

文本风格转换综述

贾熹滨¹, 刘思良¹, 胡长建², 李 让²

(1. 北京工业大学信息学部, 北京 100124; 2. 联想集团北京研究院, 北京 100085)

摘 要: 自然语言生成是人工智能系统中的重要组成部分,随着人工智能技术逐渐融入人们的生活,人们对自然语言生成技术有了更高的要求,可控自然语言生成成为新的研究热点. 文本风格转换任务作为可控自然语言生成的重要研究方向,受到学术界和工业界的广泛关注. 文本风格转换任务指的是在保留文本原有内容的基础上,有目的地生成目标风格的句子. 为了进一步展示该任务中的发展进程,首先,介绍了当前文本风格转换任务的定义及发展脉络;然后,基于模型方法将当前的研究工作总结成3类,即基于隐空间解耦、基于显性解耦和基于一映射,在简述这些工作的基础上,着重介绍了一些具有启发意义的工作;接着,梳理和分析了文本风格转换任务中的公共数据集和评估方法;最后,对文本风格转换任务的未来发展趋势提出展望.

关键词: 自然语言生成; 人工智能; 文本风格转换; 显性解耦; 隐空间解耦; 一步映射

中图分类号: U 461; TP 308

文献标志码: A

文章编号: 0254-0037(2022)04-0443-14

doi: 10.11936/bjutxb2020120008

Text Style Transfer: A Survey

JIA Xibin¹, LIU Siliang¹, HU Changjian², LI Rang²

(1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;

2. Lenovo Corporate Beijing Research & Development, Beijing 100085, China)

Abstract: Natural language generation is an important part of artificial intelligence system. As artificial intelligence technology is gradually integrated into human lives, people have higher requirements for natural language generation technology, and controllable natural language generation has become a new research hotspot. The task of text style transfer, as an important research direction of controllable natural language generation, has been widely concerned by academia and industry. The task of text style transfer refers to the purposeful generation of target-style sentences on the basis of retaining the original content of the text. To further demonstrate the development process of this task, the definition and development context of the current text style transfer task was first introduced. Then, based on the model method, the current research work was summarized into three categories: based on latent space disentangled, and based on explicit disentangled, based on one-step mapping, and on the basis of a brief description of these works, enlightening work was focused on. Furthermore, the common data sets and evaluation methods in the text style transfer task was combed and analyzed. Finally the future development trend of the text style transfer task was prospected.

Key words: natural language generation; artificial intelligence; text style transfer; explicit disentangled; latent space disentangled; one-step mapping

风格转换任务在人工智能许多子领域都有着广泛的研究,包括计算机视觉^[1-2]领域和自然语言处理^[3-15]领域。

风格转换最早的研究是在计算机视觉领域中,Gatys等^[1]认为艺术画作是由画家在内容和风格中构成复杂的图像来创造独特的视觉体验,人工智能系统无法具备同样的能力。受到计算机视觉领域的其他任务的启发,深度神经网络的视觉模型在某些计算机视觉的任务中可以达到接近于人类的表现。借助深度神经网络对图片提取内容与风格信息,再通过内容重构和风格重构可实现在艺术画作与真实照片之间的图像风格转换。

在自然语言处理领域中,风格转换的研究起步较晚。文本风格相较图像风格定义更为宽泛,例如:文章和文章的摘要^[16]、礼貌的表达和粗鲁的表达^[17]、积极的表达和消极的表达^[7]都可以视为2种不同的风格。实现方法可以大体分为解耦与基于一映射两大类。早期的文本风格转换任务受图像风格转换任务相关研究的启发,将研究的重点放在如何解耦文本中的风格与内容。近几年也有一些研究者认为文本的内容和风格特征的解耦是非必须的,提出让模型直接学习不同文本风格间的映射关系,实现文本的风格转换,推动文本风格转换研究进一步发展。另外,随着表示学习突飞猛进的发展,预训练语言模型在各项自然语言处理的任务上取得了巨大的成功^[18]。基于预训练语言模型的文本风格转换方法^[10]为文本风格转换任务提供了新的研究思路。

本文梳理了近几年文本风格转换算法的研究工作,首先,从文本风格转换任务的数据集入手,按照是否具有平行语料,介绍平行数据集和非平行数据集;然后,依次介绍平行和非平行数据集上的相关工作;最后,展望文本风格转换任务未来的发展,尝试为研究人员建立一个比较完整的研究领域视图,为相关领域的研究者提供帮助。

1 文本风格转换任务概述

自然语言生成作为人工智能领域一项重要的技术,已经广泛应用于各类实际场景中,包括机器翻译、对话生成等。以往的自然语言生成模型往往存在一些缺陷,无法根据实际需求生成文本,距离真正的人工智能还有着一定的差距。寻找可控自然语言生成方法成为近年来的研究重点。

可控自然语言生成是指模型能根据特定的风格

和内容生成文本。风格可控包括控制句子的长度^[19]、情感^[20]、时态^[20]等;内容可控包括基于对话中的角色信息^[21]、维基百科中的常识信息^[22]等控制文本的生成。文本风格转换任务作为可控自然语言生成的子任务,是一种在保留句子内容信息完整的基础上控制句子的风格,生成具有指定风格的句子的任务。通常指定的风格和句子原有风格是对立的。

文本风格转换在实际场景中有着广泛的应用。在智能客服场景中,根据用户的情绪生成具有相关情感回复能有效提高用户的体验。另外,在书面语言的修改及纠错的场景中,文本风格转换也发挥着重要作用。模型将书面语言和口语化语言作为2种相互转换的风格,通过学习二者间的不同,为用户提供书面语言的纠错或改写的服务。

然而,文本风格转换任务的发展过程是曲折的,仍存在着三大问题阻碍文本风格转换任务的发展。

其一,缺乏平行的数据集。传统的文本生成模型是在平行数据集上训练的,各类任务在有监督的训练中取得了巨大的成功^[23-24]。由于文本风格转换任务本身的复杂性,很难构建合适的平行数据集。以Yelp^[3]数据集为例,它是文本风格转换最早的主流数据集。数据源于用户在Yelp平台上对饭店服务的评价。Yelp数据集将评价分为积极和消极2种。每一条评价对应用户一次消费体验,很难找到对立风格中与之相对应的句子构建平行语料。

