软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2022,33(12):4668−4687 [doi: 10.13328/j.cnki.jos.006544]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

文本风格迁移研究综述*

陈可佳[1,2], 费子阳[1], 陈景强[1,2], 杨子农[1]

[1](南京邮电大学 计算机学院, 江苏 南京 210023)

[2](江苏省大数据安全与智能处理重点实验室 (南京邮电大学), 江苏 南京 210023)

通信作者: 陈可佳, E-mail: chenkj@njupt.edu.cn

摘　要: 文本风格迁移是近年来自然语言处理领域的热点问题之一, 旨在保留文本内容的基础上通过编辑或生成的方式更改文本的特定风格或属性 (如情感、时态和性别等). 旨在梳理已有的技术, 以推进该方向的研究. 首先, 给出文本风格迁移问题的定义及其面临的挑战; 然后, 对已有方法进行分类综述, 重点介绍基于无监督学习的文本风格迁移方法并将其进一步分为隐式和显式两类方法, 对各类方法在实现机制、优势、局限性和性能等方面进行分析和比较; 同时, 还通过实验比较了几种代表性方法在风格迁移准确率、文本内容保留和困惑度等自动化评价指标上的性能; 最后, 对文本风格迁移研究进行总结和展望.

关键词: 文本风格迁移; 自然语言处理; 对抗学习; 强化学习; 机器翻译

中图法分类号: TP18

中文引用格式: 陈可佳, 费子阳, 陈景强, 杨子农. 文本风格迁移研究综述. 软件学报, 2022, 33(12): 4668–4687. http://www.jos.org.cn/1000-9825/6544.htm

英文引用格式: Chen KJ, Fei ZY, Chen JQ, Yang ZN. Survey on Text Style Transfer Research. Ruan Jian Xue Bao/Journal of Software, 2022, 33(12): 4668–4687 (in Chinese). http://www.jos.org.cn/1000-9825/6544.htm

Survey on Text Style Transfer Research

CHEN Ke-Jia[1,2], FEI Zi-Yang[1], CHEN Jing-Qiang[1,2], YANG Zi-Nong[1]

[1](School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

[2](Jiangsu Key Laboratory of Big Data Security & Intelligent Processing (Nanjing University of Posts and Telecommunications), Nanjing 210023, China)

Abstract: Text style transfer is one of the hot issues in the field of natural language processing in recent years. It aims to transfer the specific style or attributes of the text (such as emotion, tense, gender, etc.) through editing or generating while retaining the text content. The purpose of this study is to sort out the existing methods in order to advance this research field. First, the problem of text style transfer is defined and the challenges are given. Then, the existing methods are classified and reviewed, focusing on the TST methods based on unsupervised learning and further dividing them into the implicit methods and the explicit methods. The implementation mechanisms, advantages, limitations, and performance of each method are also analyzed. Subsequently, the performance of several representative methods on automatic evaluation indicators such as transfer accuracy, text content retention, and perplexity are compared through experiments. Finally, the research of text style transfer is concluded and prospected.

Key words: text style transfer (TST); natural language processing; adversarial learning; reinforcement learning; machine translation

“风格迁移”的概念于 2015 年最先由 Gatys 等人[1]在计算机视觉领域提出. 他们使用卷积神经网络分别学习图像的内容特征和风格特征, 通过处理高层抽象特征实现图像风格迁移, 并获得了可观的视觉艺术效果. 随着生成对抗网络模型 (generative adversarial network, GAN)[2]在图像生成中的广泛应用[3,4], 图像风格迁移领域获得了进一步

* 基金项目: 国家自然科学基金 (61772284, 61876091)
收稿时间: 2021-07-04; 修改时间: 2021-09-22; 采用时间: 2021-11-12; jos 在线出版时间: 2021-12-24

的发展. 受此启发, Hu 等人[5]提出了文本风格迁移 (text style transfer, TST) 任务, 其目的是通过编辑句子来改变句子的属性 (即风格)、保留与属性无关的文本内容以及保证句子的流畅度, 其方法的迁移效果如表 1 所示. 其中, 第 1 组为文本的情感属性 (积极和消极) 转换的案例, 第 2 组为文本的时态属性 (过去、现在和将来) 转换的案例. 目前, 文本风格迁移可应用于许多现实场景中, 例如文本润色[6]、诗歌创作[7]、人机对话[8]、特定风格标题生成[9]、社区评论环境改善[10,11]等.

表 1 文本风格迁移示例[5]

negative -> positive (消极 -> 积极)		varying the code of tense (时态变换)
Success cases (成功案例)	Failure cases (失败案例)	
the film is strictly routine! (这部电影太老套了!) the film is full of imagination (这部电影充满了想象力)	the plot is not so original (这个情节不那么新颖) the plot weaves us into <unk> (这个情节将我们带入<unk>)	i thought the movie was too bland and too much (过去时: 我过去认为这部电影太长太平淡了) i guess the movie is too bland and too much (现在时: 我猜这部电影太长太平淡了) i guess the film will have been too bland (虚拟语态: 我猜这部电影将会很平淡)

由于缺乏大量的平行语料 (parallel corpus), 文本风格迁移任务难以直接使用机器翻译中常用的序列到序列 (sequence to sequence, seq2seq)[12]模型, 大部分的研究旨在从无监督学习的角度通过某种学习机制分离文本属性与文本内容的潜在表示, 再融合文本内容和目标属性以实现风格迁移. 例如: 一些方法[5,13−16]采用生成对抗网络[2]学习文本数据的平滑表示, 分离文本的属性与内容. 另一些方法[17,18]采用强化学习策略, 将文本的属性、内容、流畅度等量化为奖励加入模型以实现属性与内容的分离. 此外还有方法[19]采用机器翻译中的回译策略, 弱化句子的属性来更好地保留句子的内容. 与上述隐式地分离文本属性与内容的方法不同, Li 等人[20]观察到句子的属性通常呈现为特定的短语, 显式地分离相关短语并更换为目标属性短语便可实现文本风格迁移, 从而提出了删除、检索、生成 (delete retrieve generate, DRG) 的风格迁移框架.

本文对文本风格迁移领域的研究进行系统性地综述. 第 1 节给出该问题的定义并探讨这一领域目前面临的挑战和尚未解决的问题; 第 2 节对现有的方法、模型进行分类介绍; 第 3 节介绍已有的数据集和评价指标, 以及通过实验对其中的代表性方法进行验证并在自动化指标上进行比较; 最后给出未来可能的研究方向.

1 文本风格迁移的定义与挑战

1.1 定 义

文本风格迁移是指在保留文本内容的基础上通过编辑或生成的方式更改文本的特定属性 (注: 在已发表论文中, 常用“属性”一词来表示风格). 这里的属性可以指情感 (sentiment)、时态 (tense)、性别 (gender)、政治倾向 (political slant) 等.

该任务可形式化为: 给定数据集 $\mathcal{A}=\{(x_1,v_1),(x_2,v_2),\ldots,(x_n,v_n)\}$, 其中 x_i 代表一个句子, $v_i(v_i\in\gamma)$ 代表句子 x_i 包含的某种属性. γ 为属性的取值集合, 一般包含源属性与目标属性两种, 即 $\gamma=\{v^{\text{src}},v^{\text{tgt}}\}$. 文本风格迁移的目标是学习一个函数 $f_\theta(\cdot):(x,v^{\text{src}})\to(y,v^{\text{tgt}})$. 其中, 具有源属性 v^{src} 的句子 x 经过函数映射得到具有目标属性 v^{tgt} 的 y, 但保留了文本的内容. 针对不同的文本风格迁移任务, γ 的定义也不同. 如在情感迁移中, $\gamma=\{positive,negative\}$; 在政治倾向迁移中, $\gamma=\{republican,democratic\}$.

后文表 2 列举了文中出现的每个变量符号及其含义.

1.2 主要挑战

目前, 文本风格迁移任务主要存在以下难点.

(1) 缺少平行语料

针对不同的文本风格迁移任务需要构建不同的数据集, 而每构建一种平行语料数据集都需要大量的语言学知

识和极大的人工开销, 因此目前仅有少量的平行语料数据集, 例如“现代英语-莎士比亚英语”风格的数据集 Shakespeare[21,22]和“正式语言-非正式语言”风格的数据集 GYAFC[6]. 平行语料的缺乏导致了文本风格迁移任务难以直接使用机器翻译领域中的 seq2seq 模型, 而要在无监督学习的框架下进行建模.

表 2 符号表

符号	描述	符号	描述
a	源属性	D	模型判别器
c	目标属性	R	奖励函数
$\mathcal{A}$	带属性标注的句子集合	θ_E	编码器参数
x	含有源属性 a 的源句子	θ_G	生成器参数
y	含有目标属性 c 的目标句子	θ_D	判别器参数
E	模型编码器	α	注意力权重
G	模型生成器	z	文本的潜在表示

(2) 难以分离内容和属性

文本风格迁移不仅要转换文本的属性还需要保留文本的内容, 然而在自然语言中文本的内容和属性往往纠缠在一起, 难以显式地分离. 例如: 句子“i’ve noticed the food service sliding down hill quickly this year.”, 如何让模型在隐空间中更好地分离出属性词 (如这里的“sliding down hill”) 的潜在表示, 是该任务的主要难点之一.

(3) 缺乏公认且统一的评价指标

文本风格迁移研究是在近几年才得到广泛的关注, 不同的工作采用的评价指标也不尽相同 (本文在第 3 节介绍了各指标的具体含义). 评价指标的设计与选取对于模型性能的比较有至关重要的作用, 缺少公认且普适性的评价指标是该任务的另一个问题.

2 方法综述

近年来涌现了越来越多的文本风格迁移方法, 从不同的研究角度应对以上的挑战. 根据训练数据是否为平行语料, TST 方法可初步分为 (如图 1): 基于监督学习的方法、基于无监督学习的方法和基于半监督学习的方法. 其中, 缺少平行语料的无监督学习方法为该领域的主流方法和重点研究内容, 可分为隐式方法和显式方法. 隐式方法旨在学习并区分文本内容和属性的潜在表示, 进一步分为: 解缠策略、强化学习策略、回译策略、伪平行语料策略等; 显式方法旨在从文字本身的角度分离文本的内容和属性, 进一步分为: 基于词频的删除策略、基于注意力机制的删除策略、基于词频和注意力机制相结合的删除策略.

