的 DDL 模型来初始化 VSAR 模型参数；然后按照算法 2 所示，使用带标签数据和未标注数据进行半监督学习。其直观思路是将更好的先验信息注入 VSAR 模型的训练中。

算法 1知识强化预训练

输入：有监督训练数据 ($\mathcal{D}_T^S = \{(s_1,\ t_1)$ ..... $(s_N,\ t_N)\}$)，有监督验证数据 ($\mathcal{D}_V^S$)**参数**：
DDL 模型：$\boldsymbol{\theta}_{\boldsymbol{s}|\boldsymbol{t}}$ 和 $\boldsymbol{\theta}_{\boldsymbol{t}|\boldsymbol{s}}$**参数共享**：在整个知识强化预训练过程中，令 $\boldsymbol{\theta}_{\boldsymbol{s}|\boldsymbol{t}}$ 等于 $\boldsymbol{\theta}_{\boldsymbol{t}|\boldsymbol{s}}$

输出：$\boldsymbol{\theta}_{\boldsymbol{s}|\boldsymbol{t}}{}^*$ 和 $\boldsymbol{\theta}_{\boldsymbol{t}|\boldsymbol{s}}{}^*$
1：使用随机种子初始化 $\boldsymbol{\theta}_{\boldsymbol{s}|\boldsymbol{t}}$ 和 $\boldsymbol{\theta}_{\boldsymbol{t}|\boldsymbol{s}}$；将最大训练轮次设为 $\boldsymbol{T}$；设置 $\boldsymbol{L_2}^* = 0$2：**当**尚未达到最大轮次**时** 3：根据公式 9，使用来自 $\mathcal{D}_T^S$的小批量数据更新 $\boldsymbol{\theta}_{\boldsymbol{s}|\boldsymbol{t}}$ 和 $\boldsymbol{\theta}_{\boldsymbol{t}|\boldsymbol{s}}$ 4：**如果** 根据 $\mathcal{D}_V^S$计算的公式 9 中的 $\boldsymbol{L_2}$ 大于 $\boldsymbol{L_2}^*$ **则**5：设置 $\boldsymbol{L_2}^* \leftarrow \boldsymbol{L_2}$6：设置 $\boldsymbol{\theta}_{\boldsymbol{s}|\boldsymbol{t}}{}^* \leftarrow \boldsymbol{\theta}_{\boldsymbol{s}|\boldsymbol{t}}$7：设置 $\boldsymbol{\theta}_{\boldsymbol{t}|\boldsymbol{s}}{}^* \leftarrow \boldsymbol{\theta}_{\boldsymbol{t}|\boldsymbol{s}}$8：**结束条件**9：**结束循环 返回**：$\boldsymbol{\theta}_{\boldsymbol{s}|\boldsymbol{t}}{}^*$ 和 $\boldsymbol{\theta}_{\boldsymbol{t}|\boldsymbol{s}}{}^*$

5.2 语言模型先验的影响

在文献中 [38, 32, 65, 46], , 引入语言模型先验用于正则化，以促使样本更有可能包含“合理”的自然语言，尤其是在训练初期。因此，我们采用相同的方法，并在模型中使用先验。我们通过实证发现，当有标签数据集相对较小时，先验是有用的。然而，令人惊讶的是，当数据集较大时，使用我们的参数初始化方法，在 VSAR 模型中不使用先验进行训练能取得更好的结果。改进具有显著性（$p < .05$；Wilcoxon检验），如表 5 和表6所示。我们将不使用语言模型先验的结果记为 DDL +VSAR*，其对数似然为：

$$\begin{aligned}\boldsymbol{L}^* = &\sum_{i=1}^{N}\{\mathbb{E}_{q(\bar{t}|s_{(i)};\boldsymbol{\phi})}[\log p(s_{(i)}|\bar{t};\boldsymbol{\theta})]\} \\ &+\sum_{i=1}^{M}(\log p(s_{(i)}|t_{(i)};\boldsymbol{\theta}_{\boldsymbol{s}|\boldsymbol{t}}) + \log p(t_{(i)}|s_{(i)};\boldsymbol{\theta}_{\boldsymbol{t}|\boldsymbol{s}}))\end{aligned} \tag{11}$$

为进一步研究这一问题，我们进行了实验，比较在不同数据比例设置下，使用先验（公式 10）或不使用先验（公式 11）进行训练时半监督学习的性能。我们通过实验发现，当带标签数据比例较低时，带先验的组合模型（DDL+VSAR）能够显著提升（$p < .05$；

算法 2知识强化微调

输入:
无监督数据（$\mathcal{D}^U = \{s_1, ..., s_M\}$）有监督训练数据（$\mathcal{D}_T^S = \{$（$s_1, t_1$）...,（$s_N, t_N$）$\}$），有监督验证数据（$\mathcal{D}_V^S$）参数：VSAR 模型： $\boldsymbol{\phi}$ 和 $\boldsymbol{\theta}$；DDL 模型： $\boldsymbol{\theta_{s|t}}$ 和 $\boldsymbol{\theta_{t|s}}$参数共享：在整个知识强化微调过程中，设置 $\boldsymbol{\phi}$ 等于 $\boldsymbol{\theta_{t|s}}$； $\boldsymbol{\theta}$ 等于 $\boldsymbol{\theta_{s|t}}$；且 $\boldsymbol{\theta_{s|t}}$ 等于 $\boldsymbol{\theta_{t|s}}$ 输出： $\boldsymbol{\theta_{s|t}}^{**}, \boldsymbol{\theta_{t|s}}^{**}$； $\boldsymbol{\phi}^{**}$ 和 $\boldsymbol{\theta}^{**}$1：使用 $\boldsymbol{\theta_{t|s}}^*$初始化 $\boldsymbol{\phi}$ 和 $\boldsymbol{\theta_{t|s}}$ ；并使用 $\boldsymbol{\theta_{s|t}}^*$初始化 $\boldsymbol{\theta}$ 和 $\boldsymbol{\theta_{s|t}}$ ；将最大训练轮次设为 $\boldsymbol{T}$；设置$\boldsymbol{L_2}^* = 0$。2：当最大轮次尚未达到时，执行以下操作。3：根据公式 9 4 $\boldsymbol{\phi}$ $\boldsymbol{\theta}$ $\mathcal{D}^U$ 7 ，使用来自 $\mathcal{D}_T^S$ 5 $\boldsymbol{L_2}$ 9 $\mathcal{D}_V^S$ $\boldsymbol{L_2}^*$ 的小批量数据更新 $\boldsymbol{\theta_{s|t}}$ 和 $\boldsymbol{\theta_{t|s}}$ ：根据公式 ，使用小批量数据更新 和 ：如果公式 中的 计算结果基于 大于 ，则6：设置 $\boldsymbol{L_2}^* \leftarrow \boldsymbol{L_2}$7：设置 $\boldsymbol{\theta_{s|t}}^{**} \leftarrow \boldsymbol{\theta_{s|t}}$8：设置 $\boldsymbol{\theta_{t|s}}^{**} \leftarrow \boldsymbol{\theta_{t|s}}$9：设置 $\boldsymbol{\phi}^{**} \leftarrow \boldsymbol{\phi}$10：设置 $\boldsymbol{\theta}^{**} \leftarrow \boldsymbol{\theta}$11：结束条件判断12：结束循环返回： $\boldsymbol{\theta_{s|t}}^{**}, \boldsymbol{\theta_{t|s}}^{**}$； $\boldsymbol{\phi}^{**}$ 和 $\boldsymbol{\theta}^{**}$

Wilcoxon检验）性能更好且更加稳定。这与 [38, 32, 65, 46] 中的观察结果一致。然而，当带标签数据占比较大时，不使用先验的组合模型（DDL+VSAR）表现显著更好（$p < .05$；Wilcoxon检验）。

我们认为，这种现象与先验的选择有关，因为该先验是在未标注源文本语料库上预训练的，而非在目标文本语料库上预训练的。这种近似导致其与真实先验分布 $p(\bar{\boldsymbol{t}})$ 之间存在分布偏移。因此，在算法 1 中使用较少比例的带标签数据时，用于算法 2 中初始化 VSAR 模型的最终 DDL 参数 $\boldsymbol{\theta_{s|t}}^*$ 和 $\boldsymbol{\theta_{t|s}}^*$ 不够理想。在这种情况下，先验仍能使组合模型受益于半监督学习。然而，当带标签数据占比较大时，初始化已经足够好，此时分布偏移可能会损害组合模型。

5.3 半监督学习设置

在半监督学习设置下，我们将有监督源目标对的数量限制为小于或等于无监督源文本的数量（$M \leq N$），否则我们完全可以采用监督学习，以充分利用观测数据对。本文全面研究了 M和 N的不同规模。在此设置下的实验结果见表1、表2、表3和表4。

6 实验

这里，我们介绍数据集、实验设置、评估指标和实验结果 .