其二,无法解决梯度回传问题。传统的风格转换任务的训练可分为2个阶段:先由生成器生成目标风格的样本,再由判别器对生成的样本计算风格损失,通过反向传播,回传判别损失,实现风格转换。该模式在图像风格转换中取得了巨大的成功^[1-2]。在文本风格转换任务中,文本的表示是离散化的。在文本生成的过程中,每一个时间步都需要选择当前时间步最大概率的词作为该时间步的输出,导致梯度消失难以微分,无法通过反向传播实现梯度回传,需要借助其他方法解决梯度回传问题。

其三,缺乏全面的自动评价指标。在文本风格转换任务中,评价指标包括风格转换的准确率、内容的保存程度及流利度3个方面。在衡量模型的优劣时,会将人工评价和自动评价的结果共同作为评价的依据。人工评价是指找一批以数据集中的语言作为母语或对该语言有一定能力的人,从3个方面为生成样本打分,能够准确地测试出模型各指标的表现差异。因为人工评价存在成本高、主观性强、难以

复制的缺点,所以在衡量不同模型间的差异时,还是以自动评价标准作为主要参考.自动评价指的是借助统计模型或神经网络模型分别从单一维度评价生成结果,计算成本低,评估速度快,可信度高.当前的自动评价方法都是只对单一维度评价,在采用自动评价指标为模型打分时,每个指标只能反映单个维度模型的表现,无法全面地评价模型的表现.多维度的评价指标将有助于文本风格转换任务的进一步发展.

为了解决这些问题,研究者们寻找了许多解决办法.针对缺乏平行语料,采用无监督的自然语言生成方法在非平行语料下训练模型.另外,一些研究者还针对某些特定风格,开发平行语料的数据集,有效地缓解了缺乏平行数据的问题.针对文本的离散表示,通常的解决办法有3种:第1种采用 Gumbel-Softmax^[25],通过重参数化采用连续分布近似文本的离散分布,避免离散化数据导致的无法梯度回传问题;第2种,避开每个时间步的采样,将每个时间步生成的词表分布作为该时间步生成的词向量,该方法也是文本风格转换任务中最常用的处理离散化数据的方法;第3种,采用强化学习方法,解决不可微分问题,通过奖惩机制引导模型的生成.针对缺乏自动评价指标的问题,Zhang等^[26]将 Bert^[18]作为自动评价模型,该模型能够综合文本内容与流利度这2个评价维度对样本打分,能够更全面地评价模型.

2 数据集

研究者根据不同的实际场景,选择一些风格特征明显的场景构建数据集,包括正面与负面、浪漫与人文、学术与新闻、男性与女性等.根据数据集中是否具有平行语料,将当前的数据集分为平行数据集和非平行数据集.

平行数据集的构成包括2种风格特征的文本数据,不同风格特征的文本数据具有一对一的映射关系,每一句话都可找到风格特征不同、内容特征相同的另一句,两句互为标签.在此类数据集中,模型可进行有监督训练.与之类似,非平行数据集包含2种不同风格特征的文本数据,但不同风格的文本数据之间没有对应关系.因此,在非平行数据集中的文本风格转换模型只能在无监督的条件下训练.本文在表1中列举了一些文本风格转换研究相关的数据集.

最早的平行数据集是由 Xu等^[30]在2012年提出的基于写作风格的平行数据集,将莎士比亚的写

作风格和现代写作风格当作2种对立的风格特征,构建平行数据集.之后 Carlson等^[34]在2018年提出基于《圣经》语言风格与现代语言风格的数据集.在统计与比较多个版本的《圣经》后,总结不同版本《圣经》的语言风格,在《圣经》的语言风格与现代语言风格之间构建平行数据集. Rao等^[31]在2018年提出基于书面表达的平行数据集.针对人们的书写习惯,将书面形式的表达和非书面形式的表达作为2种不同的风格,模型可以在该数据集上学习和训练,为非母语者提供写作帮助及为儿童教育提供帮助,具有广泛的应用前景.

非平行数据集中研究最多的是 Yelp 和 Amazon 数据集. Yelp 数据集是2017年 Yelp 挑战赛中由官方提供,通过众包的方式选择数据集中风格比较明显的文本构建的积极与消极的数据集. Yelp 是国外类似于大众点评的应用,顾客会在 Yelp 上对餐厅打分. Amazon 数据集原本是在文本分类任务中应用的数据,在2018年, Li等^[8]将该数据集率先应用在文本风格转换任务中. Amazon 数据集是源自于亚马逊网购平台的产品评价,包括积极的评价和消极的评价.

从表1可以发现,平行数据集的数据量通常小于非平行数据集的数据量,规模基本小1个数量级.这是因为平行数据集的构建需要大量的人工标注.相对来说,非平行数据集更容易获取.

在文本风格转换任务的研究中,缺乏平行数据集的问题一直阻碍着文本风格转换研究的发展.因此,当前文本风格转换的研究主要集中在非平行数据集.

3 基于平行数据集的文本风格转换方法

在文本风格转换任务的方法中,大多数基于平行数据集的研究都是将文本风格转换任务当作标准的序列到序列(Seq2Seq)的文本生成任务,具体做法是将2种不同风格特征、相同语义信息的数据集类比为机器翻译任务中成对的翻译语料,参照机器翻译生成方法实现文本的风格转换.

Jhamtani等^[35]的工作是较早出现的基于平行数据集的文本风格转换研究.该工作在 Xu等^[30]提出的莎士比亚数据集上,参考生成式文本摘要的复制机制^[36],提出在标准的序列到序列模型中增加指针网络^[37],有效提升了文本风格转换效果.