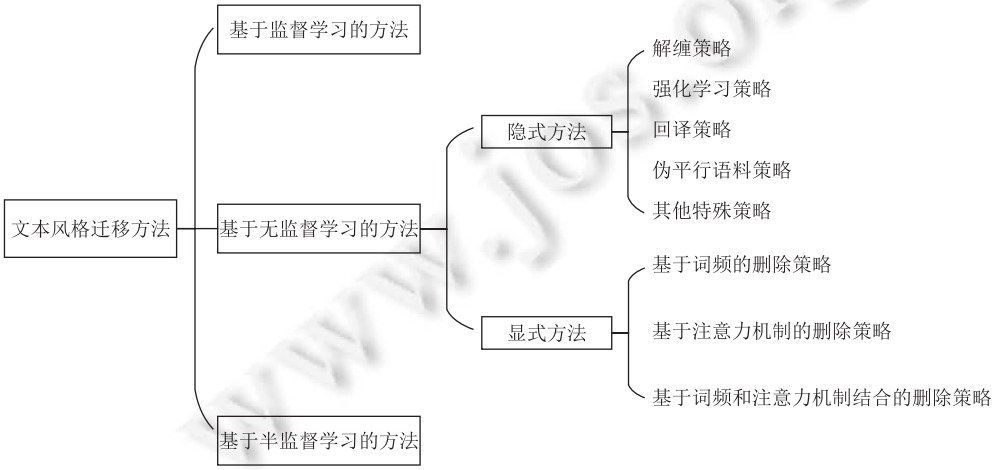


图 1 文本风格迁移方法的分类

本文对每类方法均展开详细的介绍, 并对各方法的主要机制、优缺点和性能进行比较.

2.1 基于监督学习的方法

与许多自然语言生成任务 (如文本摘要抽取、机器翻译等) 相同, 在文本风格迁移 (TST) 任务上也可以使用基于监督学习的序列到序列 (seq2seq) 模型. 通常, seq2seq 模型由编码器和解码器两个部分构成, 并使用平行语料库进行训练. 在训练过程中, 编码器端的输入是需要转换风格的文本, 解码器端的输出是转换目标风格后的文本. 在基于监督学习的 TST 方法中, Jhamtani 等人[23]、Wang 等人[24]以及 Sancheti 等人[25]的方法最具代表性.

文本风格迁移此前基本上是一个手动过程, 几乎没有自动化的 TST 方法. Jhamtani 等人[23] 探索了一种使用神经网络模型将现代英语转换为莎士比亚英语的自动化方法, 首先构建了一个“莎士比亚-现代”(Shakespearean-modern) 英语的外部词典 V, 并引入 PTB 语料[26]以获得更好的单词嵌入. 方法使用了基于注意力 (attention) 机制[27]的 seq2seq 模型, 其中编码器采用双向 LSTM 模型, 解码器采用 RNN 和指针网络 (pointer network)[28]构成的混合模型. 指针网络旨在解决词汇重叠以及专有名词、稀有词无法通过 seq2seq 模型进行预测的问题. 方法对于任意句子对 $(x,y)\in\mathcal{A}$, 预测每个时间步 $t\in\{1,\ldots,T\}$在不同单词 $w\in V$ 上的概率分布 $P_t(w)$, 采用交叉熵损失函数进行训练.

Wang 等人[24]基于 GYAFC 数据集[6]研究“正式-非正式”(formal-informal) 语言 TST 问题, 使用 GPT[29]作为编码器和解码器, 并用预训练模型 GPT-2[30]的参数进行初始化. 由于在预处理过程中可能会引入噪声 (如将专有名词“*R&B*”为“*r&b*”), 因此作者将规则加入预训练模型, 提出了 3 种微调方法, 分别是: 1) 将源句和基于规则处理后的句子进行拼接输入编码器; 2) 将源句和基于规则处理后的句子分别进行编码和解码, 计算两个解码器预测概率的均值; 3) 采用两个编码器分别编码, 但仅使用一个基于层次注意力机制的解码器. 加入规则的方法使用了更多的源句信息, 因此获得了更好的迁移准确率和文本内容保留度.

Sancheti 等人[25]则针对有平行语料的“正式-非正式”(formal-informal) 和“莎士比亚-现代”(Shakespearean-modern) 两个文本风格数据集提出基于 seq2seq 的 TST 方法. 该方法的基础模型与 Jhamtani 等人[23]的模型一致, 即编码器采用 LSTM 模型, 解码器采用 RNN 和指针网络构成的混合模型. 与前人方法不同的是, 该方法提出了一种基于强化学习的框架, 根据转换后样本的风格正确率得分和内容保留度得分作为奖励, 从而促进生成更加符合要求的样本. 实验结果表明, 强化学习的使用能够在保持风格正确率的同时进一步提升文本的内容保留度.

2.2 基于半监督学习的方法

然而如上文所言, 不同风格之间的平行语料往往是缺乏的, 难以直接训练基于 seq2seq 的 TST 模型. 半监督学习旨在研究如何同时利用少量的有类标签的样本和大量的无类标签的样本改进学习性能[31]. 因此, 研究者纷纷提出了基于半监督学习的 TST 方法.

Shang 等人[32]提出隐空间交叉投影 (cross projection in latent space, CPLS) 方法, 以 seq2seq 为基础框架, 定义了在不同风格的隐空间之间的投影函数. 针对小规模平行语料与大规模非平行语料, 分别设计了不同的约束条件来训练投影函数. 模型包含两对编码器和解码器, 分别标记为 A 和 B. 对于平行语料, 将句子对 (a,b) 中的句子 a 通过编码器 A 投影到隐空间 A 中得到 c_a, 随后通过投影函数 $f(\cdot)$ 投影到隐空间 B 中得到 $\widetilde{c}_b$, 再将 $\widetilde{c}_b$ 送入解码器 B 中得到 $\tilde{b}$, 最小化句子 b 的编码 c_b 和 $\widetilde{c}_b$ 的欧氏距离以及 b 和 $\tilde{b}$ 的负对数似然损失来进行训练; 对于非平行语料, 借鉴回译 (back-translation) 机制来回投影重建输入来训练投影函数 $f(\cdot)$ 和 $g(\cdot)$, 其中通过 $f(\cdot)$ 投影得到的 $\widetilde{c}_b$ 会通过 $g(\cdot)$ 重新投影到编码器 A 的隐空间中得到 $\widetilde{c}_a$, 再将其送入解码器 A 得到 $\tilde{a}$, 最小化 $\widetilde{c}_a$ 和 c_a 的欧氏距离以及 a 和 $\tilde{a}$ 的负对数似然损失来进行训练. 为了评估所提出方法的性能, 作者构建并发布了一个关于中文古诗词和现代诗词的风格迁移数据集.

Zhang 等人[33]针对“正式-非正式”文本风格迁移任务提出了 3 种数据增强方法来扩充平行语料, 分别是回译、正式性判别 (formality discrimination) 和多任务迁移 (multi-task transfer). 先使用扩充的平行语料对 seq2seq 模型进行预训练, 再使用原始的平行语料对模型微调. 换句话说, 增强的数据被视为模型的先验知识, 仅在预训练阶段使用. 消融实验证明, 同时使用 3 种数据增强方法取得了最好的效果.

2.3 基于无监督学习的方法

与机器翻译任务相比, 文本风格迁移任务的平行语料更难获取, 因此目前大部分工作均为基于无监督学习的方法, 旨在有效分离文本的属性和内容. 本节首先根据分离数据的形式 (表示级还是文本级) 将这类方法大致分为隐式方法和显式方法, 然后再从学习框架和策略的角度对每类方法作进一步细分.

(1) 隐式方法

该类方法是指模型自动学习句子内容和属性的潜在表示并进行风格的分离与转换. 目前, 主要采用了解缠、强化学习、回译、伪平行语料等策略, 并基于自编码器 (auto-encoder, AE)[34]、变分自编码器 (variational auto-encoder, VAE)[35]、生成对抗网络 (generative adversarial network, GAN)[2]等模型学习文本的潜在表示.

● 解缠策略

解缠策略主要是通过编码器将文本映射到隐空间得到潜在表示, 从而分离内容和属性并进行属性迁移. 图 2 包含了 3 种最常见的解缠策略.

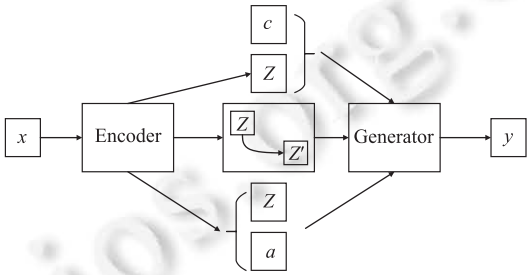


图 2 基于解缠策略的 TST 框架

第 1 种解缠策略最为常见 (图 2 框架的最上分支), 采用了对抗学习的方式, 即将生成的目标句子送入属性判别器, 再通过判别器优化生成器, 从而使得目标属性完全由编码 c 控制, 属性无关的文本内容完全由编码 z 控制.

由于 VAE 和 GAN 生成的文本在很大程度上是不可控的, 因此很少会被用来研究通用文本的生成. Hu 等人[5]最先提出将 VAE 和属性判别器相结合, 通过编码器获得与属性无关的文本内容编码 z, 使用结构化编码 c 控制目标句子的属性, 并通过属性判别器提供反馈信号以优化生成器, 将 z 和 c 结合生成符合目标属性的句子. 损失函数包括生成器 (公式 (1)) 和判别器 (公式 (2)) 两个部分:

$$\min_{\theta_G} \mathcal{L}_G = \mathcal{L}_{\text{VAE}} + \lambda_c \mathcal{L}_{\text{Attr},c} + \lambda_z \mathcal{L}_{\text{Attr},z} \tag{1}$$

$$\min_{\theta_D} \mathcal{L}_D = \mathcal{L}_s + \lambda_u \mathcal{L}_u \tag{2}$$

其中, λ 为权重系数, $\mathcal{L}_{\text{VAE}}$ 为 VAE 模型的损失函数, $\mathcal{L}_{\text{Attr},c}$ 为目标属性的损失函数, $\mathcal{L}_{Attr,z}$ 为文本内容的损失函数, $\mathcal{L}_s$ 为判别器 D 的目标函数, $\mathcal{L}_u$ 为最小熵正则化项:

$$\mathcal{L}_{\text{Attr},c}(\theta_G) = -\mathbb{E}_{\rho(z)\rho(c)}\left[\log_{q_D}\left(c|\widetilde{G_\tau}(z,c)\right)\right] \tag{3}$$

$$\mathcal{L}_{\text{Attr},z}(\theta_G) = -\mathbb{E}_{p(z)p(c)}\left[\log_{q_E}\left(z|\widetilde{G_\tau}(z,c)\right)\right] \tag{4}$$

当句子中存在多类属性时, 该方法可以控制某类属性的迁移. 例如, 在 IMDB 数据集[36]上, 给定具有消极情感的过去时态句子“the acting was also kind of hit or miss”, 该模型可以在保证过去时态的情况下, 生成具有积极情感的句子, 其结果为“his acting was impeccable”. 实验表明该方法能够生成可控的文本, 并有效提高了风格转换的准确率. 不过文中的评价指标过于单一, 缺少文本内容保留度和句子流畅度的评价.