6.1 数据集

MSCOCO [66]：该数据集已被广泛用于评估释义生成方法，其中包含人工标注的图像描述。每张图像对应五条由不同标注者撰写的描述，他们会描述图像中最显著的物体或动作。我们在实验中使用2017年版本；对于每张图像所附的五条描述，我们随机选取一条作为源字符串，一条作为目标字符串用于训练。测试时，我们随机选取一条作为源字符串，并将其余四条作为参考描述。

Quora[4]：该数据集由 150K 行问题重复对组成，自 2017年 起便被用作释义生成的基准数据集。然而，由于该数据集没有专门划分训练集和测试集，先前的模型都是根据不同规模的数据子集进行评估的。

对于两个数据集（MSCOCO 和 Quora），为了提高结果的可复现性，我们使用 [67][5] 提供的预训练分词器（“bert-base-uncased”版本），并将最大词元长度设为 20（删除前 20 个词元之后的所有词元）。按照 [68, 21, 13], 的做法，Quora数据集的训练集、验证集和测试集分别设为 100K、4K 和 20K；MSCOCO 的训练集、验证集和测试集分别设为 93K、4K 和 20K。对于表7和表8中的补充研究，为了与 [22, 6] 报告的结果进行公平比较，我们将 Quora数据集的训练集、验证集和测试集分别设为 100K、24K 和 24K；将 MSCOCO 的训练集、验证集和测试集分别设为 100K、5K 和 5K。

其他可用于释义生成任务的数据集包括 ParaBank [69] 和 PARANMT [70] , 。这两个大规模数据集均通过回译技术，利用其他非英语平行语料库创建。由于研究文献中较少采用这两个数据集，我们无法将其与现有研究直接比较。不过，作为替代方案，为了进一步证明我们提出的模型的有效性，我们在这两个数据集上开展了半监督学习实验。对于ParaBank数据集，我们采用了“ParaBank v1.0（500 万对）”[6]，并使用与 Quora数据集类似的设置。该数据集包含 500 万行重复的句对；我们随机选取相同的数据子集用于实验，并分别使用大小为 100K、4K 和 20K 的训练集、验证集和测试集。对于 PARANMT数据集，我们首先筛选释义句对（删除 paragram-phrase 分数[7]高于 0.95 或低于 0.90 的条目；将最大词元长度设为 20，最小词元长度设为 5），并按照 [70], 的建议保留中间百分位范围，以去除噪声和过于简单的释义。筛选完成后，我们采用与 Quora数据集类似的实验设置，随机选取相同的数据子集用于实验，并分别使用大小为 100K、4K 和 20K 的训练集、验证集和测试集。

6.2 基线方法

我们考虑了几种最先进的基线方法，详见表 5、表6、表7 和表8。请注意，这些实验结果直接取自 [13][8] 和 [6]。在评估中，我们首先将自行实现的 Transformer 模型作为绝对基线；正如 [13]所报告的，该模型取得了具有竞争力的性能。Transformer 模型 [64] 被视为最先进水平的模型，非常“难以超越”。我们基于与 [13] 和 [6]中类似的设置报告模型性能。

近年来，大规模预训练语言模型（PLM）已被广泛用作 NLP 领域中理解与生成任务的先进方法；然而，本文并未将 PLM 选作竞争基线模型，因为它们包含基于大规模文本语料以无监督方式训练得到的外部信息。相反，本文着重于从零开始对释义生成任务进行端到端学习。这样的实验设置使我们能够直接与其他文献中的方法进行比较。因此，本文不与任何 PLM 进行比较，将这一点留待未来工作。不过请注意，各类 PLM 模型主要都基于 Transformer架构，而我们将其用作基线模型进行比较。这意味着，我们的方法有望在带标签资源有限、未标注数据规模庞大的情况下，帮助提升 PLM 的性能。

表1： **Quora 半监督学习实验结果**。

模型	已标注	未标注	B-1	B-2	B-3	B-4	i-B	R-1	R-2	R-L
DDL	20K	–	46.68	33.44	25.46	20.18	11.08	47.57	25.42	45.50
DDL+VSAR$_1$	20K	20K	47.80 ↑	34.33 ↑	26.17 ↑	20.76 ↑	11.25 ↑	48.03 ↑	25.82 ↑	45.84 ↑
DDL+VSAR$_2$	20K	100K	**50.26** ↑	**36.87** ↑	**28.50** ↑	**22.82** ↑	**11.60** ↑	**51.51** ↑	**28.45** ↑	**49.07** ↑
DDL	50K	–	53.31	40.22	31.70	25.80	**13.80**	**55.63**	**32.15**	**53.13**
DDL+VSAR$_1$	50K	50K	53.33 ↑	39.93 ↓	31.39 ↓	25.49 ↓	13.45 ↓	55.51 ↓	31.90 ↓	52.95 ↓
DDL+VSAR$_2$	50K	100K	**53.79** ↑	**40.47** ↑	**31.86** ↑	**25.93** ↑	13.67 ↓	55.58 ↓	31.89 ↓	52.93 ↓

[4]https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs [5] https://github.com/huggingface/transformers [6]https://nlp.jhu.edu/parabank/ [7]paragram-phrase 分数用于衡量一对句子之间的语义相似度。对于完整的 PARANMT 数据集，平均分数为 0 43，0 69，0 26 0 95 09，.，标准差为 . .。我们选择均值一个标准差范围内的范围，即 (. .)。由于我们希望数据集中的句子确实构成释义，因此进一步将下界限制为 .，以确保数据的质量。[8]作者没有公开其代码。

表2： **MSCOCO 半监督学习实验结果。**

模型	已标注	未标注	B-1	B-2	B-3	B-4	i-B	R-1	R-2	R-L
DDL	20K	–	66.82	47.25	33.14	23.75	16.66	40.53	14.95	36.94
DDL+VSAR$_1$	20K	20K	66.98 ↑	47.28 ↑	33.10 ↓	23.72 ↓	16.54 ↓	40.60 ↑	14.95 ↑	36.94 ↑
DDL+VSAR$_2$	20K	93K	**67.64** ↑	**48.00** ↑	**33.96** ↑	**24.55** ↑	**16.68** ↑	**40.87** ↑	**15.12** ↑	**37.01** ↑
DDL	50K	–	69.39	50.17	36.06	26.49	18.43	42.08	16.31	38.27
DDL+VSAR$_1$	50K	50K	69.43 ↑	50.21 ↑	36.08 ↑	26.45 ↓	18.31 ↓	42.20 ↑	16.33 ↑	38.31 ↑
DDL+VSAR$_2$	50K	93K	**69.91** ↑	**50.65** ↑	**36.52** ↑	**26.93** ↑	**18.51** ↑	**42.39** ↑	**16.46** ↑	**38.40** ↑

表3： **ParaBank 的半监督学习实验结果。**

模型	已标注	未标注	B-1	B-2	B-3	B-4	i-B	R-1	R-2	R-L
DDL	20K	–	50.48	39.21	31.11	25.16	12.98	57.03	33.77	55.31
DDL+VSAR$_1$	20K	20K	50.69 ↑	39.41 ↑	31.35 ↑	25.45 ↑	13.14 ↑	56.99 ↓	33.92 ↑	55.20 ↓
DDL+VSAR$_2$	20K	100K	**54.04** ↑	**43.41** ↑	**35.44** ↑	**29.40** ↑	**14.94** ↑	**60.66** ↑	**37.50** ↑	**58.90** ↑
DDL	50K	–	58.05	48.84	41.59	35.82	18.19	66.45	44.63	64.81
DDL+VSAR$_1$	50K	50K	58.29 ↑	49.24 ↑	42.10 ↑	36.40 ↑	18.45 ↑	66.88 ↑	45.23 ↑	65.23 ↑
DDL+VSAR$_2$	50K	100K	**59.22** ↑	**50.33** ↑	**43.23** ↑	**37.51** ↑	**18.89** ↑	**67.64** ↑	**46.05** ↑	**65.98** ↑