在 Rao等^[31]提出基于书写正式的数据集后,大量基于平行数据集文本风格转换的研究都集中

表 1 文本风格转换数据集
Table 1 Text style transfer datasets

数据集名称	语言	来源	类别	数据集大小/ 10^3	平行/非平行
Yelp	英文	Hu 等 ^[27]	积极/消极	444/63/126	非平行
IMDb	英文	Maas 等 ^[28]	积极/消极	366/4/2	非平行
Amazon	英文	Li 等 ^[8]	积极/消极	554/2/1	非平行
Caption	英文	Li 等 ^[8]	浪漫/幽默	12/ /1	非平行
Paper	英文	Fu 等 ^[3]	学术/新闻	392/20/20	非平行
Gender	英文	Prabhumoye 等 ^[4]	男性/女性	2 000/4/534	非平行
Politics	英文	Prabhumoye 等 ^[4]	民主党/共和党	537/4/56	非平行
Reddit	英文	Dos 等 ^[17]	礼貌/粗鲁	10 000/19/47	非平行
Twitter	英文	Dos 等 ^[17]	礼貌/粗鲁	3 000/18/18	非平行
Word substitution decipherment	英文	Dou 等 ^[29]	密文/原文	200/100/100	非平行
Authorship	英文	Xu 等 ^[30]	现代风格/莎士比亚风格	18/ /1. 4	非平行
GYAFC	英文	Rao 等 ^[31]	书面语言/非书面语言 (娱乐与音乐)	52/5/2	平行
GYAFC	英文	Rao 等 ^[31]	书面语言/非书面语言 (家庭与友谊)	51/5/2	平行
TCFC	英文	Wu 等 ^[32]	推特/书面语言	1 000/0. 9/0. 9	平行
MTFC	中文	Wu 等 ^[32]	书面语言/非书面语言	1 000/2. 8/1. 4	平行
Enron-Context	英文	Cheng 等 ^[33]	书面语言/非书面语言	13/0. 5/1	平行
Reddit-Context	英文	Cheng 等 ^[33]	冒犯/非冒犯	22/0. 5/1	平行

在该数据集. Niu 等^[38]在 2018 年将多语言引入文本风格转换任务中,提出借助书写形式敏感的机器翻译模型在英语与法语翻译的过程中加入风格控制. 采用多任务的训练方式,一方面约束机器翻译模型的翻译能力,另一方面约束生成的文本具有指定的风格,通过学习不同语言中包含的共同的语言知识有效地提升文本风格转换的质量. 这也是自然语言生成领域中常见的提升模型生成质量的方式. 与 Niu 等^[38]的研究工作相同,Xu 等^[39]同样采用了多任务的思想,通过多个损失函数训练模型在保留足够多语义内容的基础上,实现文本风格的转换.

随着 GPT2^[40]在自然语言生成任务中显示出的卓越性能,Wang 等^[41]将 GPT2 模型应用于文本风格转换任务,与以往的研究不同,该项工作主要关注的是非正式文本到正式文本的改写. 首先采用基于规则的方式处理非正式的文本,得到基于规则的正式文本,再将非正式的文本和基于规则的结果作为 GPT2 模型的输入,通过解码器解码得到生成结果.

Zhang 等^[42]认为书写正式数据集的规模较小,阻碍了书写纠错这一文本风格转换任务的发展,因此,在 Wang 等^[41]的工作基础上,提出对现有数据集进行数据增强,通过多个现有的模型在数据增强数据集训练后在多个评价指标上的提升,证明了该数据增强手段的有效性,同时在该数据集上提出新的方法.

4 基于非平行数据集的文本风格转换方法

基于非平行数据集的文本风格转换方法,主要采用基于无监督文本生成方法. 整体的研究思路可以分为两大类.

第 1 类研究思路认为句子应当包含语义内容和文本风格两部分,文本风格转换任务实质上就是二者分离、重组,即先分离语义内容和文本风格,保留句子中的语义内容,修改句子的文本风格,再重新组合. 如何成功将两部分完整分离,实现语义内容与风格特征的解耦,就成为需要主要解决的问题. 沿着语义内容与风格特征解耦的研究思路,一些研究人员认为可

在提取文本特征时采取一些手段解耦,从而获取风格独立的语义内容. 文本特征分布的空间称为隐空间,因此,这类方法称为基于隐空间解耦的方法. 另外一些研究者认为可利用句子本身的构成特点及人的语言习惯,通过对句子中词的修改实现内容特征与风格特征的解耦. 例如,在“这里做的菜很难吃”“这里做的菜很棒”这两句话中具有相同的语义内容,而表达的风格却是相反的. 决定整个句子风格倾向的就是“难吃”和“棒”这2个词. 将决定句子风格倾向的词汇与句子中剩余部分分离,就能实现内容特征与风格特征的分离. 因为是通过分割句子实现的解耦,类比隐空间解耦的方法,所以此类方法称为显性解耦的方法.

第2类研究思路认为句子中的内容特征和风格特征是相互包含的,无法彻底分离. 模型可以在不同风格的数据间学习映射关系,避免对语义与风格解耦,实现文本风格的转换. 因此,这类文本风格转换方法称为基于一部映射的方法.

4.1 基于隐空间解耦的方法

在基于隐空间解耦的文本风格转换的工作中,主要解决2个问题:一是如何从文本特征中分离出与风格无关的内容特征,二是如何控制生成句子的风格. 在这些工作中,第1个问题的解决方法包括对抗学习、基于机器翻译模型、隐空间分离等方法;第2个问题的解决方法可分为基于风格属性码和基于隐空间分裂的风格控制方法. 下面根据风格控制方法的差异,依次介绍基于隐空间解耦的文本风格

转换研究.

4.1.1 基于风格属性码的风格控制方法

此类方法的实现过程是采用额外的结构化的风格属性码控制生成文本的风格. 在实现内容特征与风格特征解耦后,将风格独立的内容特征的隐空间特征与目标风格的风格属性码拼接作为解码器的输入,或者采用不同的解码器生成对应风格的句子.

Fu 等^[3]采用循环神经网络(recurrent neural network, RNN)提取文本特征,将序列到序列模型作为基本模型,引入多任务框架与对抗学习的思想实现风格与内容的解耦. 模型结构如图1所示. 在模型训练过程中,采用迭代的方式逐步实现解耦,增加多层感知机作为风格特征判别器,对文本特征进行分类. 如果判别器能正确分类,说明文本特征仍然包含风格信息. 为了进一步分离内容与风格,可将判别器和特征提取器进行对抗训练. 首先,训练特征判别器,固定编码器参数,更新判别器参数;然后,使判别器能够准确判断文本特征中是否还包含着风格,再固定判别器参数,更新编码器,降低判别器分类的准确率,减少文本特征中包含的风格信息. 通过重复这2步,文本特征中的风格信息不断减少,实现内容与风格的解耦. 在风格的控制方法上,该工作提供了2种思路:方案一为多解码器模型,采用的是初始化状态的方法控制风格,为每个不同的风格都分配了对应的解码器;方案二为风格向量模型通过风格向量进行控制,将目标的风格向量 s 与内容特征组合作为解码器的输入,控制生成样本的风格.