Shen 等人[13]为了解决缺乏平行语料的问题, 假设不同风格的语料库共享文本内容的潜在语义分布, 提出了交叉对齐自编码器 (cross-aligned auto-encoder, CAAE), 学习公式 (5) 的风格迁移函数, 实现内容和风格的分离.

$$p(x_1|x_2;y_1,y_2) = \int p(x_1,z|x_2;y_1,y_2)dz = \int p(z|x_2,y_2)\cdot p(x_1|y_1,z)dz = \mathbb{E}_{z\sim p(z|x_2,y_2)}[p(x_1|y_1,z)] \tag{5}$$

其中, x_1、x_2 分别为含有不同风格 y_1、y_2 的句子, z 为共享的文本内容的潜在表示. 模型主要包括两个步骤: 第 1 步为编码, 将风格 y_1 的句子 x_1 进行编码映射到共享的隐空间中, 得到文本内容 z; 第 2 步为解码, 当 z 联合源风格 y_1

送入生成器时则进行重构操作, 当z联合目标风格y_2送入生成器时则通过判别器进行优化. 方法使用了两个判别器$D1$和$D2$: 其中$D1$用于区分真实样本x_1和转换后的样本x_2, $D2$用于区分真实样本x_2和转换后的样本x_1. 对编码器进行训练, 使其混淆判别器, 学习目标是将不同风格的文本映射到同一个隐空间中. 文中首次加入了人工评价指标, 包括情感风格准确率 (sentiment accuracy)、流畅度 (fluency) 和整体迁移质量 (overall transfer). 其中, 前两个指标用于评价由该模型获得的目标句, 而整体转移质量则是对同一源句的两个目标句 (由该模型及其对比模型分别获得) 进行比较性评估.

随后, Yang 等人[14]又提出一种与 CAAE 相似的方法, 不同之处在于其判别器采用的是目标域的语言模型, 因此可以为学习过程提供更丰富、更稳定的 token 级的反馈.

为了进一步提升深度隐变量模型在文本序列这类离散结构上的表现, Zhao 等人[15]扩展了 Wasserstein 自编码器 (Wasserstein autoencoder, WAE)[37], 提出对抗正则化自编码器 (adversarially regularized autoencoders, ARAE). 该模型首先扩展了 WAE 来对离散序列进行建模, 然后针对不同的可控表征学习先验知识, 通过训练属性分类器D进而训练编码器以获得该分类器D无法区分的潜在表示z. ARAE 能够在隐空间中更改源句的离散属性为目标属性, 从而生成符合要求的目标句.

Fu 等人[16]提出了基于多解码器 (multi-decoder) 和风格嵌入 (style-embedding) 的模型, 并采用对抗网络更好地分离文本内容和风格. 该模型将编码器学习到的潜在表示送入风格分类器, 通过最大化风格预测值的熵 (即最小化负熵) 使得分类器无法判别出其对应的风格. 损失函数定义为训练数据中风格标签的负对数似然函数. 这里, 多解码器是指对于不同的风格使用不同的解码器. 风格嵌入模块是将得到的文本内容和目标属性进行拼接送入解码器得到目标风格的句子. 文中提出了迁移程度 (transfer strength) 和内容保留 (content preservation) 两个新的评价指标. 前者采用 LSTM-sigmoid 分类器的得分来评价生成句子是否符合目标风格, 后者定义为源句和目标句的余弦相似度.

为了能在保持文本内容的情况下同时提高风格转换的准确率, Yi 等人[38]提出了一种通过学习多个目标句子实例的方法来提取潜在的风格属性. 该方法首先从目标域随机选取一定数量的句子, 使用编码器进行编码, 然后映射到风格空间中, 从而得到目标句子的风格表示, 再将其和源句子的编码一起送入生成器, 得到目标句子, 最后通过判别器来优化生成器. 实验表明, 通过增强风格属性信号, 该方法再风格转移准确率和内容保留之间取得更好的平衡.

第 2 种解缠策略 (图 2 框架的中间分支) 则是在属性分类器的指导下对潜在表示进行编辑, 迭代执行这一过程直到潜在表示具有目标属性类别为止. 其中, z表示为经过编码器编码后得到的潜在表示, z'表示在属性分类器监督下优化后得到的潜在表示.

Wang 等人[39]采用多任务学习对隐空间内的潜在表示进行迭代更新, 使得属性分类器对目标属性的预测置信度得分最大化. 整个框架可以分为 3 个子模型: 一个编码器, 将文本x编码为潜在表示z; 一个解码器, 将z解码为输出文本y; 以及一个属性分类器, 用于对潜在表示z的属性进行分类. 多个学习任务包括优化编码器、优化解码器和优化属性分类器, 通过优化后的潜在表示z能够生成更加符合目标属性的句子.

Liu 等人[40]提出渐进式地优化文本在隐空间中的表示z以实现文本风格的迁移. 该方法由 3 个关键组件组成: 变分自编码器 (VAE), 属性预测器 (每种属性对应一个) 和文本内容预测器. 首先, 将离散空间中的句子映射到连续空间; 然后, 对 VAE 和两种属性预测器执行基于梯度的优化, 最终找到具有目标属性且保留内容的目标句子的潜在表示z'. 此外, 该方法具有同时处理多粒度 (如句子级和单词级) 属性的能力, 因此具有更好的可解释性.

第 3 种解缠策略 (图 2 框架中的最下分支) 是先将输入的文本编码为两个潜在表示, 一个包含源属性信息 (即a), 另一个包含文本的内容信息 (即z), 然后将a替换为目标属性 (即c), 最后使用z和c的组合进行解码.

John 等人[41]探讨了在语言模型中分离风格和内容的潜在表征的问题, 将辅助性多任务和对抗性目标相结合. 通过设计面向风格的辅助性损失以确保风格信息包含在z中, 再通过设计面向内容的辅助性损失来对内容信息进行正则化, 分别在风格预测和词袋预测两个不同任务中进行训练. 实验结果表明, 该方法能够在潜在空间中有效分离出文本的内容特征, 在风格迁移准确率、内容保留度和语言流畅性 3 个方面均取得了不错的效果.

● 强化学习策略

强化学习是指智能系统在与环境的连续互动中学习最优行为策略的机器学习问题, 其目标是最大化长期累积的奖励[42]. 在文本风格迁移领域, 可通过设计不同的奖励函数机制来促进模型更好地学习文本内容和风格表示. 基于强化学习策略的模型框架如图 3 所示. 其中, Generator 是指根据需要选择的某种编码解码模型, 目的是将源句转换成目标句, 奖励可基于迁移准确率、文本内容保留、句子流畅度等方面 (即图 3 右侧的 3 个分支) 进行定义.

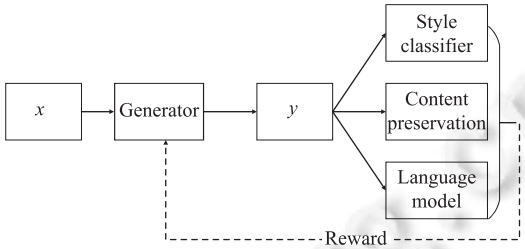


图 3 基于强化学习的 TST 框架

Luo 等人[17]把源句到目标句以及目标句到源句的映射看作一个双重学习任务, 以 seq2seq 为基准模型, 通过强化学习的训练方式获得反馈以解决平行语料缺乏的问题. 他们设计了风格准确率 R_s 和内容保留度 R_c 两个奖励机制, 见公式 (6) 和公式 (7):

$$R_s = P\left(s_y|y';\varphi\right) \tag{6}$$

$$R_c = P\left(x|y';\phi\right) \tag{7}$$

这里, x 为重构后的句子, y' 为模型生成的句子, φ和ϕ 分别为可学习参数. 整个奖励函数定义为:

$$R = \left(1+\beta^2\right)\frac{R_c \cdot R_s}{\left(\beta^2 \cdot R_c\right)+R_s} \tag{8}$$

其中, β 是一个调和权重参数.

随后, Gong 等人[18]提出基于强化学习的生成器和评估器架构, 该方法由以下模块组成: 生成器、风格鉴别器、语义模块和语言模型. 生成器使用了结合注意力机制的 RNN 模型, 将源域的句子编码作为输入并将其转换为目标域的句子; 风格鉴别器是一个具有注意力机制的双向 RNN 分类器; 语义模块则通过计算词移距离 (word mover's distance, WMD)[43,44]来测量基于词嵌入的输入句子和输出句子之间的语义差异; 语言模型使用 RNN 来判别句子的流畅度; 最后, 将风格鉴别器、语义模块和语言模型的结果作为奖励反馈给生成器进行进一步的优化. 该模型能够提高风格转换后的句子在语法上的正确性.

● 回译策略

回译 (back translation)[45,46]是机器翻译中常用的一种策略, 用于更好地利用单语语料以辅助翻译模型的训练. 假设有目标语言句子 y, 用训练好的目标语言到源语言的翻译模型得到伪平行句子对 (x', y), 然后加入到平行句子对集合中一起训练. 尽管 x' 可能包含一些未知词 (UNK) 或者错误的句法, 然而由于 y 是高质量的单语语料, 因此这样训练可以想像成一种去噪声的训练形式, 即在有噪声的情况下, 通过训练 $x \to y$ 方向的翻译模型更好地学习源句的潜在表示. 由于文本风格迁移任务也可视为一类翻译任务, 因此可借鉴回译策略, 基于语言翻译模型学习源句的潜在表示, 从而弱化句子的属性.