表4： **PARANMT 的半监督学习实验结果。**

模型	已标注	未标注	B-1	B-2	B-3	B-4	i-B	R-1	R-2	R-L
DDL	20K	–	62.49	49.47	39.92	32.84	17.40	63.13	40.43	61.15
DDL+VSAR$_1$	20K	20K	63.48 ↑	50.70 ↑	41.32 ↑	34.33 ↑	18.16 ↑	64.59 ↑	41.97 ↑	62.58 ↑
DDL+VSAR$_2$	20K	100K	**66.46** ↑	**54.20** ↑	**44.80** ↑	**37.60** ↑	**19.42** ↑	**67.55** ↑	**45.25** ↑	**65.49** ↑
DDL	50K	–	70.55	59.53	50.79	43.78	22.82	71.97	51.30	69.97
DDL+VSAR$_1$	50K	50K	70.41 ↓	59.48 ↓	50.81 ↑	43.87 ↑	23.02 ↑	72.06 ↑	51.45 ↑	70.07 ↑
DDL+VSAR$_2$	50K	100K	**70.91** ↑	**60.12** ↑	**51.48** ↑	**44.53** ↑	**23.25** ↑	**72.57** ↑	**52.03** ↑	**70.57** ↑

6.3 实验设置

本节介绍我们的主要部分实验设置。我们不使用任何外部词嵌入（例如 Glove [71], word2vec[72] 或 BERT [67] ）进行初始化；相反，为了不使用任何先验知识并更好地理解模型的影响，我们通过端到端训练获得词嵌入。我们使用 Transformer 模型的“base”版本 [64], ，该模型包含 6 层、512 个隐藏单元，并且每个编码器和解码器网络各有 8 个注意力头。在每个编码器和解码器中，我们都使用独立的可学习位置嵌入及其对应的词嵌入组件。

我们采用贪心解码方案进行释义生成，该方案计算速度快且成本低。对于模型优化，我们使用 Adam [73] 作为优化器，并采用默认超参数（$\beta_1 = 0.9$、 $\beta_2 = 0.999$、 $\epsilon = 1e-8$）。对于 Quora 和 MSCOCO 数据集，所有实验均采用 512 的批量大小。根据经验实验结果，我们将 MSCOCO 的学习率设为 $1e-4$，将 Quora 的学习率设为 $2e-4$。所有实验均在 NVidia GPU 集群的 A100 GPU 上运行，最多进行 30 个周期。实验使用不同的随机种子（1000、2000 和 3000）重复三次，并在表 1—6 中报告平均结果。

表 5： **Quora 的实验结果。**

模型	B-1	B-2	B-3	B-4	i-B	R-1	R-2	R-L
上界（复制源文本）	63.36	49.99	40.47	33.54	-	63.04	38.15	59.64
下界（随机选择）	16.10	4.50	1.94	0.79	-	9.13	1.54	8.79
残差长短期记忆网络 [74]	53.59	39.49	30.25	23.69	15.93	55.10	33.86	53.61
β-变分自编码器 [38]	47.86	33.21	24.96	19.73	10.28	47.62	25.49	45.46
Transformer [64]	53.56	40.47	32.11	25.01	17.98	57.82	32.58	56.26
LBOW-TOPk [21]	55.79	42.03	32.71	26.17	19.03	58.79	34.57	56.43
IANet+X [13]	56.06	42.69	33.38	26.52	19.62	59.33	35.01	57.13
Transformer（我们的实现）	54.73	41.59	32.96	26.94	14.50	56.90	33.28	54.29
DDL（我们的模型）	55.97 ↑	43.02 ↑	34.32 ↑	28.19 ↑	14.83 ↑	**58.80** ↑	35.00 ↑	56.11 ↑
DDL + SVAR（我们的模型）	55.79 ↑	42.79 ↑	34.11 ↑	28.01 ↑	**14.92** ↑	58.61 ↑	34.75 ↑	55.91 ↑
DDL + SVAR* （我们的模型）	**55.99** ↑	**43.05** ↑	**34.37** ↑	**28.23** ↑	14.81 ↑	58.79 ↑	**35.02** ↑	**56.14** ↑

表6： **MSCOCO 实验结果**。

模型	B-1	B-2	B-3	B-4	i-B	R-1	R-2	R-L
上界（复制源文本）	64.97	44.90	30.69	21.30	-	39.18	12.96	34.61
下界（随机选择）	32.34	10.99	3.81	1.68	-	17.58	1.51	16.27
残差长短期记忆网络 [74]	70.24	48.65	34.04	23.66	18.72	41.07	15.26	37.35
β-变分自编码器 [38]	70.04	47.59	32.29	22.54	18.34	40.72	14.75	36.75
Transformer [64]	71.31	49.86	35.55	24.68	19.81	41.49	15.84	37.09
LBOW-TOPk [21]	72.60	51.14	35.66	25.27	21.07	42.08	16.13	38.16
IANet+X [13]	72.10	52.22	37.39	26.06	21.28	43.81	16.35	39.65
Transformer（我们的实现）	68.72	49.64	35.87	26.63	18.59	42.09	16.53	38.35
DDL（我们的模型）	70.75 ↑	51.72 ↑	37.62 ↑	27.95 ↑	19.37 ↑	43.00 ↑	17.01 ↑	39.06 ↑
DDL + SVAR（我们的模型）	70.84 ↑	51.84 ↑	37.75 ↑	28.04 ↑	19.39 ↑	**43.05** ↑	**17.04** ↑	**39.07** ↑
DDL + SVAR* （我们的模型）	**70.99** ↑	**51.91** ↑	**37.82** ↑	**28.12** ↑	**19.39** ↑	43.00 ↑	17.03 ↑	39.02 ↑

6.4 评估

在本文中，我们首先基于 定量指标评估我们的模型：BLEU [75][9], 、ROUGE [76][10], 和改进型双语评估替补 [77]。这样做的主要理由是直接与文献中已有的工作进行比较；这些工作侧重于使用深度神经网络模型，从头开始进行释义生成任务的端到端学习。BLEU（双语评估替补）和 ROUGE（面向摘要评估的召回率替补）分数基于系统生成的释义与参考句子之间的“n元语法”覆盖率。它们已被广泛用于自动评估自然语言生成任务的质量和准确性。

以往研究表明，自动评估指标在释义识别任务中表现良好 [78]，并且在评估生成的释义时与人工判断高度相关 [79]。近期论文引入了其他改进型双语评估替补 [77] 指标，以兼顾生成结果与参考释义的忠实度（ BLEU）以及所引入的多样性程度（self-B）。除 self-B 外，所有指标的值越高，模型表现越好。

此外，我们在表9和表10中展示了基于Quora数据集的定性评估结果。我们的定性评估旨在考察以下两个方面：（1）在带标签的句对数据集规模不同时，我们提出的监督模型（DDL）与表现强劲的监督学习基线模型Transformer相比如何；（2）在带标签的句对数据集规模保持不变时，纳入更多未标注数据对我们提出的半监督模型（DDL+VSAR）有何益处。

6.5 结果与讨论

6.5.1 仅使用未标注数据进行学习

在一项初步实验中，我们探究了VSAR模型仅使用未标注数据执行释义生成任务的能力。开展这项实验是为了检验模型能否在没有带标签数据的帮助下，准确捕捉释义生成所需的信息。然而，初步实验结果表明，VSAR模型单独使用时无法生成高质量的释义，生成的句子往往难以理解或毫无意义。这些结果促使我们采用半监督学习方案来进行释义生成，为模型提供带标签数据的指导。对于使用VSAR进行无监督学习的情况，我们发现，尽管在完全无监督的设置下， $\boldsymbol{L}_1$ (公式7)所表示的下界在训练过程中有所下降，模型生成的释义质量仍然较低。这进一步证实了采用半监督学习方案的必要性；本文第5节将介绍该方案。

6.5.2 使用部分数据进行学习

本节展示基于部分带标签数据得出的结果，见表 1、2、3 和 4。在这四个表中，我们展示了两个模型的结果：监督学习模型 DDL 和半监督学习模型 DDL + VSAR。在半监督学习设置中，VSAR 使用未标注数据进行训练，而 DDL 使用带标签数据进行训练。DDL+VSAR$_1$ 模型使用规模相同、来自相同源和目标对的有标签数据集与未标注数据集，因此在这种情况下没有引入额外信息。DDL+VSAR$_2$ 模型则在现有有标签数据集之外，使用完整的未标注数据集，这才是真正的半监督设置。

结果表明，与仅使用带标签数据训练的监督式 DDL 模型相比，DDL+VSAR$_1$ 模型在大多数指标上表现相当或更好；尤其是在数据占比较低时（例如，20K 时的显著提升比 50K 时更明显）。