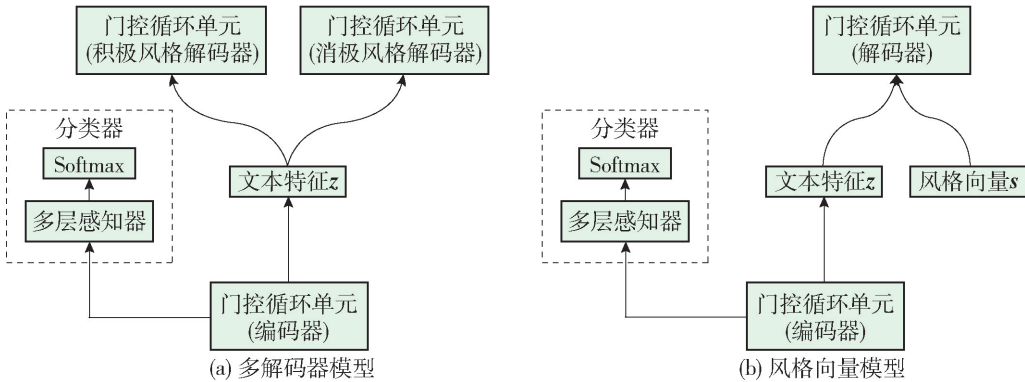


图1 多解码器模型和风格向量模型^[4]

Fig. 1 Multi-decoder model and style-embedding model^[4]

Prabhumoye 等^[4]提出借助神经机器翻译模型得到与风格无关的内容向量,模型结构如图2所示. 在模型训练前,先在具有中性风格的翻译语料中训练英译法和法译英神经机器翻译模型;将源风格的句子(英语)

通过预训练的英译法模型翻译成法语的句子作为法译英的模型的输入,通过法译英模型提取文本特征作为风格独立的内容特征. 作者认为2个翻译模型是在中性风格语料中训练的,可以减少2种风格语料对模型

的影响,从而学习到与风格无关的内容特征. 因此,通过法译英模型提取的文本特征就是风格独立的内容特

征. 在风格的控制上,采用了初始化状态的方法,通过对应风格的解码器控制目标风格样本的生成.

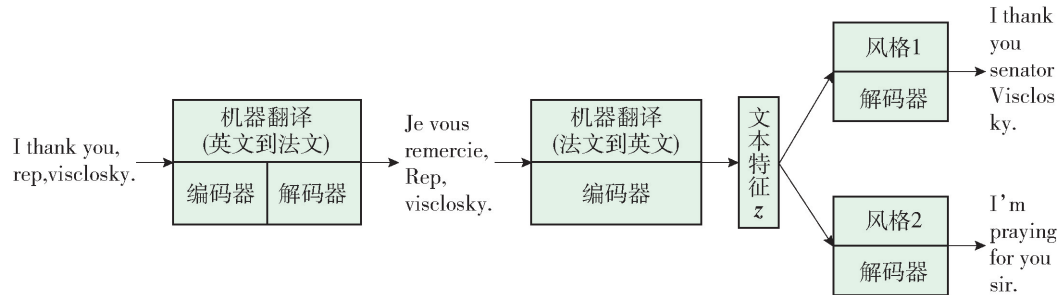


图2 基于机器翻译解耦的文本风格转换模型

Fig. 2 Text style transfer model based on machine translation disentangle

Dai 等^[6]的模型结构如图 3 所示. 这项工作的思想源于图像风格转换 Cycle-GAN^[2], 将转换风格后的样本重新转换为原来的风格, 计算与原有样本的差异, 从而保证转换过程中信息没有丢失. 该模型的训练分为 2 步: 1) 对文本建模, 学习提取文本特征. 与一般的自编码器不同, 将文本特征与随机初始化的风格向量拼接作为解码器的输入, 计算重构损失 L_{self} ; 2) 控制风格实现风格转换. 采用 2 个损失函数 L_{style} 、 L_{cycle} 控制句子风格的转换和转换过程中内容的保留. 在训练中, 将文本特征与目标风格向量拼接生成目标风格的句子, 再利用判别器判断生成样本的风格, 计算判别风格与实际的目标风格之间的差异 L_{style} , 引导生成文本的风格向目标风格靠近. 为了保证转换过程中语义信息不丢失, 将生成的文本与原有的风格向量结合重新生成原有的句子, 计算重新生成的句子与输入句子的差异 L_{cycle} . 另外, 还采用 Transformer 代替原有的 RNN 作为编解码器, 有效地提升了文本风格转换

的生成质量.

4.1.2 基于隐空间分裂的风格控制方法

此类方法认为同一风格的数据集具有相同的风格特征, 可对相同风格的句子提取统一风格特征, 将文本特征的隐空间表示分裂, 一部分代表风格独立的内容特征, 另一部分代表风格特征. 然后, 再通过对风格特征的替换和文本特征的重组, 控制文本生成的风格.

John 等^[7]对隐空间解耦做了进一步的改进和提升, 提出了 Disentangled 模型. 将内容特征与风格特征视为 2 个独立的隐空间分布, 并将文本特征投影到 2 个隐空间, 分别代表内容特征和风格特征. 模型结构如图 4 所示.