Prabhumoye 等人[19]最先将回译思想引入文本风格迁移任务中. 首先, 使用预训练好的翻译模型将源句翻译成另一种语言, 再对其进行编码得到 z:

$$z = E\left(x_f;\theta_E\right) \tag{9}$$

其中, x_f 为 f 语言的句子, θ_E 为编码器的参数. 然后, 使用不同的解码器生成对应风格的句子, 再通过风格分类器来反作用于解码器, 促进生成更符合特定风格的句子. 例如, 给定英文源句"i thank you, rep. visclosky", 该方法首先

经过翻译模型翻译成法文“je vous remercie rep. visclosky”, 再进行编码, 最后通过不同风格的解码器生成具有源风格属性的英文句子“i thank you, senator visclosky.”以及目标风格属性的句子“i’m praying for you sir”.

● 伪平行语料策略

与上述回译策略的目标不同, 另一类方法构建伪平行语料, 通过迭代的方法优化翻译模型, 实现风格迁移. 主要包括两个过程: 先通过构建伪平行语料, 再在相应的数据集上进一步训练这风格迁移模型.

Liao 等人[47]通过从 YELP 数据集中搜索内容相似但情感不同的参考语料, 生成了一个伪平行语料库. 例如, 给定句子“食物很糟糕”, 匹配的语料可能是“食物很好吃”, 其中“食物”是相同的内容而两个句子具有不同的情感. 生成的伪并行语料数据集随后用于训练变分自编码器以执行 TST 任务. 该框架通过对伪平行句子的内容相似性和属性差异性进行建模, 从而更好地分离文本的内容和属性, 并生成文本内容保留度较高的目标风格句子.

Zhang 等人[48]利用风格偏好信息和单词嵌入相似度, 在统计机器翻译 (statistical machine translation, SMT) 框架下首先学习不同风格词间的转换表, 并基于此生成初始的伪平行语料. 接着, 用该语料预训练源域和目标域的双向映射, 并在此基础上设计了迭代回译 (iterative back-translation, IBT) 的方法进一步优化模型, 生成更加符合目标属性的句子.

Jin 等人[49]随后进一步提出了基于迭代匹配和翻译 (iterative matching and translation, IMAT) 的 TST 方法, 避免了学习单词翻译表的复杂过程, 而是先对源语料和目标语料中语义相似的句子子集进行对齐、构建伪平行语料库, 再采用标准的 seq2seq 模型来学习文本风格迁移函数, 最后通过迭代的方式改进迁移函数以细化伪平行语料对齐中的缺陷.

● 其他特殊策略

此外, 还有一些采用特殊策略 (如域自适应、概率模型等) 的 TST 方法. 这些方法较为独立, 且存在特定的假设或前提.

Li 等人[50]提出了一种基于域自适应的 TST 模型, 利用来自源域 (如餐馆评论) 的大量可用数据, 解决在目标域 (如电影评论) 中数据稀缺且与源域数据分布不匹配的问题, 使得文本风格迁移能够以域感知的方式进行. 在这一设定下, 所提出的模型除了学习文本的风格之外, 还要学习源域和目标域中文本的域向量. 这里, 域向量用于区分不同的域, 例如电影评论还是餐厅评论, 可以激励模型以域感知的方式执行风格迁移, 而不是直接共享风格迁移模型, 有效避免了生成诸如“电影是美味的！”之类不符合要求的句子.

He 等人[51]提出了一种基于概率的深度生成模型来推断 TST 中句子的潜在表示, 将非平行语料库视为部分可观测的平行语料库, 并减弱了独立性假设. 假设初始数据集为: 风格 D_1 的可观测数据 $X=\left\{x^{(1)},x^{(2)},\ldots,x^{(m)}\right\}$ 和风格 D_2 的可观测数据 $Y=\left\{y^{(m+1)},y^{(m+2)},\ldots,y^{(n)}\right\}$, 目前无平行语句, 即不存在上标相同的语句. 文中引入隐语句 (latent sentence), 将数据集补充为平行语料库, 即引入 $\bar{X}=\left\{\bar{x}^{(m+1)},\bar{x}^{(m+2)},\ldots,\bar{x}^{(n)}\right\}$ 表示 D_1 中不可见的部分, $\bar{Y}=\left\{\bar{y}^{(1)},\bar{y}^{(2)},\ldots,\bar{y}^{(m)}\right\}$ 表示 D_2 中不可见的部分. 因此模型的学习目标为基于可观测的 X、Y, 推断不可观测的 $\overline{X}$、$\overline{Y}$, 通过求解联合概率学习 $p(\bar{y}|x)$ 和 $p(\bar{x}|y)$. 方法采用了基于注意力机制的 seq2seq 模型, 并采用循环语言模型建模先验信息以减弱独立性假设, 以此推断假设的平行语料库中句子的潜在表示, 然后送入解码器生成特定风格的句子.

Li 等人[52]首次将文本风格迁移任务建模为一个先搜索后学习的问题, 提出基于搜索学习的无监督文本生成 (unsupervised text generation by learning from search, TGLS) 框架. 其包含两个模块, 一个在句子空间中的搜索模块, 和一个基于搜索结果的学习模块. 搜索模块采用模拟 SA 算法[53], 将搜索生成得到的文本作为“伪标注文本”, 得到一个搜索器, 然后根据启发式的评分函数来评估一个句子质量, 包括流畅度得分、语义得分和特定任务得分 3 个方面; 学习模块采用单词级的交叉熵损失和序列级的最大边距损失来训练条件生成模型, 得到新文本. 在 GYAFC 数据集上的实验验证了该方法在内容保留度和流畅度上均取得了提升.

(2) 显式方法

该类方法认为, 句子的属性通常体现在独特的短语中, 如在句子“we sit down and we got some really slow and lazy service”中, “slow”和“lazy”是句子中的形容词, 能够体现该句子的消极属性. 因此一种简单有效的方法是只替

换属性词, 而不是从头生成一个新句子, 即只需要改变风格相关的词或短语而保留风格无关的部分就可以达到风格迁移的目的.

该类方法一般分成 3 步: 1) 删除 (delete), 即找到并删除句子中的属性词; 2) 检索 (retrieve), 即检索与文本内容最相似的目标句子; 3) 生成 (generate), 即结合目标属性生成目标句子. 本文将这类方法统称为 DRG 方法, 框架如图 4 所示. 其中虚线表示删除属性词之后也可不通过检索步骤而直接生成目标句子.

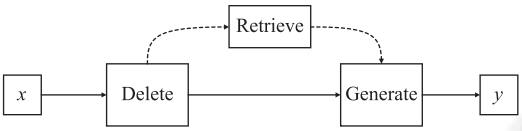


图 4 基于 DRG 的 TST 框架

DRG 方法的关键在于第 1) 步, 即如何更好地定位属性词. 本文根据不同的属性词删除策略将 DRG 方法进一步细分为以下 3 类.

● 基于词频的删除策略

Li 等人[20]最先提出了 DRG (delete, retrieve, generate) 的框架. 属性词的判别是基于词频计算获得的:

$$s(u,v)=\frac{count(u,D_v)+\lambda}{\left(\sum_{v'\in V,v'\neq v}count(u,D_{v'})\right)+\lambda} \tag{10}$$

其中, D_v 为具有属性 v 的句子集合, $D_{v'}$ 为不具有属性 v 的句子集合, $count(u,D_v)$ 表示 n-gram(u) 在 D_v 中出现的次数, λ 为平滑参数. 如果 $s(u,v)$ 值大于阈值 γ, 则认为 u 是属性词. 例如, 源属性为积极情感的句子“i have had this mount for about a year and it works great”中, 属性词为“works great”. 通过删除步骤, 可得到文本内容 $c(x,v^{\text{src}})$ (i have had this mount for about a year and it). 经过检索步骤, 可得到目标属性 $a(x,v^{\text{tgt}})$ (短语“barely used”). 检索方法定义为:

$$\begin{cases}x^{\text{tgt}}=\text{argmin}_{x'\in D_{v^{\text{tgt}}}}F\\ F=d(c(x,v^{\text{src}}),c(x',v^{\text{tgt}}))\end{cases} \tag{11}$$

其中, d 是计算两个句子序列的距离度量函数, 文中分别采用 TF-IDF 加权词重叠和欧氏距离两种度量. 生成方法采用了 4 种策略. 前两种结合深度学习模型, 分别为: 仅删除策略 (delete-only), 即将删除后的文本内容和目标属性词相结合送入 RNN, 通过最大化目标函数来重建训练语料句子的内容和源属性值, 损失函数为公式 (12), 对于上例生成目标句“i have had this mount for about a year and it still works”; 删除检索策略 (delete-retrieve), 即将删除后的文本内容和检索后得到的目标属性词结合送入 RNN, 损失函数为公式 (13), 对于上例生成目标句“i have had this mount for about a year and barely used it”. 后两种采用传统的非深度学习方法, 分别为: 基于模板的策略 (template-based), 即不经过生成器, 直接结合文本内容和属性词, 对于上例生成目标句“i have had this mount for about a year and it barely used”; 仅检索策略 (retrieve-only), 即直接将检索得到的目标句子当作最终输出, 对于上例生成目标句“i have had it for a while but barely used it”. 实验结果表明, 基于模板的策略在文本内容保留上表现最好, 仅检索策略在迁移准确率上表现最好.

$$L(\theta)=\sum_{(x,v^{\text{src}})\in D}\log p(x|c(x,v^{\text{src}}),v^{\text{src}});\theta) \tag{12}$$

$$L(\theta)=\sum_{(x,v^{\text{src}})\in D}\log p(x|c(x,v^{\text{src}}),\alpha'(x,v^{\text{src}});\theta) \tag{13}$$

● 基于注意力机制的删除策略

注意力机制最早应用于 RNN 模型中进行图像分类任务[54]. Bahdanau 等人[55]在机器翻译模型的中间增加了一层注意力机制, 同时进行源语言与目标语言的短语翻译和对齐. 最近, 注意力机制也出现在文本风格迁移中, 用于属性词的删除.