[9]https://www.nltk.org/ [10]
https://github.com/huggingface/datasets/tree/master/metrics/rouge

此外，在带标签数据规模固定的情况下，DDL+VSAR$_2$ 模型利用额外的未标注数据，性能显著优于两张表中报告的所有其他模型（$p < .05$；Wilcoxon检验），这意味着半监督学习在这一场景中确实有效。

表7：**Quora 补充结果**。

模型	B-4	self-B	i-B
Separator [22]	23.68	24.20	14.10
HRQ-VAE [6]	33.11	40.35	18.42
Transformer（我们的实现）	26.92	**35.33**	14.47
DDL + SVAR（我们的模型）	28.15 ↑	38.92 ↓	**14.73** ↑
DDL + SVAR* （我们的模型）	**28.16** ↑	39.07 ↓	14.71 ↑

表8：**MSCOCO 的补充结果**。

模型	B-4	self-B	i-B
Separator [22]	20.59	12.76	13.92
HRQ-VAE [6]	27.90	16.58	19.04
Transformer（我们的实现）	26.87	**13.50**	18.79
DDL +SVAR（我们的模型）	27.87 ↑	15.42 ↓	19.21 ↑
DDL +SVAR* （我们的模型）	**27.92** ↑	15.21 ↓	**19.29** ↑

6.5.3 使用完整数据进行学习

本节根据表 5 和表 6 中的所有带标签数据展示结果。每个表分为三个部分。第一部分展示基于测试集划分计算得出的上界[11]（将来源直接复制为释义）和下界（随机选择真实值作为释义），如 [80]所示。这些结果用于衡量模型的表现。第二部分展示近年来发表的主要最先进模型。第三部分展示我们自行实现的 Transformer 模型，我们将其视为绝对基线，并给出我们模型的结果。我们的实现与近期论文中报告的实现相比具有竞争力。对于我们的模型，DDL 是监督模型，DDL+VSAR 是半监督模型，而 DDL+VSAR*是未使用先验的模型。与最先进的监督模型相比，总体而言，我们的模型在两个数据集上均取得了统计显著更高的 BLEU分数和具有竞争力的 ROUGE分数（$p < .05$；Wilcoxon检验）。我们的补充实验结果见表7和表8，其中我们将结果与另外两个较新的最先进模型进行了比较。我们的模型再次取得了统计显著更优或具有竞争力的性能，高于所报告的结果（$p < .05$；Wilcoxon检验），这表明我们的半监督模型与最先进的监督基线具有竞争力。

6.5.4 带标签数据监督学习的定性评估

在表 9 中，我们展示了 Quora 测试数据集中的示例，以及我们提出的监督学习模型 DDL（见第 4 节）和一个性能很强的基线模型 Transformer（记为 TRANS）在使用不同数量训练数据时生成的相应模型输出。表中首先列出来源与黄金参考答案组成的文本对，随后给出分别使用 20K、50K 和 100K 个带标签的数据集对训练的模型（DDL 和 TRANS）的输出。每个示例均基于随机种子 1000 生成，以确保定性评估公平；此外，我们始终确保较小规模的带标签文本对是较大数据规模组中示例的子集，从而能够更好地量化增加未标注数据所带来的益处。

结果清楚地表明，使用的带标签数据越多，生成的释义在语义信息方面就越准确，与参考文本也越吻合。不过与此同时，我们也发现，带标签文本对数量较少时，DDL 的优势更加显著（即使用 20K 数据时，效果提升比使用 50K 和 100K 数据时更明显）。此外，与 TRANS 模型相比，我们的 DDL 模型在捕捉信息要旨方面展现出明显优势，能够捕捉关键细节（例如，在第二个示例中识别出哮喘和吸入器；在第四个示例中识别出 35000 而非 50000

[11]请注意，此处的上界并非指性能上界 而非指性能上界，而不是指一个完整的..... 满意度水平表现的指标。

表 9：**Quora数据集上，Transformer (TRANS) 与 DDL 模型在不同数量的带标签数据（括号内标示）下的部分释义生成结果。**

来源:	新闻学有哪些最好的课程?
参考:	新闻学方面最好的课程有哪些?
TRANS（20K）	哪款软件最适合初学者?
DDL（20K）	最好的数字营销课程是什么?
TRANS（50K）	那格浦尔最好的课程有哪些?
DDL（50K）	新闻学有哪些最好的课程?
TRANS（100K）	有哪些关于新闻学的最佳课程?
DDL（100K）	新闻学有哪些最好的课程?
来源:	没有吸入器时，什么方法可以缓解哮喘?
参考:	没有吸入器时，有哪些方法可以帮助哮喘患者?
TRANS（20K）	一个范围的范围与一个范围的 000r 发生碰撞时会怎样
DDL (20K)	怎样才能在一个月内不做波浪就存下我的钱?
TRANS (50K)	在 PPR 中可以做些什么工作?
DDL (50K)	没有吸入器的哮喘有什么用途?
TRANS（100K）	要成为紧急国际标准化比值，需要经过什么程序?
DDL（100K）	没有吸入器时，我该如何缓解哮喘?
来源:	怎样才能让 100 万用户注册我的应用程序?
参考:	怎样才能让我的社交应用程序获得一百万用户?
TRANS（20K）	怎样才能开一家朋友商店?
DDL（20K）	怎样才能在我的应用程序上获得一款游戏应用程序?
TRANS (50K)	我怎样才能让我的创业公司赚到一百万个百万?
DDL (50K)	我怎样才能让我的应用程序拥有第一百万用户?
TRANS (100K)	我怎样才能让一百万用户来编写应用程序?
DDL（100K）	怎样才能让一百万用户注册我的应用程序
来源:	有人能推荐一款印度售价低于 3.5 万的最佳笔记本电脑吗?
参考:	哪款笔记本电脑最好，售价低于 35,000 印度卢比?
TRANS（20K）	有哪些适合笔记本电脑的最佳选择?
DDL（20K）	在印度买哪款笔记本电脑最好?
TRANS（50K）	在印度，50000卢比以下哪款笔记本电脑最好?
DDL（50K）	印度 35,000 以内最值得购买的笔记本电脑有哪些?
TRANS（100K）	印度 50,000 卢比以内最值得购买的笔记本电脑是哪一款?
DDL（100K）	印度 35,000 以内最值得购买的笔记本电脑是哪一款?
来源:	如何管理我的 Microsoft 账户?
参考:	如何管理我的 Microsoft 账户?
TRANS (20K)	如何管理我的 Google 账户?
DDL（20K）	如何管理我的 Microsoft Office?
TRANS（50K）	如何管理我的 IP 地址?
DDL（50K）	如何管理 Microsoft Microsoft Office?
TRANS（100K）	如何管理我的 Microsoft 账户?
DDL（100K）	如何管理我的 Microsoft 账户?

个示例）。DDL 模型的学习效率也更高：在若干示例中，它使用 50K 数据便取得了与 TRANS 模型使用 100K 数据相当的结果。这些观察进一步印证了我们提出的 DDL 模型的有效性和效率。

6.5.5 纳入未标注数据的定性评估

在表10中，我们展示了 Quora 测试数据集中的定性示例，以及我们提出的半监督学习模型 DDL +VSAR（见第 5 节）在相同数量的带标签数据（20K，与 6.5.4 节中的数据实例相同）和不同规模的未标注数据（20K、50K 和 100K）条件下生成的相应模型输出。作为对比，我们还提供了使用相同的 20K 个示例训练的 DDL 模型（第 4 节）的模型输出。与表9类似，我们使用相同的随机种子 1000 生成这些示例。在表10中，我们首先列出了来源和黄金参考答案，

表10：**在使用不同数量的未标注数据（括号内所示）和相同数量的已标注数据（20K）时，半监督模型（DDL+VSAR）的部分释义生成结果，以 Quora数据集为例。**