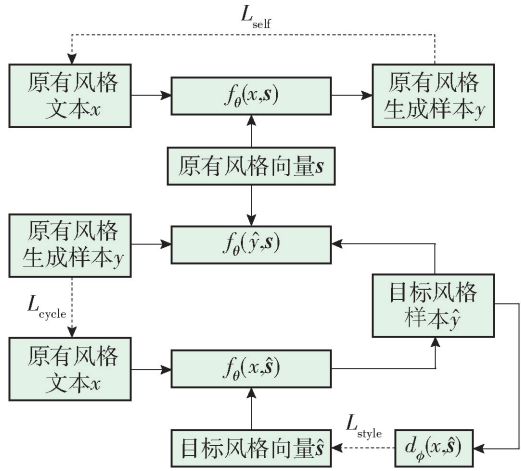


图3 Style Transformer 模型

Fig. 3 Style Transformer model

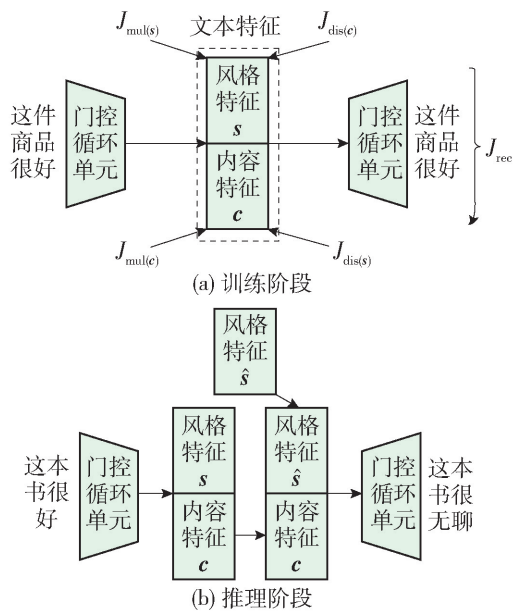


图4 Disentangled 模型

Fig. 4 Disentangled model

在训练阶段, 将文本特征分别投影到不同的隐

空间中,得到风格特征 s 和内容特征 c . 然后,为了保证 2 个隐空间相互独立,设计 2 个判别器用于计算某个隐空间特征中包含的风格信息和内容信息. 针对内容隐空间特征,最小化内容损失 $J_{\text{mul}(c)}$,最大化风格损失 $J_{\text{dis}(s)}$. 针对风格隐空间,最小化风格损失 $J_{\text{dis}(c)}$,最大化内容损失 $J_{\text{mul}(s)}$,促使风格空间中包含更多的风格信息和更少的内容信息. 另外,计算隐空间内容损失采用的是词袋损失^[43]的方法. 将输入句子去除停用词和风格词,得到词袋特征,将词袋特征当作标签,再基于隐空间特征预测词袋特征. 如果隐空间包含的内容信息足够多,就能完整地预测出文本的词袋特征,同时去除输入文本的停用词和情感词也保证了得到的词袋特征没有包含风格信息. 计算隐空间内容损失的方法与 Fu 等^[3]的工作类似,在此就不做赘述.

在推理阶段,需要用目标风格的特征代替原有风格的特征,模型采用对大量相同风格句子的风格特征求和取平均的方法,作为该种风格的特征表示,替换原有风格特征,控制生成句子的风格.

4.2 基于显性解耦的方法

在基于显性解耦的研究工作中研究者将决定句子整体风格的词汇称作风格标记,通过改变句子中的风格标记实现句子的风格转换.

显性解耦的方法可以分为 3 步:第 1 步,删除风格标记,对数据进行预处理,通过删除句子中原有风格的标记获取风格独立语义内容的文本表示. 第 2 步,检索目标风格的风格标记,用于替换原有风格标记. 第 3 步,生成目标风格句子,将第 2 步检索到的目标风格标记填入原有风格标记的位置.

4.2.1 删除原有风格标记

Li 等^[8]提出一种简单易行的方法以找到句子中的风格标记. 在数据集中,计算每个词在 2 种风格数据集中出现的次数,其中风格标记与剩余词在不同数据集中出现的频率存在显著的差异,再根据这一差异区分出风格标记与剩余词. 具体公式为

$$s(u, v) = \frac{\text{count}(u, D_v) + \lambda}{\left(\sum_{v' \in V, v' \neq v} \text{count}(u, D_{v'}) \right) + \lambda} \quad (1)$$

式中:Count()代表词 u 在数据集 D_v 中出现的次数; λ 代表平滑稀疏,保证数据集 D_v 中没有出现的词也能分配一定的权重; $s(u, v)$ 代表词 u 在数据集 D_v 中对于风格 v 的贡献度,当 $s(u, v)$ 大于某个阈值时,将词 u 作为风格 v 的风格标记.

采用自注意力机制的风格分类器,为文本中的每个词分配权重,权重越高的词对句子的风格贡献度越大,将文本风格贡献度高的词作为风格标记,公式为

$$a = \text{Softmax}(\mathbf{w} \cdot \tanh(\mathbf{WH}^T)) \quad (2)$$

$$\mathbf{r} = \mathbf{a} \cdot \mathbf{H} \quad (3)$$

$$y = \text{Softmax}(\mathbf{W}' \cdot \mathbf{r}) \quad (4)$$

先为句子中的每个词分配权重 a ,由每一时间步的隐状态通过全连接和 Softmax 得到;再将权重与每一时间步的隐状态相乘,得到加权后的隐状态 c ;最后,通过 Softmax 得到句子每个词汇的贡献度,即句子的风格概率分布. 将所有权重大于阈值的词作为风格标记,其中将阈值设定为平均权重. Xu 等^[9]还将上述的 2 种方法融合,将词频与自注意力词权重相乘,更精确地找到句子中的风格标记. 另外, Wu 等^[10]认为长短期记忆(long short-term memory, LSTM)网络模型不够鲁棒,提取特征效果有待提高,并用实验证明删除风格标记过程中会删掉一些重要的上下文或遗漏一些风格标记. 因此,提出采用 Transformer^[44]选择删除对象,结果显示能够更准确地识别句子中风格标记.

Sudhakar 等^[11]提出一个判断方法,公式为

$$p_x = p_{\theta_c}(s|x) \quad (5)$$

$$p_{x, t_i} = p_{\theta_c}(s|x, t_i) \quad (6)$$

$$S_{t_i}^k = p_{x^k} - p_{x^k, t_i} \quad (7)$$

先通过预训练好的风格分类器(基于自注意力机制的分类器)计算句子 x 的风格 s 的概率 p_x . 再计算去掉词 t_i 后文本的风格概率 p_{x, t_i} ,将 2 个风格概率的差 S_{t_i} 作为删除词 t_i 的风格重要性. 之后按照重要性从大到小的顺序依次删除 k 个词汇得到剩余文本的风格概率 p_{x^k, t_i} ,计算 k 个词对句子风格的影响,直到整个文本的风格概率小于阈值 α . 最后,将删除的 k 个词作为该文本的句子. 为了避免无法完整删除风格标记情况的出现,设定了删除的风格标记应当大于阈值 β ,并详细论证了 2 个阈值的大小对删除风格标记的影响.

4.2.2 寻找目标风格标记

在删除输入文本中原有风格标记后,需要在另一种风格的数据集中寻找能够与之相对应的目标风格标记. 在此处将去除原有风格标记的剩余部分称为模板,一种常用的寻找目标风格标记的方法是用原有的风格数据集中的模板匹配目标风格数据集中的模板. 如果目标风格数据集中有一些模板与处理

后的模板的语义相似度高,二者对应风格标记也应是相互对应的. 为了将模板与目标风格数据集中其相似的模板匹配,一般都是通过句子向量的余弦相似度来寻找. 常用的计算句子向量的方式包括基于词频的 TF-IDF, 以及采用预训练词向量的 GloVe^[45], 通过对句子中所有词的词向量求和取平均后, 得到句子向量.