Xu 等人[56]提出的一种情感属性迁移方法, 包含中立模块 (neutralization module) 和情感模块 (emotionalization module). 前者通过训练一个基于注意力机制的分类器来显式分离表示文本内容的中性词和表示属性的情感词, 情感词往往获得较高的注意力权重, 而中性词通常获得较低的注意力权重. 其隐藏层注意力权重定义为:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i=0}^{T}\exp(e_i)} \tag{14}$$

其中, $e_i = f(h_i, h_T)$ 是一个对齐模型, 用来评估每个词对情感分类的贡献, 最后一个隐藏状态 h_T 视为上下文向量, 包含输入序列的所有信息. 最后将分离出的文本内容和目标属性进行结合. 文章使用循环强化学习对这两个模块进行训练, 奖励函数为:

$$R = 1 + \beta^2 \frac{BLEU \cdot Confid}{(\beta^2 \cdot BLEU) + Confid} \tag{15}$$

其中, BLEU[57]是衡量机器翻译质量的指标, 在文本风格迁移中常用于评价内容保留度, *Confid* 表示转换后达到目标属性的正确率, β 是超参数.

Zhang 等人[58]针对情感属性迁移任务提出了基于自注意力机制的情感属性词删除方法. 文中将注意力的权重定义为:

$$\alpha_t = sigmoid\left(v^{\mathrm{T}} h_t\right) \tag{16}$$

其中, v 是参数向量, sigmoid 函数给每个输入单词一个接近 1 或 0 的可分辨权重, h_t 为隐藏层状态. 作者将注意力权重映射到离散值 0 或 1 上, 注意力权重大于平均值的词被认定为情感词, 小于平均注意力权重的词被认定为文本内容.

由于文本预训练模型具有强大的特征提取能力[59], Sudhakar 等人[60]将其应用到 DRG 框架中, 提出了两个 Transformer 的变种模型 B-GST (blind generative style transformer) 和 G-GST (guided generative style transformer). 其中, G-GST 模型在生成部分中的输入为删除属性词后得到的文本内容和检索得到的目标句子属性, 因此该模型适合目标语料与源语料有相似句子的情况, 并可以在推理阶段手动指定想要的目标属性词来获得风格更准确的句子. 而 B-GST 模型则没有检索步骤, 适合目标语料与源语料没有相似句子的情况, 但是可能会检索出内容不一致的目标句子. 实验结果表明两个模型均能够较好地分离文本内容和属性词. 与之前的方法相比, B-GST 模型在文本内容保留度和流畅性上效果更优.

Malmi 等人[61]提出了一个新的模型 MASKER, 通过在源域和目标域训练掩蔽语言模型 (MLM), 找到两个模型在似然性方面最不一致的文本跨度 (text span), 从而识别源句中待删除的分词 (token). 删除的 tokens 通过使用填充的 MLM (padded masked language model) 替换, 无需预先确定插入 tokens 的数量. 在句子融合和情感迁移上的实验验证了该方法在完全无监督的情况下性能较好, 并且在低资源环境中比以前的方法表现更好.

● 基于词频和注意力机制结合的删除策略

Wu 等人[62]认为文本风格迁移和完型填空任务非常相似, 因此提出了一种“掩蔽-填充”(mask-infill) 两步法. 首先定位属性词并用掩蔽 (mask) 来代替, 达到分离文本内容和情感属性的目的; 然后, 改进掩蔽语言模型 (mask language model, MLM) 得到属性条件 MLM 模型 (attribute conditional MLM). 文中使用了词频和注意力机制结合的方法定位属性词:

$$s(u,a) = s_c(u,a) * p \tag{17}$$

其中, p 为每个候选属性词在基于注意力机制分类器下得到的概率. 然后以被掩蔽的属性词上下文和目标属性为条件预测单词或短语来填充被掩蔽的位置. 如给定积极情感属性的源句子“beautiful scenery and good service”, 经过基于词频和注意力机制结合的删除策略后, 变为“[MASK] scenery and [MASK] service”, 再通过填充, 使其成为具有消极情感属性的目标句子“terrible scenery and poor service”. 消融实验表明, 基于注意力机制的删除策略在迁移准确率上效果最好, 而基于词频和注意力机制结合的删除策略在文本内容保留度上表现较好. 表 3 从动机、关键性技术、优缺点和性能等方面总结了主流的无监督文本风格迁移方法.

表 3 无监督的文本风格迁移方法对比

方法	动机	关键性技术	优点	局限性	性能		
					ACC	BLEU	PPL
Ctrl-Gen[5]	生成属性可控的高质量文本句子	基于属性编码控制的解缠、可控生成	高质量可控	可能生成语法错误的句子	很好	较好	一般
CAAE[13]	跨不同文本语料库共享潜在文本内容分布	基于属性编码控制的解缠、交叉对齐编码	基本符合目标属性且流畅度较好	文本内容可能会发生改变	较好	较差	较好
ARAE[15]	训练离散结构的深度潜变量模型	基于属性编码控制的解缠、对抗正则化编码	基本符合目标属性且流畅度较好	文本内容可能会发生改变	较好	一般	较好
MD[16]	学习单独的内容表示和属性表示	基于属性编码控制的解缠、多解码器	句子流畅性较好	可能生成不符合目标属性的句子	较差	一般	较好
SE[16]		基于属性编码控制的解缠、风格编码	句子流畅性较好	极可能生成不符合目标属性句子	较差	一般	较好
Ctrl-LRE[39]	基于属性分类器以最少的潜在表示编辑取代对属性建模的过程	潜在表示迭代优化的解缠	综合性较好	可能会产生失败的案例	很好	很好	较好
DAE[41]	学习语言模型中属性和内容的潜在表示解缠	源属性和文本内容潜在表示分离的解缠	句子流畅性较好	文本内容很可能会发生改变	一般	较差	很好
DualRL[17]	一步映射直接迁移文本的风格, 而不需要任何内容和风格的分离	对偶强化学习	文本内容保留度高	可能生成不符合目标属性的句子	一般	很好	较好
RLS[18]	在强化学习过程中解缠文本内容和属性	强化学习	基本符合目标属性且流畅度较好	文本内容可能会发生改变	较好	较差	较好
BST[19]	弱化句子属性更好保留文本内容	回译	基本符合目标属性且流畅度较好	在各个数据集上的表现有差异	较好	较差	较好
IBT[48]	将TST看作无监督机器翻译任务	迭代回译	生成的句子符合目标属性	没有衡量句子的流畅度	很好	较好	—
IMAT[49]	更好的保留文本内容	迭代匹配和翻译	生成的句子符合目标属性	没有衡量句子的流畅度	很好	较好	—
DAST[50]	利用来自其他域的大量可用数据	领域自适应文本转换	综合性较好	句子流畅性不是很好	很好	较好	较好
DAST-C[50]			句子流畅性不好		很好	较好	较差
Template[20]	文本属性常由特定的词表示	基于词频的属性次删除	构建了多种方法, 满足不同的需求	缺乏普适性较好的方法	一般	较好	较差
RO[20]					很好	较差	—
DO[20]					较好	一般	很好
DAR[20]					很好	一般	很好
Cycle-RL[56]	分为中和模块和情感模块解决无平行语料问题	基于注意力机制的属性词删除	文本内容保留度较好	可能生成不符合目标属性的句子	较差	较好	—
SMAE[58]	使用非情感的上下文为情感词的出现提供了指引	基于注意力机制的属性词删除	基本符合目标属性且流畅度较好	文本内容可能会发生改变	较好	较差	较好
B-GST[60]	引入预训练模型Transformer更好的删除属性词	基于注意力机制的属性词删除	综合性较好	可能会产生失败的案例	很好	较好	较好
G-GST[60]			文本内容保留度较好	可能生成不符合目标属性的句子	较差	较好	一般
AC-MLM-SS[62]	将TST任务看作完型填空任务进行遮蔽和填充	词频和注意力结合的属性词删除	符合目标属性且流畅度较好	没有衡量句子的流畅度	很好	较好	—
MASKER[61]	根据源域和目标域的文本跨度不同查找属性词	基于注意力机制的属性词删除	符合目标属性且流畅度较好	没有衡量句子的流畅度	一般	较好	—
StyIns[38]	从多个目标句子中实例提取潜在的属性	基于多目标句子属性编码	符合目标属性且流畅度较好	可能生成语法错误的句子	很好	较好	一般

本文将各方法的性能从高到低分为 4 个层次: 很好、较好、一般、较差. 该划分参考了各方法的原文以及 Hu 等人[63]的实验结果, 计算了所有方法在 3 个评价指标上的均值, 即 ACC=81.5%, BLEU=34.0, PPL=325.7, 并将其作为基准指标. 表 2 中, “很好”表示该方法的指标与基准指标相比 ACC 提高 5% 以上、BLEU 值提高 20 以上、PPL 值降低 200 以上; “较好”表示该方法的指标和基准指标相近, 即 ACC 提高在 5% 以内、BLEU 值提高在 20 以内、PPL 值降低在 200 以内; “一般”表示该方法的指标与基准指标相比 ACC 降低了 5% 以内、BLEU 值降低了 20 以内、PPL 值增加了 200 以内; “较差”表示该方法的指标与基准指标相比 ACC 降低 5% 以上、BLEU 值降低 20 以上、PPL 值增加 200 以上.

3 实 验

3.1 数据集

表 4 汇总了文本风格迁移领域常用的数据集及其风格示例.

表 4 TST 数据集

Methods	Dataset	Attributes	Application examples
Supervised learning	Shakespeare	Modern Shakespearean	Give me one kiss and I’ll *go down*. One kiss I’ll *descend*.
	GYAFC	Informal Formal	*i* tried to like him, *but i just can’t*. I tried to like him, but *I cannot*.
Unsupervised learning	YELP	Positive Negative	I would *definitely recommend* this for a cute case. I would *not recommend* this for a cute case.
	Amazon	Positive Negative	the food is *fresh* and the environment is *good*. the food is *bland* and the food is the *nightmare*.
	Tense	Past Present Future	I *thought* the movie *was* too bland and too much. I guess the movie *is* too bland and too much. I guess the film *will* have been too bland.
	Topic	Music Science Politics	what is your *favorite sitcom* with adam sandler ? what is an event horizon with regards to *black holes* ? what is an event with *black people* ?
	Paper-News	Paper News	*foreign banking* carbon on fitness technology. *luxury fashion* takes on fitness technology.
	Gender	Male Female	my wife ordered *country fried steak* and *eggs*. my *husband* ordered the *chicken salad* and the *fries*.
	Caption	Factual Humorous	a young man dances by a fountain. a young man sits by a fountain *like a monkey with a smiley face* .
	Political	Republican Democratic	I absolutely agree with *senator Paul’s* actions. I absolutely agree with *Elizabeth warren’s* actions .