来源:	有可能到达地球核心吗?
参考:	如果我真的想这么做，可以一路挖到地球核心吗?
DDL(20K)	可以去地球吗?
DDL(20K) + VSAR(20K)	我该怎么环游地球?
DDL(20K) + VSAR(50K)	我该怎么环游地球?
DDL(20K) + VSAR(100K)	有可能到达地球核心吗?
来源:	新闻学有哪些最好的课程?
参考:	新闻学有哪些最好的课程?
DDL(20K)	最好的数字营销课程是什么?
DDL(20K) + VSAR(20K)	最好的数字营销机构是什么?
DDL(20K) + VSAR(50K)	最好的数字营销课程是什么?
DDL(20K) + VSAR(100K)	新闻学最好的课程有哪些?
来源:	我该如何不再想念某个人?
参考:	我该如何不再想念某个人?
DDL(20K)	我怎样才能不再想着自己?
DDL(20K) + VSAR(20K)	我怎样才能不再思考思考这件事?
DDL(20K) + VSAR(50K)	我怎样才能不再想某件事?
DDL(20K) + VSAR(100K)	如何停止想念一个人?
来源:	是什么激励着所有人?
参考:	是什么激励着人们?
DDL(20K)	为什么人们总是把所有人都留下?
DDL(20K) + VSAR(20K)	人们为什么要纹身?
DDL(20K) + VSAR(50K)	是什么激励人们如此坚强?
DDL(20K) + VSAR(100K)	是什么激励着人们?
来源:	怎样才能让 100 万用户注册我的应用程序?
参考:	怎样才能让我的社交应用程序拥有 100 万用户?
DDL(20K)	如何在我的应用程序上下载一款游戏?
DDL(20K) + VSAR(20K)	如何通过一个应用程序找到一个人?
DDL(20K) + VSAR(50K)	如何通过 YouTube 成为亿万富翁?
DDL(20K) + VSAR(100K)	如何让我的应用程序重新获得 100 万用户?
来源:	吃素有益健康，还是吃荤有益健康?
参考:	哪种食物更有益健康：素食还是荤食?
DDL(20K)	吸烟算素食吗，素食?
DDL(20K) + VSAR(20K)	素食有益健康吗?
DDL(20K) + VSAR(50K)	素食比素食更好吗?
DDL(20K) + VSAR(100K)	素食对健康是好还是坏?
来源:	我该如何管理我的 Microsoft 账户?
参考:	我该如何管理我的 Microsoft 账户?
DDL(20K)	我该如何管理我的 Microsoft Office?
DDL(20K) + VSAR(20K)	我该如何管理我的微软大小?
DDL(20K) + VSAR(50K)	我该如何管理我的 Google 账户?
DDL(20K) + VSAR(100K)	如何管理我的 Microsoft 账户?

随后列出基于 20K 个带标签文本对训练的模型输出：DDL 模型使用的 20K 个带标签文本对；DDL+VSAR 模型使用的相同 20K 个带标签文本对（类似于表1、2、3和4中的 DDL+VSAR$_1$ 设置）；DDL+VSAR模型使用的相同 20K 个带标签文本对及 50K 个未标注数据（额外增加 30K）；以及 DDL+VSAR 模型使用的相同 20K 个带标签文本对及 100K 个未标注数据（额外增加 80K）（类似于表1、2、3和4中的 DDL+VSAR$_2$ 设置）。

我们可以清楚地观察到，在 DDL+VSAR 模型的训练过程中使用更多未标注数据时，生成的释义与参考文本更加匹配。在 VSAR 模型中加入未标注数据后生成的潜在序列，提升了释义生成性能。总体而言，我们观察到一种逐步

提升：纳入更多未标注样例后，模型能够更好地捕捉信息的核心。与 VSAR 模型在完全无监督场景中未能学习的初步实验（第6.5.1节）相比，通过表10中的定性结果以及表1、2、3和4中的定量结果，我们表明可以结合无监督模型 VSAR 和监督模型 DDL 来进行半监督学习。

表11： **Quora数据集误差分析中选取的释义生成结果**。

来源:	什么是网络营销？我想学习网络营销，该怎么做?
参考:	开始学习网络营销的最佳方式是什么?
DDL(20K) + VSAR(20K)	互联网上最有价值的网站是什么?
DDL(20K) + VSAR(50K)	学习互联网的最佳方式是什么?
DDL(20K) + VSAR(100K)	印度线上和线下有哪些最好的数字营销课程，以及为什么?
来源:	你觉得人生的意义是什么?
参考:	在你看来，人生的意义是什么?
DDL(20K) + VSAR(20K)	生命的意义是什么?
DDL(20K) + VSAR(50K)	你生命的意义是什么?
DDL(20K) + VSAR(100K)	你认为生命的意义是什么?
来源:	你为什么相信鬼魂和超自然现象?
参考:	你相信鬼魂吗？为什么?
DDL(20K) + VSAR(20K)	你相信鬼魂吗?
DDL(20K) + VSAR(50K)	你相信鬼魂吗?
DDL(20K) + VSAR(100K)	你觉得自己曾遭遇过超自然事件吗?
来源:	为什么那么多人在 Quora 上提问，而这些问题只需在任意搜索引擎中输入，就能轻松而详尽地得到解答?只需在任意搜索引擎中输入问题，就能轻松而详尽地得到解答?
参考:	为什么这么多人在 Quora 上问些矫情的问题，而这些问题明明用谷歌就能轻松搜到??
DDL(20K) + VSAR(20K)	为什么人们要在 Quora 上提问，而不是直接用谷歌搜索?
DDL(20K) + VSAR(50K)	为什么人们要在 Quora 上提问，而不是直接用谷歌搜索?
DDL(20K) + VSAR(100K)	为什么人们会在 Quora 上提出只需快速搜索网页就能回答的问题??
来源:	在黄金海岸哪里能找到美味的纸杯蛋糕?
参考:	在黄金海岸哪里能找到不同口味的纸杯蛋糕?
DDL(20K) + VSAR(20K)	在黄金海岸哪里可以买到品质最好的精品纸杯蛋糕?
DDL(20K) + VSAR(50K)	在黄金海岸哪里可以买到品质最好的纸杯蛋糕?
DDL(20K) + VSAR(100K)	在黄金海岸哪里可以找到口味、设计和装饰最好的纸杯蛋糕?
来源:	素食有益健康，还是非素食有益健康?
参考:	哪种食物对我们的健康有益：素食还是非素食?
DDL(20K)	吸烟算是素食者吗?
DDL(20K) + VSAR(20K)	素食是否被认为有益健康?
DDL(20K) + VSAR(50K)	素食是否比素食更好?
DDL(20K) + VSAR(100K)	素食对健康是好还是坏?

6.5.6 误差分析

本节对我们提出的模型进行误差分析。总体而言，正如表1、2、3和4所示，我们的模型在半监督设置下表现良好；然而，在某些情况下，即使使用更多未标注对，模型生成示例的质量也会变差，或没有提升。我们发现，对于 DDL+VSAR 模型而言，在使用更多未标注数据后性能没有提升的情况中，测试示例的黄金参考答案相对容易改写，且句子较短。不过，我们还观察到，如表11所示，当使用更多未标注数据点时，DDL+VSAR 模型有时会生成过于复杂的释义。我们注意到，使用未标注数据会提升模型处理更复杂释义的能力，但同时也会促使模型给出复杂的答案。这一发现颇有意思，表明未标注数据的质量也在半监督学习过程中发挥着重要作用。

6.5.7 算法运行时间

本节进一步讨论我们所提模型的算法运行时间和 GPU 内存需求。对于 DDL 模型而言，它相当于训练一个标准 Transformer 模型，并额外增加了成本项（包括两个方向：从来源到目标，以及从目标到来源）。因此，训练 DDL 模型无需额外运行时间，也不需要额外的 GPU 内存。对于单独的 VSAR 模型，由于来源重建和潜在推断共享参数（如图 1 所示），其训练成本与 DDL 模型相近，此外还需要一些额外的 GPU 内存，用于保存潜变量样本的梯度以执行反向传播。由于训练过程中只使用一个样本，额外的 GPU 内存开销并不显著，而且我们没有发现训练 VSAR 模型需要额外的运行时间。对于半监督学习模型 DDL+VSAR，我们需要先用带标签数据预训练 DDL 模型，再以 DDL 模型的权重初始化模型，并使用带标签和未标注数据共同微调 DDL+VSAR 模型。与预训练和微调的过程相比，总运行时间增加了一倍。尽管我们发现，预训练已使模型处于接近最优的参数空间，因此，为确保算法能够扩展到大规模数据，我们或许可以在半监督学习中减少训练轮数。

7 局限性与未来工作

这里简要讨论本研究发现的两个主要局限性，以及各自相关的未来工作。第一个局限性涉及本研究使用的定量评估指标，例如 BLEU 和 ROUGE 分数。这些指标基于生成内容与参考文本之间 n 元语境的重叠。虽然它们通常用于与已发表的结果进行比较（如表 5、6、7 和 8 所示），但无法直接评估生成文本的质量。已有研究提出了较新的评估指标，例如 BERTScore [81], ，但对于哪些定量指标最佳，目前仍未达成普遍共识。未来的一个研究方向是深入探索这一领域，并提出更好的评估指标，以衡量生成释义的质量，最终为自然语言生成研究带来切实影响。