4.2.3 生成目标风格句子

生成目标风格句子的方法大致可以分为 2 类: 一类是直接将检索到的目标风格模板对应的风格标记填入原有目标风格的位置; 另一类是采用编解码器作为生成模型, 以原有模板和检索到的目标风格标记作为输入, 生成目标风格句子.

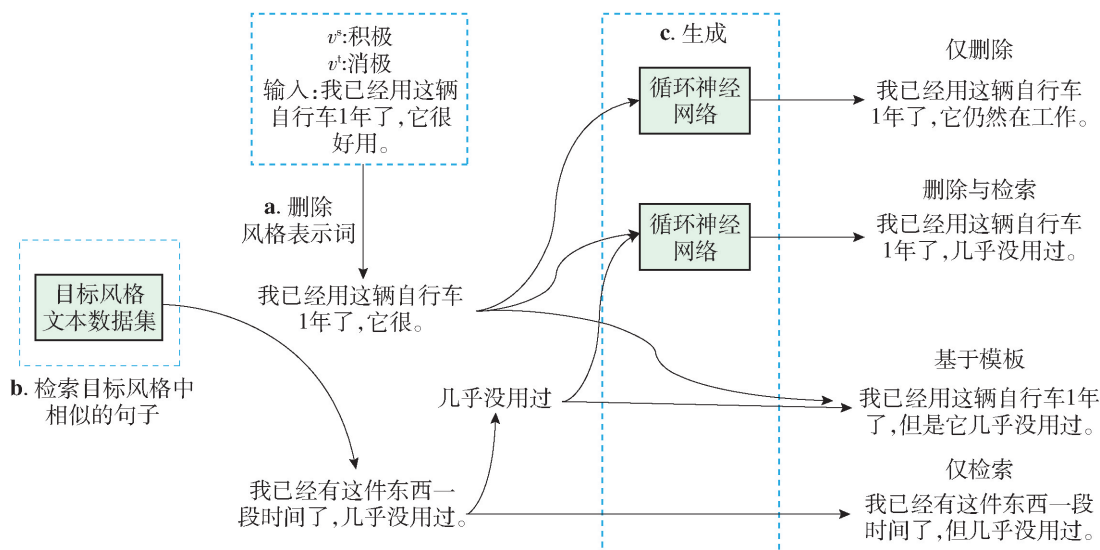


图 5 4 种文本风格转换方法^[8]

Fig. 5 Four methods in text style transfer^[8]

Wu 等^[10]借助预训练语言模型 Bert, 结合 Bert 的预训练任务掩盖式语言模型 (mask language model, MLM), 将原有风格标记视为需要掩盖的词, 目标风格标记视为需要预测的词. 以 Bert 作为基础模型, 借助预训练语言模型中的知识, 将检索与填充融合成一步, 提升了文本风格转换的质量.

4.3 基于一映射的方法

基于一映射的方法指的是采用端到端的生成方法, 将文本特征视为一个整体, 避免对文本中风格与内容特征的分离. 通过学习不同风格句子之间存在的映射关系, 实现风格转换. 具体实现可以分为 3 种方式: 基于机器翻译、基于强化学习和基于隐空间编辑的生成方法.

4.3.1 基于机器翻译的方法

Zhang 等^[12]受到无监督机器翻译研究的启发,

Li 等^[8]提出 4 种策略生成目标风格句, 按照生成方式的不同可以分为“基于模板”“仅检索”“仅删除”“删除与检索”, 如图 5 所示. 前 2 种方法属于第 1 类生成方法, 其中“基于模板”方法是将目标风格标记 $c(x^t, v^t)$ 替换输入文本中的原有风格标记 v^s ; “仅检索”方法是将对立风格数据集中模板间相似度最高的句子 x^t 直接作为输出结果. 后 2 种方法属于第 2 类生成方法. “仅删除”方法是直接将原有模板 $c(x, v^s)$ 作为编解码器的输入, 生成目标风格句子; “删除与检索”方法是在目标风格数据集中找到模板间相似度最高的模板, 将匹配到模板的目标风格标记 $c(x^t, v^t)$ 与原有模板 $c(x, v^s)$ 组合作为编解码输入, 生成目标风格句子.

将基于非平行语料数据集的文本风格转换任务视为无监督的机器翻译. 采用迭代回译的方式提升翻译质量, 先构建伪平行语料, 作为迭代回译的初始输入. 通过计算词向量的余弦相似度学习不同风格词之间的对应关系, 建立对应词表. 按照词表替换原文本中的词, 得到伪平行语料. 构建完成的伪平行语料作为平行语料, 由模型再生成另一种风格的句子, 再将回译生成的结果作为第 2 轮模型的输入. 在回译过程中通过语言模型与分类器引导迭代回译, 在迭代中不断提升风格转换质量.

Jin 等^[13]同样采用迭代回译的方法, 区别在于该模型通过检索构建伪平行语料. 在计算不同风格数据集间文本的语义相似度后, 将超过某个预先设定阈值的句子与原有文本构成伪平行语料对. 在迭代回译的过程中, 比较该轮生成的结果与检索结果

和源句之间的词移距离 (word mover's distance, WWD)^[46],将距离更近的句子作为下一轮的伪平行语料,避免在迭代回译过程中出现模型崩溃的情况。

4.3.2 基于强化学习的方法

Luo 等^[14]认为在没有并行数据的情况下解耦很难分离句子的内容与风格,并提出了一个双向强化学习框架(如图6所示),将2种风格间文本相互转换的

过程视为一种双向映射的学习任务.针对文本的风格转换和文本内容保持设计2个奖励函数,由得分引导模型学习风格间的映射.同时,参考迭代回译的思想,提出伪退火教师教学算法,每次将上一轮的生成结果作为伪平行语料,再借助伪平行语料进行下一轮的训练,迭代提升模型的生成质量.在2个基准数据集的实验模型中取得了很好的效果。