平行语料数据集包括:

- Shakespeare: 该数据集包括现代风格英语和莎士比亚风格英语之间的平行语料.
- GYAFC: 该数据集包括娱乐音乐 (E&M) 和家庭关系 (F&R) 两个领域的数据, 含有正式语句和非正式语句之间的平行语料.

非平行语料数据集包括:

- YELP: 该数据集来源于美国最大的点评网站 YELP, 语句具有正负情感标签.
- Amazon: 该数据集来源于亚马逊购物网站的商品评论, 语句具有正负情感标签.
- Tense: 该数据集来源于 TimeBank 网站 (timeml.org), 包含过去、现在、未来 3 个时态的语句.
- Topic: 该数据集来源于 Yahoo QA, 包含科学、音乐、政治 3 个类别的语句.
- Paper-News Title: 该数据集包含论文风格的语句和新闻风格的语句, 前者是从学术网站抓取的论文标题, 后

者是 UCL 数据集中选取的新闻标题.

● Gender: 该数据集包含 YELP 网站上对食品企业的评论, 每条评论都具有性别标签.

● Caption: 该数据集为图片的文字说明或标题, 分别标记为真实、浪漫或幽默 3 种标签.

● Political: 该数据集为美国参众两院的政客在 Facebook 上发表的评论, 每条评论都被贴上了共和党或民主党的标签.

本文的验证性实验选取了目前广泛使用的 YELP、Amazon 和 GYAFC 数据集. 其中, YELP 和 Amazon 来源于 Li 等人[20]的工作, GYAFC 来源于 Rao 等人[6]的工作, 且数据集的处理与划分方式均与原文一致. 各数据集的统计信息如表 5 所示.

表 5 各数据集统计信息

Dataset	Attributes	Train (k)	Dev	Test
YELP	Positive	270	2 000	500
	Negative	180	2 000	500
AMAZON	Positive	277	985	500
	Negative	279	1 015	500
GYAFC	Formal	50	1 019	500
	Informal	50	1 019	500

3.2 评价指标

一个好的文本风格迁移模型应该满足生成的句子符合目标属性、文本内容保留度高、语言流畅性好等不同方面的性能. 不同的工作采用或定义的评价指标也不尽相同. 本文对已有工作使用的评价指标以及 Mir 等人[64]总结的评价指标进行概括 (如表 6 所示).

表 6 评价指标

Automatic Evaluation	Human Evaluation
Style Accuracy	Transfer Strength
BLEU	Content Preservation
PPL	Fluency

本文将文本风格迁移任务的评价指标分为两大类: 自动化评价 (automatic evaluation) 和人工评价 (human evaluation), 并且这两大类都分别从 3 个方面去评价, 分别是: 转换后的句子是否满足目标属性、文本内容是否得到保留、转换后的句子是否流畅或语法是否有错误.

自动化评价一般包含准确率 (accuracy, ACC)、BLEU 以及困惑度 (perplexity, PPL) 等. 其中, 准确率是通过预训练好的分类器 (如 TextCNN[65]、FastText[66]等) 来判断转换后的句子是否满足目标属性. BLEU 是机器翻译中评价内容一致性的指标, 它的计算方式为:

$$BLEU = BP \times \exp \int_{n=1}^{N} W_n \log P_n \tag{18}$$

其中, BP (brevity penalty) 为长度惩罚因子, $W_n = 1/N$ (N 的上限为 4), P_n 表示 n-gram 的精度. PPL 是通过预训练好的语言模型 (如 GPT-1[29]、GPT-2[30]、BERT[59]等) 来判断转换后句子的语法正确性和流畅度, PPL 值小, 语句越流畅, 计算方式为:

$$PPL(S) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{p(w_1 w_2 \dots w_N)}} = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{p(w_i | w_1 w_2 \dots w_{i-1})}} \tag{19}$$

其中, S 代表一个句子, N 是句子长度, $p(w_i)$ 是第 i 个词的概率.

人工评价具有一定的主观性且耗时耗力, 一般作为自动化评价的补充. 常用方法是随机挑选一些转换后的句子, 将其和源句子一起交给语言学家评判, 但不透露句子的风格信息. 和自动化评价一样, 人工评价也是从迁移准

确率、文本内容保留度、语法正确性 3 个方面进行打分, 打分范围在 0~5, 最后计算平均得分.

3.3 实验结果

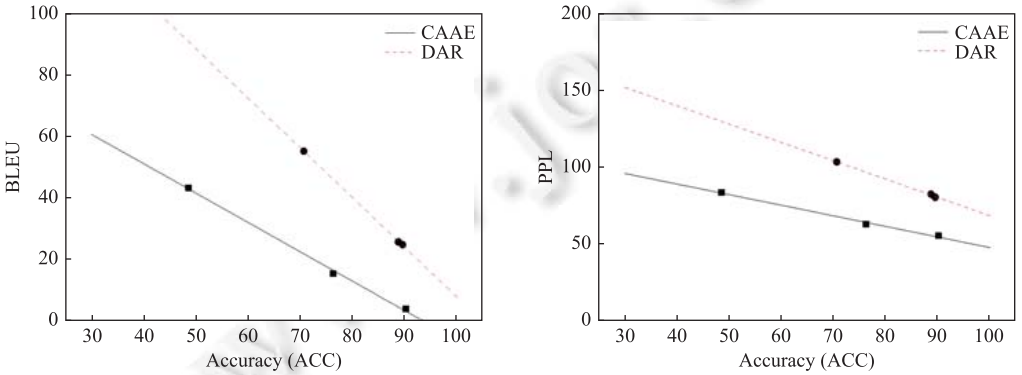
本文在 3 个数据集上对基于无监督学习的文本风格迁移代表性方法进行了实验, 并列出各方法在自动化评价指标上的结果 (如表 7 所示).

实验结果表明, 数据集的分布对模型的表现有一定的影响, 以迁移准确率 (ACC) 为例, 没有一种方法在 3 个数据集上均取得最优的表现. 此外, 也没有一种方法在 3 个指标上均表现最优. 大多数模型更趋近于在某些指标上表现好, 而在另外指标上表现较差. 例如: RO[20]、BST[19]模型尽管获得了很高的迁移准确率, 但 BLEU 值很低, 说明其在保留源文本内容方面是无效的. 而 SE 模型[16]正好相反, 在文本内容保留度上表现不错, 但在迁移准确率和流畅度上表现很差. 从结果中还观察到, 隐式 TST 方法 (表 7 中上半部分) 与显式 TST 方法 (表 7 中下半部分) 相比在文本内容保留度上表现相对较差, 总体来说, DAR 模型[20]和 B-GST 模型[60]在 3 个指标上均有较好的表现.

表 7 自动化评价实验结果

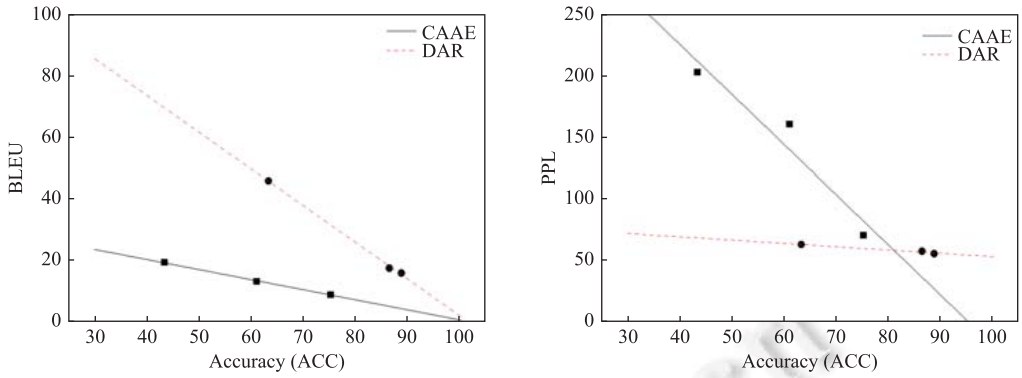
模型	YELP			AMAZON			GYAFC		
	ACC (%)	BLEU	PPL	ACC (%)	BLEU	PPL	ACC (%)	BLEU	PPL
CAAE[13]	76.2	15.2	62.9	75.1	8.9	70.2	66.8	3.6	35.2
SE[16]	8.6	24.5	163.6	38.2	15.0	60.1	23.5	8.2	86.3
MD[16]	52.3	20.5	90.3	55.3	15.3	76.1	24.9	11.5	97.3
Template[20]	81.1	28.9	185.6	67.8	30.6	80.3	50.6	34.3	100.5
RO[20]	95.2	4.7	25.7	70.3	10.4	60.2	13.4	15.6	96.8
DO[20]	87.5	24.9	81.4	46.1	28.3	94.5	20.2	28.2	103.4
DAR[20]	89.5	24.7	80.4	88.6	15.9	55.4	61.1	20.2	110.3
BST[19]	90.8	6.8	32.8	76.7	7.5	48.3	70.5	1.3	50.2
Cycle-RL[56]	80.1	12.5	145.3	68.7	14.2	183.2	70.2	22.5	66.1
B-GST[60]	86.7	57.2	102.0	61.1	70.1	60.1	78.2	39.5	99.5
G-GST[60]	77.2	43.9	134.4	58.6	72.5	165.8	76.3	36.7	90.5

除此之外, 本文借鉴了 Mir 等人[64]和 Hu 等人[63]的工作, 从隐式方法和显式方法中分别选取了 CAAE 和 DAR 模型作为代表, 对评价指标进行了度量权衡分析. 包括风格迁移准确率 (ACC) 和文本内容保留 (BLEU) 的权衡分析, 以及风格迁移准确率 (ACC) 和流畅度 (PPL) 的权衡分析. 在 YELP、Amazon 和 GYAFC 这个数据集上的分析结果分别如图 5、图 6、图 7 所示.