第二个局限性是，缺少基于用户的研究来评估依据我们提出的方法实现的系统，而这类研究通常被认为是衡量模型性能更全面的方式。尽管本文在第 6.5 节中观察到了令人鼓舞的结果，但本研究中的评估依赖于公开基准数据集，而这些数据集可能偏向于模型训练所用数据的某个子集。由于资源有限，本研究未开展基于用户的评估，因此这也是一个值得未来研究的方向。

作为未来工作，本文探索了一种更具结构化的方法，将未观测到的释义建模为离散潜变量。一个值得探索的未来方向是将这一技术应用于预训练语言模型（PLMs），并进一步加以拓展。预训练语言模型的最新进展表明，即使只使用相对少量的数据，它们也能取得出色的性能；在文档检索任务中，甚至在少样本或零样本场景下也能表现良好。然而，自然语言生成（NLG）任务仍面临挑战，本文提出的方法可作为一块垫脚石，助力在自然语言生成任务中同时利用带标签数据和未标注数据。

8 结论

本文提出了一种用于释义生成的半监督深度生成模型。无监督模型（VSAR）基于变分自动编码框架，为处理缺失标签提供了一种有效方法。监督模型（DDL）采用双向学习，并将监督信息注入无监督模型。通过新颖的知识强化学习（KRL）方案，我们以实证方式证明，在使用未标注数据和一部分配对数据的情况下，半监督学习有利于提升组合模型的性能。评估结果表明，在半监督设置下，我们的组合模型优于一个非常强大的基线模型。我们还发现，即使使用完整数据集，我们的组合模型在两个释义生成基准数据集上的性能也能与最先进模型相媲美。此外，我们能够将语言建模为用于释义生成任务的离散潜变量序列。重要的是，最终的生成模型能够在序列到序列任务中有效利用监督数据和无监督数据。

参考描述

[1] Li Dong、Jonathan Mallinson、Siva Reddy 和 Mirella Lapata。面向问答的释义学习。载于*2017*年自然语言处理实证方法会议论文集，第875–886页，丹麦哥本哈根，2017年。计算语言学协会。

[2] Minsuk Lee、James Cimino、Hai Ran Zhu、Carl Sable、Vijay Shanker、John Ely 和 Hong Yu。超越信息检索——医学问答。载于*AMIA* 年度研讨会论文集，第2006卷，第469页。美国医学信息学会，2006年。[3] Xuchen Yao 和 Benjamin Van Durme。结构化数据上的信息抽取：使用 Freebase 进行问答。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956–966页，美国马里兰州巴尔的摩，2014年。计算语言学协会。[4] Yuanjie Liu、Shasha Li、Yunbo Cao、Chin-Yew Lin、Dingyi Han 和 Yong Yu。理解并总结社区问答服务中的答案。载于第*22*届国际计算语言学会议论文集（*Coling 2008*），第497–504页，英国曼彻斯特，2008年。Coling 2008 组织委员会。[5] Tim Miller。人工智能中的解释：来自社会科学的见解。《人工智能》，267：1–38，2019。[6] Tom Hosking、Hao Tang 和 Mirella Lapata。用于释义生成的层次化草图归纳。*arXiv* 预印本 *arXiv:2203.03463*，2022年。[7] David Kauchak 和 Regina Barzilay。用于自动评估的释义。载于*NAACL* 人类语言技术会议主会论文集，第455–462页，美国纽约市，2006年。计算语言学协会。[8] Shashi Narayan、Siva Reddy 和 Shay B. Cohen。基于潜变量概率上下文无关文法的释义生成，用于语义解析。载于第*9*届国际自然语言生成会议论文集，第153–162页，英国爱丁堡，2016年。计算语言学协会。[9] George A. Miller。WordNet：英语词汇数据库。载于《语音与自然语言：*1992*年*2*月*23*日至*26*日在纽约州哈里曼举行的研讨会论文集》，1992年。[10] Ashutosh Kumar、Kabir Ahuja、Raghuram Vadapalli 和 Partha Talukdar。语法引导的受控释义生成。计算语言学协会会刊，8：329–345，2020年。[11] Jianing Zhou 和 Suma Bhat。释义生成：最先进水平综述。载于*2021*年自然语言处理实证方法会议论文集，第5075–5086页，2021年。[12] Yuxian Meng、Xiang Ao、Qing He、Xiaofei Sun、Qinghong Han、Fei Wu、Jiwei Li 等。Conrpg：利用上下文作为正则化项进行释义生成。*arXiv* 预印本 *arXiv:2109.00363*，2021年。[13] Yixuan Su、David Vandyke、Simon Baker、Yan Wang 和 Nigel Collier。保留主要部分，改写次要部分：一种释义生成的两阶段方法。载于计算语言学协会会议论文集：*ACL-IJCNLP 2021*，第560–569页，线上，2021年。计算语言学协会。[14] Ilya Sutskever、Oriol Vinyals 和 Quoc V. Le。神经网络的序列到序列学习。载于 Zoubin Ghahramani、Max Welling、Corinna Cortes、Neil D. Lawrence 和 Kilian Q. Weinberger 编，神经信息处理系统进展*27*：*2014*年神经信息处理系统年度会议，*2014*年*12*月*8*日至*13*日，蒙特利尔，魁北克，加拿大，第3104–3112页，2014年。[15] Xiaojin Jerry Zhu。半监督学习文献综述。2005年。[16] Jesper E Van Engelen 和 Holger H Hoos。半监督学习综述。机器学习，109(2)：373– 440，2020年。[17] Diederik P. Kingma 和 Max Welling。自编码变分贝叶斯。载于 Yoshua Bengio 和 Yann LeCun 编，第*2*届国际学习表征会议，*ICLR 2014*，班夫，艾伯塔省，加拿大，*2014*年*4*月*14*日至*16*日，会议论文集，2014年。[18] Danilo Jimenez Rezende、Shakir Mohamed 和 Daan Wierstra。深度生成模型中的随机反向传播与近似推断。载于第*31*届国际机器学习会议论文集，*ICML 2014*，中国北京，*2014*年*6*月*21*日至*26*日，第32卷*JMLR*研讨会与会议论文集，第1278–1286页。JMLR.org，2014年。[19] Andriy Mnih 和 Karol Gregor。信念网络中的神经变分推断与学习。载于第*31*届国际机器学习会议论文集，*ICML 2014*，中国北京，*2014*年*6*月*21*日至*26*日，第32卷*JMLR*研讨会与会议论文集，第1791–1799页。JMLR.org，2014年。[20] Geoffrey E Hinton 和 Ruslan R Salakhutdinov。利用神经网络降低数据的维度。《科学》，313(5786)：504–507，2006年。[21] Yao Fu、Yansong Feng 和 John P. Cunningham。基于潜在词袋的释义生成。载于 Hanna M. Wallach、Hugo Larochelle、Alina Beygelzimer、Florence d’Alché-Buc、Emily B. Fox 和 Roman Garnett 编，