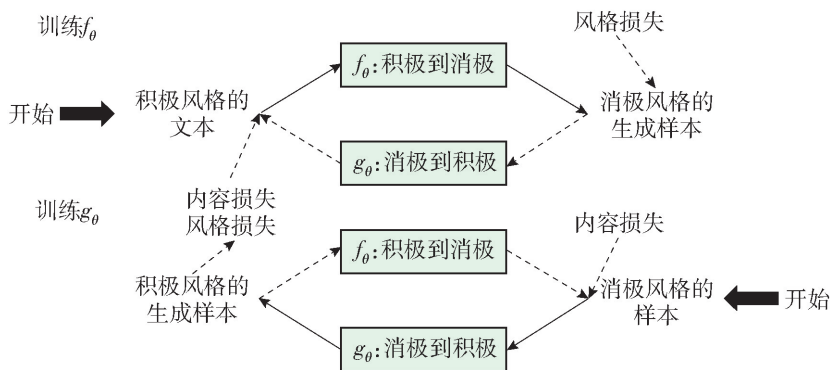


图6 基于双向强化学习约束的文本风格转换模型训练流程^[14]

Fig. 6 Training process of text style transfer model based on two-way reinforcement learning constraints^[14]

4.3.3 基于隐空间编辑的方法

Liu 等^[15]认为可以在隐空间中直接学习不同风格文本隐空间特征存在的映射关系,编辑隐空间特征,实现风格转换.其创新点在于可以直接利用分类器的梯度回传影响隐空间特征.通过迭代的方式不断修改隐空间特征,直到分类器能将转换后的特征识别成目标风格,从而学习到隐空间中不同风格的分布间的映射.为了限制隐空间编辑的迭代过程中的参数的更新幅度,在损失函数中增加正则项,避免由于更新幅度过大导致的信息丢失。

4.4 各类研究工作的比较与分析

在表2中列举了各类风格转换的研究工作在Yelp数据集上的实验结果,其中参考-BLEU指的是计算生成结果与预设的参考结果的相似度(BLEU是一种衡量生成样本质量的评价标准,会在5.1.2节详细介绍),输入-BLEU指的是计算生成样本与输入数据间的相似度,这2项指标用于衡量生成样本中内容的保留度.准确率指的是生成样本中具有指定风格的比例,由训练好的分类器对生成样本打分,衡量风格转换的成功率。

由表2可得到一些结论:第一,对实验结果分析后可推断模型的准确率与原文-BLEU值存在负相关,即包含原有内容越多越容易被判断为原风格.这意味着这些模型在训练时可能会采取牺牲一个指标来换取另一个指标的提升的策略,需要在模型训

练的时候加以平衡.第二,无论是按照哪一种研究思路的工作,当前最优的实验结果差别不大,但不同的研究思路在各个指标上的表现有所不同.隐空间解耦在训练过程中会采用重构损失来减少风格转换过程中内容的损失,此类生成方式可比较自然地完成本风格转换,在参考-BLEU指标上表现出较大优势.显性解耦中的一些风格转换方法是通过修改或添加风格词汇实现的,在风格的转换上能表现出较高成功率,如果原文与参考结果差异过大,可能会导致参考-BLEU得分较低.基于一映射的风格转换方法是通过迭代训练实现的,在保证语义内容不丢失的情况下不断向目标风格靠近,因此,此类工作在2个指标上表现比较均衡.第三,Transformer模型的加入能够有效地提升生成样本的质量,尤其是在Wu等^[10]中引入了基于Transformer的预训练语言模型,借助丰富的语言知识,在不同指标上都有所提升。

在模型的可解释性上,基于隐空间解耦和基于一映射的方法模型存在可解释性较差、不易训练、训练过程不可控制等问题,而基于显性解耦的方法模型具有可解释性好,容易调优、调试和改进的优点。

在模型时空复杂度的比较上,3种方法如果采用同种编解码器,空间复杂度是一样的.在时间复杂度的比较中,因为基于一映射的方法需要大量的迭代训练,所以该方法的时间复杂度高于基于隐

表 2 文本风格转换在 Yelp 数据集上的实验结果

Table 2 Experimental results of text style transfer on Yelp dataset

来源	研究思路	模型	方法	参考-BLEU	原文-BLEU	准确率/%
原文				23	100	3.8
Fu 等 ^[3]	基于隐空间解耦	RNN	多编码器模型	12.9	40.1	46.9
		RNN	风格向量模型	9.2	37.9	49.8
Prabhumoye 等 ^[4]	基于隐空间解耦	RNN		6.8		87.2
Dai 等 ^[6]	基于隐空间解耦	Transformer	条件约束	17.1	45.3	93.7
		Transformer	多分类约束	20.3	54.9	87.7
John 等 ^[7]	基于隐空间解耦	RNN				93.4
Li 等 ^[8]	基于显性解耦	RNN	仅检索	0.4	0.7	95.4
		RNN	基于模板	11.8	44.1	81.7
		RNN	仅删除	7.5	28.6	85.7
		RNN	删除与检索	8.4	29.1	87.7
Wu 等 ^[10]	基于显性解耦	Transformer		20.7		97.3
Zhang 等 ^[12]	基于一步映射	RNN		22.79		96.6
Jin 等 ^[13]	基于一步映射	RNN		22.46		93.2
Luo 等 ^[14]	基于一步映射	RNN		17.71	58.72	85.6
Liu 等 ^[15]	基于一步映射	RNN		18.8		92.3

空间解耦的方法。基于显性解耦的方法需要对数据进行预处理,该方法的时间复杂度高于基于隐空间解耦的方法。另外,决定模型的时空复杂度的另一个重要因素是编解码器采用何种模型。采用 Transformer 模型会带来更大的参数量,同时,会大幅增加模型的训练时间。

在文本风格转换任务中存在各种类型的数据集,一些句子的风格可能是由句子整体决定的。此类数据集难以分离句子中的内容特征与风格特征,可以采用基于一部映射的方法。对于一些句子结构比较简单的数据集,基于隐空间解耦和基于显性解耦的方法都可以应用在这些数据集中。

5 评价指标

评价文本风格转换生成质量的方法通常分为人工评价和自动评价 2 种,均是从生成文本的风格转换准确率、内容的保留度以及文本的流利度 3 个方面进行评价。这一观点得到了广泛的认可^[47-48]。