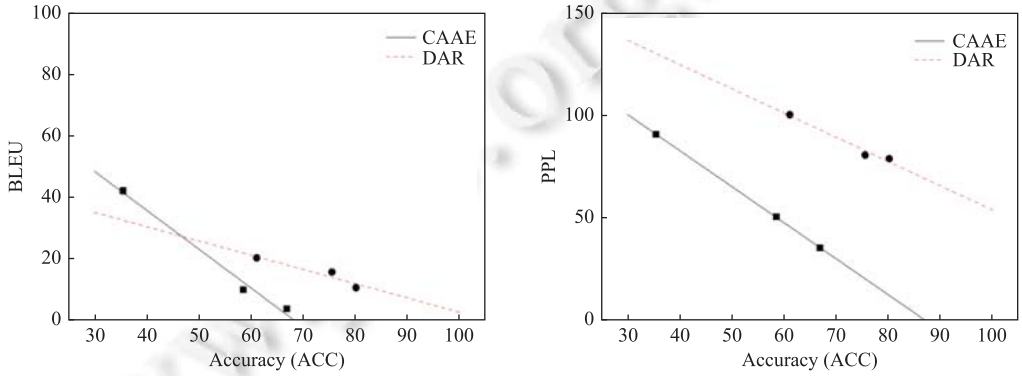
(a) Style and content trade-off on YELP (b) Style and fluency trade-off on YELP

图 5 YELP 数据集上情感迁移的度量权衡分析



(a) Style and content trade-off on Amazon (b) Style and fluency trade-off on Amazon

图 6 Amazon 数据集上情感迁移的度量权衡分析



(a) Style and content trade-off on GYAFC (b) Style and fluency trade-off on GYAFC

图 7 GYAFC 数据集上的正式语句迁移度量权衡分析

图 5、图 6 展示了情感迁移任务上多个指标的权衡分析结果. 对于两个模型, 随着 ACC 的增加, BLEU 和 PPL 均会下降, 说明随着迁移准确率的提升, 句子也会更加流畅, 但内容保留度会降低. 在两个数据集上, 随着 ACC 值的增加, DAR 模型的 BLEU 值的下降趋势更为显著, 但总体来说 ACC 和 BLEU 均优于 CAAE 模型. DAR 模型的 PPL 值随着 ACC 的提升并无显著下降, 说明显式方法中提升迁移准确率对于句子的流畅度的影响不大. 对于 CAAE 模型, 不同的数据集上随着 ACC 的提升, PPL 呈现了不同的下降趋势, 如在 Amazon 数据集上的 PPL 下降较快, 说明该模型在提升迁移准确率的同时也能很好地提高句子的流畅度. 在图 7 的正式-非正式风格迁移任务上, 即使风格迁移准确率降到最低, CAAE 模型仍然难以很好地保留文本的内容, 而 DAR 模型则能够实现.

总体来说, 当风格迁移准确率增加时, 内容保留度会有所下降, 可能的原因是隐式方法解缠时丢失了部分的内容信息, 而显式 TST 方法在保证较高迁移准确率的同时, 具有较好的文本内容保留度. 这从其做法上容易理解, 即仅替换了源属性相关的关键字, 可以更好地保留源句子的文本内容. 另外, 数据集的不同也会对结果产生一定的影响, 可能的原因是和数据集中的句子质量有所关联.

4 结 论

文本风格迁移是一个具有挑战性的新兴课题, 受到了学术界和工业界的广泛关注, 具有重要的研究意义和广阔的应用前景. 本文对文本风格迁移任务及其挑战进行了详细的介绍, 分类梳理并总结了各方法的基本原理、优缺点和性能. 并对目前的主流模型进行了对比实验, 分析了它们在 3 个自动化评价指标上的表现.

尽管目前该领域已经出现了一些有效的方法, 但任务类型比较固定, 以情感属性转换、语言正式性转换等为

主, 从实验结果来看离实际的应用还有较大距离. 除了 Jin 等人[67]所提到的将 TST 融入到更多的 NLP 任务中和更具体的实际生活中等, 本文也总结了以下几个未来可能的研究方向.

(1) 更好地分离内容与属性

如何在隐空间中更好地分离内容和属性依然是需要重点研究的问题. 除了改进基于 GAN 的解缠方法之外, 采用如对比学习[68,69]等自监督学习方法或者引入附加额外的信息, 根据具体的上下文情景[70]来获得更易分离的潜在表示, 值得进一步的研究.

(2) 将预训练模型应用在 TST 上

使用预训练模型处理 NLP 任务是目前非常热门的研究方向, 其体现了迁移学习的概念, 本质是在一个数据集上训练模型, 然后使该模型能够适应不同的数据集以执行不同的 NLP 操作. 在 TST 任务中, 最近出现了少量采用预训练方的法, 如 Sudhakar 等人[60]利用预训练模型 Transformer 的强大特征提取能力来更好地删除属性词, Duan 等人[71]采用预训练和插件 VAE 的方式, 将文本生成模块与条件表示模块解耦, 允许进行“一对多”条件风格转换. 如何更好地运用预训练模型和预训练方法到 TST 任务中, 也是未来值得探索的问题.

(3) 超越两种风格之间的转换

大多 TST 方法专注于文本在两种风格之间的转换. 未来 TST 研究应该探索二元风格迁移之外更丰富的任务. 例如, Lai 等人[72]提出了一种多属性 TST 任务, 通过指定多个风格属性 (例如作者的情感、性别等) 来迁移文本. Goyal 等人[73]在通用语料库上预训练基于 Transformer 的语言模型来初始化编码器-解码器, 然后以多种风格语言模型作为鉴别器来增强其对多个目标风格维度的转换能力.

(4) 其他语种的风格迁移任务

现有的 TST 模型大多应用于英语语料库. 然而, 不同的语种可能有其独特的文本样式属性. Mizukami 等人[74]设计了一个对话系统, 使用基于统计机器翻译的技术模拟日本作家的文本风格. 在非英语语言环境下, TST 方法需要捕获特定语种的文体属性, 这一类的研究可以提高我们对文本风格和风格表示的理解.

(5) 设计新的自动评价指标

现有的评估方法有一定的局限性. 例如迁移准确率的评估通常受限于属性分类器的性能. 此外, 实验结果表明风格转换强度与内容保留度往往成反比, 难以得到方法优劣与否的综合评价. 因此, 需要进一步探索新的综合性的自动评估指标.

致　谢　Fu 等人收集并持续性更新了近年来文本风格迁移领域相关的工作, 为本文的撰写提供很大帮助, 论文列表地址为: https://github.com/fuzhenxin/Style-Transfer-in-Text

References:

[1] Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2414–2423.

[2] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Cambridge: MIT Press, 2014. 2672–2680.

[3] Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2016. 2172–2180.

[4] Chen FJ, Zhu F, Wu QX, Hao YM, Wang ED, Cui YG. A survey about image generation with generative adversarial nets. Chinese Journal of Computers, 2021, 44(2): 347–369 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2021.00347]

[5] Hu ZT, Yang ZC, Liang XD, Salakhutdinov R, Xing EP. Toward controlled generation of text. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR.org, 2017. 1587–1596.

[6] Rao S, Tetreault J. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans: Association for Computational Linguistics, 2018. 129–140. [doi: 10.18653/v1/N18-

1012]

[7] Yi XY, Sun MS, Li RY, Li WH. Automatic poetry generation with mutual reinforcement learning. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 3143–3153. [doi: 10.18653/v1/D18-1353]

[8] Zhou H, Huang ML, Zhang TY, Zhu XY, Liu B. Emotional chatting machine: Emotional conversation generation with internal and external memory. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 730–739.

[9] Jin D, Jin ZJ, Zhou JT, Orii L, Szolovits P. Hooks in the headline: Learning to generate headlines with controlled styles. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 5082–5093. [doi: 10.18653/v1/2020.acl-main.456]

[10] Dos Santos CN, Melnyk I, Padhi I. Fighting offensive language on social media with unsupervised text style transfer. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne: ACL, 2018. 189–194. [doi: 10.18653/v1/P18-2031]

[11] Laugier L, Pavlopoulos J, Sorensen J, Dixon L. Civil rephrases of toxic texts with self-supervised transformers. In: Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, 2021. 1442–1461. [doi: 10.18653/v1/2021.eacl-main.124]

[12] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Cambridge: MIT Press, 2014. 3104–3112.

[13] Shen TX, Lei T, Barzilay R, Jaakkola T. Style transfer from non-parallel text by cross-alignment. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017. 6833–6844.

[14] Yang ZC, Hu ZT, Dyer C, Xing EP, Berg-Kirkpatrick T. Unsupervised text style transfer using language models as discriminators. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2018. 7298–7309.

[15] Zhao JB, Kim Y, Zhang K, Rush A, LeCun Y. Adversarially regularized autoencoders. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 5897–5906.

[16] Fu ZX, Tan XY, Peng NY, Zhao DY, Yan R. Style Transfer in text: Exploration and evaluation. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2018. 663–670.

[17] Luo FL, Li P, Zhou J, Yang PC, Chang BB, Sun X, Sui ZF. A dual reinforcement learning framework for unsupervised text style transfer. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: IJCAI. org, 2019. 5116–5122. [doi: 10.24963/ijcai.2019/711]

[18] Gong HY, Bhat S, Wu LF, Xiong JJ, Hwu WM. Reinforcement learning based text style transfer without parallel training corpus. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 3168–3180. [doi: 10.18653/v1/N19-1320]

[19] Prabhumoye S, Tsvetkov Y, Salakhutdinov R, Black AW. Style transfer through back-translation. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Melbourne: Association for Computational Linguistics, 2018. 866–876. [doi: 10.18653/v1/P18-1080]

[20] Li JC, Jia RB, He H, Liang P. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers). New Orleans: Association for Computational Linguistics, 2018. 1865–1874. [doi: 10.18653/v1/N18-1169]

[21] Xu W, Ritter A, Dolan B, Grishman R, Cherry C. Paraphrasing for style. In: Proc. of the COLING 2012. Mumbai: The COLING 2012 Organizing Committee, 2012. 2899–2914.