神经信息处理系统进展*32*：神经信息处理系统年度会议*2019，NeurIPS 2019，2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，第13623–13634页，2019年。[22] Tom Hosking 和 Mirella Lapata。面向意图保持型释义的意义与形式分解。载于计算语言学协会第*59*届年会暨自然语言处理国际联合会议第*11*届会议论文集（第*1*卷：长篇论文），第1405–1418页，线上，2021年。计算语言学协会。[23] Di He、Yingce Xia、Tao Qin、Liwei Wang、Nenghai Yu、Tie-Yan Liu 和 Wei-Ying Ma。用于机器翻译的双向学习。载于 Daniel D. Lee、Masashi Sugiyama、Ulrike von Luxburg、Isabelle Guyon 和 Roman Garnett（编），神经信息处理系统进展*29*：神经信息处理系统年度会议*2016，2016*年*12*月*5*日至*10*日，西班牙巴塞罗那，第820–828页，2016年。[24] Shang-Yu Su、Chao-Wei Huang 和 Yun-Nung Chen。用于自然语言理解与生成的双重监督学习。载于计算语言学协会第*57*届年会论文集，第5472–5477页，意大利佛罗伦萨，2019年。计算语言学协会。[25] Shang-Yu Su、Yung-Sung Chuang 和 Yun-Nung Chen。通过双重推理改进语言理解与生成。载于计算语言学协会会议论文集：*EMNLP 2020*，第4930–4936页，线上，2020年。计算语言学协会。[26] Shang-Yu Su、Chao-Wei Huang 和 Yun-Nung Chen。通过联合双重学习实现无监督语言理解与生成。载于计算语言学协会第*58*届年会论文集，第671–680页，线上，2020年。计算语言学协会。[27] Ellie Pavlick、Pushpendre Rastogi、Juri Ganitkevitch、Benjamin Van Durme 和 Chris Callison-Burch。PPDB 2.0：更优的释义排序、细粒度蕴含关系、词嵌入和风格分类。载于计算语言学协会第*53*届年会暨自然语言处理国际联合会议第*7*届会议论文集（第*2*卷：短篇论文），第425–430页，中国北京，2015年。计算语言学协会。[28] Ankush Gupta、Arvind Agarwal、Prawaan Singh 和 Piyush Rai。用于释义生成的深度生成框架。载于 Sheila A. McIlraith 和 Kilian Q. Weinberger（编），第三十二届*AAAI*人工智能会议论文集（*AAAI-18*）、第*30*届人工智能创新应用会议（*IAAI-18*）暨第*8*届*AAAI*人工智能教育进展研讨会（*EAAI-18*），美国路易斯安那州新奥尔良，*2018*年*2*月*2*日至*7*日，第5149–5156页。AAAI出版社，2018年。[29] Yi Chen、Haiyun Jiang、Lemao Liu、Rui Wang、Shuming Shi 和 Ruifeng Xu。MCPG：一种灵活的多层可控无监督释义生成框架。载于计算语言学协会会议论文集：*EMNLP 2022*，第5948–5958页，2022年。[30] Jiayuan Xie、Jiali Chen、Yi Cai、Qingbao Huang 和 Qing Li。保留关键信息的视觉释义生成。*ACM*多媒体计算、通信与应用汇刊，2023。[31] André F. T. Martins、Tsvetomila Mihaylova、Nikita Nangia 和 Vlad Niculae。自然语言处理的潜在结构模型。载于计算语言学协会第*57*届年会论文集：教程摘要，第1–5页，意大利佛罗伦萨，2019年7月。计算语言学协会。[32] Yishu Miao 和 Phil Blunsom。作为潜变量的语言：用于句子压缩的离散生成模型。载于*2016*年自然语言处理实证方法会议论文集，第319–328页，得克萨斯州奥斯汀，2016年。计算语言学协会。[33] Yishu Miao、Lei Yu 和 Phil Blunsom。用于文本处理的神经变分推断。载于 Maria-Florina Balcan 和 Kilian Q. Weinberger（编），第*33*届国际机器学习会议论文集，*ICML 2016*，美国纽约市，*2016*年*6*月*19*日至*24*日，*JMLR*研讨会与会议论文集第48卷，第1727–1736页。JMLR.org，2016年。[34] Yoon Kim、Sam Wiseman 和 Alexander M Rush。自然语言深度潜变量模型教程。*arXiv*预印本 *arXiv:1812.06834*，2018年。[35] Samuel R. Bowman、Luke Vilnis、Oriol Vinyals、Andrew Dai、Rafal Jozefowicz 和 Samy Bengio。从连续空间生成句子。载于第*20*届*SIGNLL*计算自然语言学习会议论文集，第10–21页，德国柏林，2016年。计算语言学协会。[36] Adji B. Dieng、Yoon Kim、Alexander M. Rush 和 David M. Blei。通过生成式跳跃模型避免潜变量坍塌。载于 Kamalika Chaudhuri 和 Masashi Sugiyama（编），第*22*届人工智能与统计学国际会议，*AISTATS 2019*，日本冲绳那霸，*2019*年*4*月*16*日至*18*日，第89卷论文集

gs of 机器学习　　*g* 研究　，pag第 2397–2405 页。PMLR　，2019年。

[37] Junxian He、Daniel Spokoyny、Graham Neubig 和 Taylor Berg-Kirkpatrick。变分自编码器中的滞后推理网络与后验坍塌。载于 第七届国际学习表征会议，*ICLR 2019*，美国路易斯安那州新奥尔良，*2019*年*5*月*6*日至*9*日。OpenReview.net，2019年。[38] Irina Higgins、Loïc Matthey、Arka Pal、Christopher Burgess、Xavier Glorot、Matthew Botvinick、Shakir Mohamed 和 Alexander Lerchner。β-VAE：通过受约束变分框架学习基本视觉概念。载于 第五届国际学习表征会议，*ICLR 2017*，法国土伦，*2017*年*4*月*24*日至*26*日，会议论文集。OpenReview.net，2017年。[39] Ali Razavi、Aäron van den Oord、Ben Poole 和 Oriol Vinyals。使用δ-VAE防止后验坍塌。载于第七届国际学习表征会议，*ICLR 2019*，美国路易斯安那州新奥尔良，*2019*年*5*月*6*日至*9*日。OpenReview.net，2019年。[40] Yixin Wang、David Blei 和 John P Cunningham。后验坍塌与潜变量不可辨识性。神经信息处理系统进展，34，2021年。[41] Tsung-Hsien Wen、Yishu Miao、Phil Blunsom 和 Steve J. Young。潜在意图对话模型。载于 Doina Precup 和 Yee Whye Teh（编），第*34*届国际机器学习会议，*ICML 2017*，澳大利亚新南威尔士州悉尼，*2017*年*8*月*6*日至*11*日，机器学习研究论文集第70卷，第3732–3741页。PMLR，2017年。[42] Volodymyr Mnih、Nicolas Heess、Alex Graves 和 Koray Kavukcuoglu。视觉注意力的循环模型。载于 Zoubin Ghahramani、Max Welling、Corinna Cortes、Neil D. Lawrence 和 Kilian Q. Weinberger（编），神经信息处理系统进展*27*：*2014*年神经信息处理系统年度会议，*2014*年*12*月*8*日至*13*日，加拿大魁北克省蒙特利尔，第2204–2212页，2014年。[43] Eric Jang、Shixiang Gu 和 Ben Poole。使用Gumbel-Softmax进行分类重参数化。载于 第五届国际学习表征会议，*ICLR 2017*，法国土伦，*2017*年*4*月*24*日至*26*日，会议论文集。OpenReview.net，2017年。[44] Chris J. Maddison、Andriy Mnih 和 Yee Whye Teh。Concrete分布：离散随机变量的连续松弛。载于 第五届国际学习表征会议，*ICLR 2017*，法国土伦，*2017*年*4*月*24*日至*26*日，会议论文集。OpenReview.net，2017年。[45] Jihun Choi、Kang Min Yoo 和 Sang-goo Lee。学习组合任务特定树结构。载于 Sheila A. McIlraith 和 Kilian Q. Weinberger（编），第三十二届*AAAI*人工智能会议（*AAAI-18*）、第*30*届人工智能创新应用会议（*IAAI-18*）以及第八届*AAAI*人工智能教育进展研讨会（*EAAI-18*）论文集，美国路易斯安那州新奥尔良，*2018*年*2*月*2*日至*7*日，第5094–5101页。AAAI出版社，2018年。[46] Junxian He、Xinyi Wang、Graham Neubig 和 Taylor Berg-Kirkpatrick。无监督文本风格迁移的概率公式。载于 第八届国际学习表征会议，*ICLR 2020*，埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。OpenReview.net，2020年。[47] Giangiacomo Mercatali 和 André Freitas。使用离散变分自编码器解耦自然语言中的生成因素。*arXiv*预印本 *arXiv:2109.07169*，2021年。[48] Sang Michael Xie 和 Stefano Ermon。通过连续松弛实现可重参数化的子集采样。载于 Sarit Kraus（编），第二十八届国际人工智能联合会议论文集，*IJCAI 2019*，中国澳门，*2019*年*8*月*10*日至*16*日，第3919–3925页。ijcai.org，2019年。[49] Vlad Niculae、Caio F Corro、Nikita Nangia、Tsvetomila Mihaylova 和 André FT Martins。神经网络中的离散潜在结构。*arXiv*预印本 *arXiv:2301.07473*，2023年。[50] Rongjie Huang、Jiawei Huang、Dongchao Yang、Yi Ren、Luping Liu、Mingze Li、Zhenhui Ye、Jinglin Liu、Xiang Yin 和 Zhou Zhao。Make-an-audio：使用提示增强扩散模型进行文本到音频生成。*arXiv*预印本 *arXiv:2301.12661*，2023年。[51] Flavio Schneider、Zhijing Jin 和 Bernhard Schölkopf。Mo\^ usai：使用长上下文潜在扩散进行文本到音乐生成。*arXiv*预印本 *arXiv:2301.11757*，2023年。[52] Minh-Thang Luong、Quoc V. Le、Ilya Sutskever、Oriol Vinyals 和 Lukasz Kaiser。多任务序列到序列学习。载于 Yoshua Bengio 和 Yann LeCun（编），第四届国际学习表征会议，*ICLR 2016*，波多黎各圣胡安，*2016*年*5*月*2*日至*4*日，会议论文集，2016年。[53] Han Guo、Ramakanth Pasunuru 和 Mohit Bansal。用于句子简化的动态多层级多任务学习。载于 第二十七届国际计算语言学会议论文集，第462–476页，美国新墨西哥州圣菲，2018年。计算语言学协会。

[54] Han Guo、Ramakanth Pasunuru 和 Mohit Bansal。结合蕴含与问题生成的软层特定多任务摘要。载于计算语言学协会第*56*届年会论文集（第*1*卷：长篇论文），第687–697页，澳大利亚墨尔本，2018年。计算语言学协会。[55] Yiren Wang、ChengXiang Zhai 和 Hany Hassan。多语言神经机器翻译的多任务学习。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第1022–1034页，线上，2020年。计算语言学协会。[56] Jingfei Du、Edouard Grave、Beliz Gunel、Vishrav Chaudhary、Onur Celebi、Michael Auli、Veselin Stoyanov 和 Alexis Conneau。自训练改进自然语言理解的预训练。载于*2021*年计算语言学协会北美分会会议：人类语言技术论文集，第5408–5418页，线上，2021年。计算语言学协会。[57] Ernie Chang、Vera Demberg 和 Alex Marin。通过自动标注的质量加权弱监督，联合提升语言理解与生成。载于欧洲计算语言学协会第*16*届会议论文集：主会场，第818–829页，线上，2021年。计算语言学协会。[58] Ronald J Williams 和 David Zipser。用于持续运行的全循环神经网络的学习算法。神经计算，1(2):270–280，1989年。[59] Steven Burrows、Martin Potthast 和 Benno Stein。通过众包和机器学习获取释义。*ACM*智能系统与技术汇刊（*TIST*），4(3):1–21，2013年。[60] Ziqiang Cao、Chuwei Luo、Wenjie Li 和 Sujian Li。用于释义的联合复制与受限生成。载于 Satinder P. Singh 和 Shaul Markovitch（编），第三十一届*AAAI*人工智能会议论文集，*2017*年*2*月*4*日至*9*日，美国加利福尼亚州旧金山，第3152–3158页。AAAI出版社，2017年。[61] Jianbo Chen、Le Song、Martin J. Wainwright 和 Michael I. Jordan。学习解释：模型解释的信息论视角。载于 Jennifer G. Dy 和 Andreas Krause（编），第*35*届国际机器学习会议（*ICML 2018*）论文集，*2018*年*7*月*10*日至*15*日，瑞典斯德哥尔摩斯德哥尔摩国际展览中心，机器学习研究第80卷机器学习研究论文集，第882–891页。PMLR，2018年。[62] Muhammed Fatih Balin、Abubakar Abid 和 James Y. Zou。Concrete自编码器：可微特征选择与重建。载于 Kamalika Chaudhuri 和 Ruslan Salakhutdinov（编），第*36*届国际机器学习会议（*ICML 2019*）论文集，*2019*年*6*月*9*日至*15*日，美国加利福尼亚州长滩，机器学习研究第97卷机器学习研究论文集，第444–453页。PMLR，2019年。[63] Yoshua Bengio、Nicholas Léonard 和 Aaron Courville。针对条件计算，估计或传播随机神经元中的梯度。*arXiv*预印本 *arXiv:1308.3432*，2013年。[64] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。注意力就是你所需的一切。载于 Isabelle Guyon、Ulrike von Luxburg、Samy Bengio、Hanna M. Wallach、Rob Fergus、S. V. N. Vishwanathan 和 Roman Garnett（编），神经信息处理系统进展第*30*卷：*2017*年*12*月*4*日至*9*日于美国加利福尼亚州长滩举行的神经信息处理系统年会，第5998–6008页，2017年。[65] Zichao Yang、Zhiting Hu、Chris Dyer、Eric P. Xing 和 Taylor Berg-Kirkpatrick。利用语言模型作为判别器进行无监督文本风格迁移。载于 Samy Bengio、Hanna M. Wallach、Hugo Larochelle、Kristen Grauman、Nicolò Cesa-Bianchi 和 Roman Garnett（编），神经信息处理系统进展第*31*卷：*2018*年神经信息处理系统年会（*NeurIPS 2018*），*2018*年*12*月*3*日至*8*日，加拿大蒙特利尔，第7298–7309页，2018年。[66] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。载于欧洲计算机视觉会议，第740–755页。施普林格，2014年。[67] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向Transformer预训练。载于*2019*年计算语言学协会北美分会会议：人类语言技术论文集，第*1*卷（长篇与短篇论文），第4171–4186页，美国明尼苏达州明尼阿波利斯，2019年。计算语言学协会。[68] Zichao Li、Xin Jiang、Lifeng Shang 和 Qun Liu。可分解神经释义生成。载于计算语言学协会第*57*届年会论文集，第3403–3414页，意大利佛罗伦萨，2019年。计算语言学协会。[69] J Edward Hu、Rachel Rudinger、Matt Post 和 Benjamin Van Durme。ParaBank：通过词汇约束神经机器翻译进行单语双语文本生成与句子释义。载于*AAAI*人工智能会议论文集，第33卷，第6521–6528页，2019年。

[70] John Wieting 和 Kevin Gimpel。ParaNMT-50M：借助数百万条机器翻译，突破释义句嵌入的极限。*arXiv*预印本 *arXiv:1711.05732*，2017年。[71] Jeffrey Pennington、Richard Socher 和 Christopher Manning。GloVe：用于词表示的全局向量。载于*2014*年自然语言处理实证方法会议（*EMNLP*）论文集，第1532–1543页，卡塔尔多哈，2014年。计算语言学协会。[72] Tomás Mikolov、Ilya Sutskever、Kai Chen、Gregory S. Corrado 和 Jeffrey Dean。词与短语的分布式表示及其组合性。载于 Christopher J. C. Burges、Léon Bottou、Zoubin Ghahramani 和 Kilian Q. Weinberger（编），神经信息处理系统进展第*26*卷：第*27*届神经信息处理系统年会论文集，会议于*2013*年*12*月*5*日至*8*日在美国内华达州太浩湖举行，第3111–3119页，2013年。[73] Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于 Yoshua Bengio 和 Yann LeCun（编），第三届国际学习表征会议（*ICLR 2015*）论文集，*2015*年*5*月*7*日至*9*日，美国加利福尼亚州圣迭戈，会议论文集，2015年。[74] Aaditya Prakash、Sadid A. Hasan、Kathy Lee、Vivek Datla、Ashequl Qadir、Joey Liu 和 Oladimeji Farri。利用堆叠残差 LSTM 网络进行神经释义生成。载于*COLING 2016*论文集：第*26*届国际计算语言学会议技术论文，第2923–2934页，日本大阪，2016年。COLING 2016组织委员会。[75] Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第*40*届年会论文集，第311–318页，美国宾夕法尼亚州费城，2002年。计算语言学协会。[76] Chin-Yew Lin。ROUGE：用于摘要自动评估的软件包。载于文本摘要拓展，第74–81页，西班牙巴塞罗那，2004年。计算语言学协会。[77] Hong Sun 和 Ming Zhou。联合学习用于释义生成的双向统计机器翻译系统。载于计算语言学协会第*50*届年会论文集（第*2*卷：短篇论文），第38–42页，韩国济州岛，2012年。计算语言学协会。[78] Nitin Madnani、Joel Tetreault 和 Martin Chodorow。重新审视用于释义识别的机器翻译指标。载于*2012*年计算语言学协会北美分会会议：人类语言技术论文集，第182–190页，加拿大蒙特利尔，2012年。计算语言学协会。[79] Sander Wubben、Antal van den Bosch 和 Emiel Krahmer。将释义生成视为单语翻译：数据与评估。载于第六届国际自然语言生成会议论文集，2010年。[80] Wenqing Chen、Jidong Tian、Liqiang Xiao、Hao He 和 Yaohui Jin。用于释义生成的语义一致、句法可变编码器-解码器框架。载于第*28*届国际计算语言学会议论文集，第1186–1198页，西班牙巴塞罗那（线上），2020年。国际计算语言学委员会。[81] Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q Weinberger 和 Yoav Artzi。BERTScore：使用BERT评估文本生成。载于国际学习表征会议。