[22] Xu W. Data-driven approaches for paraphrasing across language variations [Ph.D. Thesis]. New York: New York University, 2014.

[23] Jhamtani H, Gangal V, Hovy E, Nyberg E. Shakespearizing modern language using copy-enriched sequence to sequence models. In: Proc. of the 2017 Workshop on Stylistic Variation. Copenhagen: Association for Computational Linguistics, 2017. 10–19. [doi: 10.18653/v1/W17-4902]

[24] Wang YL, Wu Y, Mou LL, Li ZJ, Chao WH. Harnessing pre-trained neural networks with rules for formality style transfer. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 3573–3578. [doi: 10.18653/v1/D19-1365]

[25] Sancheti A, Krishna K, Srinivasan BV, Natarajan A. Reinforced rewards framework for text style transfer. In: Proc. of the 42nd European Conf. on Information Retrieval. Lisbon: Springer, 2020. 545–560.

[26] Marcus MP, Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: The Penn treebank. Computational Linguistics, 1993, 19(2): 313–330.

[27] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017. 6000–6010.

[28] Merity S, Xiong CM, Bradbury J, Socher R. Pointer sentinel mixture model. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[29] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

[30] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog, 2019, 1(8): 9.

[31] Liu JW, Liu Y, Luo XL. Semi-Supervised learning methods. Journal of Computers, 2015, 38(8): 1592–1617 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2015.01592]

[32] Shang MY, Li PJ, Fu ZX, Bing LD, Zhao DY, Shi SM, Yan R. Semi-supervised text style transfer: Cross projection in latent space. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 4937–4946. [doi: 10.18653/v1/D19-1499]

[33] Zhang Y, Ge T, Sun X. Parallel data augmentation for formality style transfer. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 3221–3228. [doi: 10.18653/v1/2020.acl-main.294]

[34] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507. [doi: 10.1126/science.1127647]

[35] Kingma DP, Welling M. Auto-encoding variational Bayes. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff: ICLR, 2014.

[36] Diao QM, Qiu MH, Wu CY, Smola AJ, Jiang J, Wang C. Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2014. 193–202. [doi: 10.1145/2623330.2623758]

[37] Tolstikhin I, Bousquet O, Gelly S, Schölkopf B. Wasserstein auto-encoders. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[38] Yi XY, Liu ZH, Li WH, Sun MS. Text style transfer via learning style instance supported latent space. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. IJCAI.org, 2020. 3801–3807 [doi: 10.24963/ijcai.2020/526]

[39] Wang K, Hua H, Wan XJ. Controllable unsupervised text attribute transfer via editing entangled latent representation. In: Proc. of the Advances in Neural Information Processing Systems 32: Annual Conf. on Neural Information Processing Systems 2019. Vancouver: NeurIPS, 2019. 11034–11044.

[40] Liu DH, Fu J, Zhang YD, Pal C, Lv JC. Revision in continuous space: Unsupervised text style transfer without adversarial learning. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2020. 8376–8383.

[41] John V, Mou LL, Bahuleyan H, Vechtomova O. Disentangled representation learning for non-parallel text style transfer. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 424–434. [doi: 10.18653/v1/P19-1041]

[42] Li H. Statistical Learning Methods. 2nd ed., Beijing: Tsinghua University Press, 2019 (in Chinese).

[43] Kusner MJ, Sun Y, Kolkin NI, Weinberger KQ. From word embeddings to document distances. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: JMLR.org, 2015. 957–966.

[44] Wu LF, Yen IEH, Xu K, Xu FL, Balakrishnan A, Chen PY, Witbrock MJ. Word mover's embedding: From word2vec to document embedding. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 4524–4534. [doi: 10.18653/v1/D18-1482]

[45] Ganitkevitch J, Callison-Burch C. The multilingual paraphrase database. In: Proc. of the 9th Int'l Conf. on Language Resources and Evaluation (LREC2014). Reykjavik: European Language Resources Association (ELRA), 2014. 4276–4283.

[46] Ganitkevitch J, Callison-Burch C, Napoles C, Van Durme B. Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In: Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing. Edinburgh: Association for Computational Linguistics, 2011. 1168–1179.

[47] Liao Y, Bing LD, Li PJ, Shi SM, Lam W, Zhang T. QuaSE: Sequence editing under quantifiable guidance. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 3855–3864. [doi: 10.18653/v1/D18-1420]

[48] Zhang ZR, Ren S, Liu SJ, Wang JY, Chen P, Li M, Chen EH. Style transfer as unsupervised machine translation. arXiv:1808.07894v1, 2018.

[49] Jin ZJ, Jin D, Mueller J, Matthews N, Santus E. IMAT: Unsupervised text attribute transfer via iterative matching and translation. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 3097–3109. [doi: 10.18653/v1/D19-1306]

[50] Li DQ, Zhang YZ, Gan Z, Cheng Y, Brockett C, Dolan B, Sun MT. Domain adaptive text style transfer. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 3304–3313. [doi: 10.18653/v1/D19-1325]

[51] He JX, Wang XY, Neubig G, Berg-Kirkpatrick T. A probabilistic formulation of unsupervised text style transfer. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[52] Li JJ, Li ZC, Mou LL, Jiang X, Lyu MR, King I. Unsupervised text generation by learning from search. In: Proc. of the Advances in Neural Information Processing Systems 33: Annual Conf. on Neural Information Processing Systems 2020. 2020.

[53] Liu XG, Mou LL, Meng FD, Zhou H, Zhou J, Song S. Unsupervised paraphrasing by simulated annealing. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 302–312. [doi: 10.18653/v1/2020.acl-main.28]

[54] Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent models of visual attention. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Cambridge: MIT Press, 2014. 2204–2212.

[55] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[56] Xu JJ, Sun X, Zeng Q, Zhang XD, Ren XC, Wang HF, Li WJ. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: Association for Computational Linguistics, 2018. 979–988. [doi: 10.18653/v1/P18-1090]

[57] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002. 311–318. [doi: 10.3115/1073083.1073135]

[58] Zhang Y, Xu JJ, Yang PC, Sun X. Learning sentiment memories for sentiment modification without parallel data. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 1103–1108. [doi: 10.18653/v1/D18-1138]

[59] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[60] Sudhakar A, Upadhyay B, Maheswaran A. "Transforming" delete, retrieve, generate approach for controlled text style transfer. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 3269–3279. [doi: 10.18653/v1/D19-1322]

[61] Malmi E, Severyn A, Rothe S. Unsupervised text style transfer with padded masked language models. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 8671–8680. [doi: 10.18653/v1/2020.emnlp-main.699]

[62] Wu X, Zhang T, Zang LJ, Han JZ, Hu SL. Mask and Infill: Applying masked language model for sentiment transfer. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: IJCAI.org, 2019. 5271–5277. [doi: 10.24963/ijcai.2019/732]

[63] Hu ZQ, Lee RKW, Aggarwal CC, Zhang A. Text style transfer: A review and experimental evaluation. arXiv:2010.12742, 2020

[64] Mir R, Felbo B, Obradovich N, Rahwan I. Evaluating style transfer for text. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 495–504. [doi: 10.18653/v1/N19-1049]

[65] Kim Y. Convolutional neural networks for sentence classification. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014. 1746–1751. [doi: 10.3115/v1/D14-1181]

[66] Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. In: Proc. of the 15th Conf. of the European Chapter of the Association for Computational Linguistics (Vol. 2, Short Papers). Valencia: Association for Computational Linguistics, 2017. 427–431.

[67] Jin D, Jin ZJ, Hu ZT, Vechtomova O, Mihalcea R. Deep learning for text style transfer: A survey. arXiv:2011.00416, 2021.

[68] van den Oord A, Li YZ, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.

[69] He KM, Fan HQ, Wu YX, Xie SN, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proc. of the 2020

IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 9726–9735. [doi: 10.1109/CVPR42600.2020.00975]

[70] Cheng Y, Gan Z, Zhang YZ, Elachqar O, Li DQ, Liu JJ. Contextual text style transfer. In: Proc. of the 2020 Findings of the Association for Computational Linguistics (EMNLP 2020). Association for Computational Linguistics, 2020. 2915–2924. [doi: 10.18653/v1/2020.findings-emnlp.263]

[71] Duan Y, Xu CW, Pei JX, Han JL, Li CL. Pre-train and Plug-in: Flexible conditional text generation with variational auto-encoders. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 253–262. [doi: 10.18653/v1/2020.acl-main.23]

[72] Lai CT, Hong YT, Chen HY, Lu CJ, Lin SD. Multiple text style transfer by using word-level conditional generative adversarial network with two-phase training. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 3579–3584. [doi: 10.18653/v1/D19-1366]

[73] Goyal N, Srinivasan BV, Anandhavelu N, Sancheti A. Multi-style transfer with discriminative feedback on disjoint corpus. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021. 3500–3510. [doi: 10.18653/v1/2021.naacl-main.275]

[74] Mizukami M, Neubig G, Sakti S, Toda T, Nakamura S. Linguistic individuality transformation for spoken language. In: Lee GG, Kim HK, Jeong M, Kim JH, eds. Natural Language Dialog Systems and Intelligent Assistants. Cham: Springer, 2015. 129–143. [doi: 10.1007/978-3-319-19291-8_13]

附中文参考文献:

[4] 陈佛计, 朱枫, 吴清潇, 郝颖明, 王恩德, 崔芸阁. 生成对抗网络及其在图像生成中的应用研究综述. 计算机学报, 2021, 44(2): 347–369. [doi: 10.11897/SP.J.1016.2021.00347]

[31] 刘建伟, 刘媛, 罗雄麟. 半监督学习方法. 计算机学报, 2015, 38(8): 1592–1617. [doi: 10.11897/SP.J.1016.2015.01592]

[42] 李航. 统计学习方法. 第2版, 北京: 清华大学出版社, 2019.

陈可佳(1980－), 女, 博士, 副教授, CCF 专业会员, 主要研究领域为机器学习, 图数据挖掘.

陈景强(1983－), 男, 博士, 副教授, 主要研究领域为自动摘要, 自然语言处理.

费子阳(1996－), 男, 硕士生, 主要研究领域为文本风格迁移, 自然语言处理.

杨子农(1997－), 男, 硕士生, 主要研究领域为机器翻译, 自然语言处理.