5.1 自动评价指标

5.1.1 风格转换准确率

风格转换准确率是评价风格转换模型的一个重要指标,它是指在所有的生成文本中,风格转换成功

的句子所占的比例。准确率越高代表模型的风格转换效果越好。一般采用的方法是用预训练的文本分类器对生成的句子进行分类。在实际应用中,研究者则会采用不同的模型作为分类器。Shen 等^[5]使用 TextCNN^[49]模型作为分类器,测试表明,在数据集中进行预训练可以获得 90% 以上的分类精度。

5.1.2 内容的保留度

随着对生成句子进一步分析,研究者发现即使生成的句子在风格转换准确率这一指标上得分较高,转移风格后的句子内容与输入句子的内容并不相关,文本风格转换也是失败的。对于保存程度的测量,研究者们提出了很多方法,其中 BLEU^[50]是使用最广泛的方法。BLEU 是评估输出句子与参考句子之间字或词重叠程度的指标,代表了 2 个句子的相似程度,BLEU 分数越高,表示 2 条文本越相似,代表模型内容保存得越好。此外,还有一种基于词嵌入的度量方式,通过 Word2Vector^[51]或 GloVe^[45]获得风格转换前后句子的向量表示,将向量表示之间的余弦相似度作为内容保留度的评价标准。在 Bert^[18]模型被提出后,一些研究者提出可以采用语境化的词嵌入代替静态的词嵌入,并采用 Bert 模型对内容的保留度打分,可以获得更准确的评价标准,

即 BertScore^[26].

5.1.3 流利度

仅从风格转换准确度和内容给出的保留度评价,仍然无法保证生成的句子足够好.如果生成的句子完全不符合人类的阅读习惯,风格转换模型还是失败的.因此,判断句子是否流畅也很有必要.Mir等^[47]在真实句子和风格转换句子样本上训练了一个基于单字的神经逻辑回归分类器,判断生成的句子是真实的还是模型生成的.最后,计算通过分类器的句子数占所有转换生成句子的比例,并给出流畅度分数.另外,采用困惑度来衡量生成句子的流利度也是一种评价流利度的方法.一般采用预训练的语言模型对句子打分,越符合人类阅读习惯的困惑度越低.

5.1.4 自动评价指标存在的缺点与改进

上述3种方法均在单个维度评价模型的生成质量,无法从多个维度反映整体模型的质量,因此,可以联合多种评价方法来评价模型.比如:计算准确率和BLEU的几何平均值和谐波平均值,对模型的综合能力打分.

除此之外,在以困惑度作为评价流利度的指标时,可能会遇到生成样本中存在一些比较罕见的单词,导致作为评价指标的困惑度升高,错误地评价模型.除了采用人工评价作为补充,目前没有更好的解决办法.

5.2 人工评价

自动评价的结果可以快速地提供模型的评价指标,但是自动评价无法提供模型进一步训练的修正意见,无法直观地反映风格转换过程中出现的问题.人工评价可以弥补这一缺陷,人工分析生成样本可为模型训练提供更多意见.关于人工评价,研究人员需要先找到一些具有语言背景的工作人员评价模型.一般来说,研究人员会找到3~6个人对每个句子进行打分.使用Stent等^[52]提出的1~5利克特量表对句子的3个指标进行打分.评估过程需要对样本充分打乱顺序,从中取出足够的样本,同时,保证每个评估者无法知道句子的来源.

6 展望

风格转换的任务实质上是改写句子的风格.从狭义上讲,风格转换的任务是二值属性间的风格转换,比如:积极到消极、书面语言与非书面语言.从广义上讲,风格转换可以包括多个风格之间的转换^[53],以及文章与文章主题间的转换^[54],或者细粒

度上的风格转换^[55].除此之外,目前大多数的工作都是基于单句的文本风格转换.在实际应用中,句子通常还有上下文,转换单句的时候需要考虑上下文才能实现更自然的风格转换样本.Cheng等^[33]提出基于上下文内容一致性的文本风格转换任务.

随着大规模预训练语言模型的发展,预训练模型在许多下游任务中显示出优越性^[18],将预训练语言模型应用到风格转换任务中,可能会推动风格转换任务的发展.

本文基于当前文本风格转换任务的发展现状总结出以下3个方面的挑战.

6.1 更复杂的风格转换任务

在自然语言处理的实际应用中,二元属性的风格转换不能满足所有实际应用场景的需要.因此,研究者提出一些更复杂的文本风格转换任务.第1类,多风格间的转换任务.Helbig等^[53]将句子中的情感属性作为一种风格,提出一种端到端的基于词典替换的六分类情感转换模型.模型分为2部分:一部分负责找到需要替换的词,另一部分负责填词,保证新的句子包含目标情感.另外,增加了打分模块,对多个生成结果从情感风格、内容保留以及流利度这三方面打分,将得分最高的句子作为最后的结果.从模型的实验结果上看,只能处理一些简单的句子,主要目的是将二元的文本风格转换任务扩展到多元.第2类,将不同的文体作为风格.Yang等^[54]提出在摘要和文章这2种文体之间实现风格转换.与一般的文本风格转换任务不同,文本风格是由整段文字决定的.为了弥补不同风格间数据的差异,补充上下文信息,该研究还引入了一个外部数据库,实现了在摘要和正文间的转换.第3类,Luo等^[55]提出一种细粒度的风格转换模型,在传统粗粒度(正向或负向风格)的基础上,增加细粒度的控制,可以控制风格表达的程度,表达更细腻的风格.

6.2 基于上下文一致性的文本风格转换

以前风格转换任务的工作主要研究对象是单句之间的风格转换,没有考虑句子在风格转换后对上下文的影响.因此,Cheng等^[33]提出一个新的文本风格转换任务,即基于上下文的文本风格转换任务.将上下文也作为控制样本生成的条件,在单句风格转换的基础上,保持上下文的一致性.在该研究中,模型将上下文与原句信息共同作为编解码器的输入,另外,增加了判断生成样本是否与上下文一致的分类器,引导模型生成样本.

6.3 预训练语言模型

预训练语言模型已经证明了它们在各种自然语言处理任务中的能力,包括机器翻译、文本分类、命名实体识别。特别是 GPT2^[40] 在自然语言生成方面取得了显著的成功。如果将预训练模型所学知识应用于语体迁移任务中,能有效地提高迁移句子的质量,同时保证风格转换句子的流畅性。

参考文献:

- [1] GATYS L A, ECKER A S, BETHGE M. A neural algorithm of artistic style[EB/OL]. [2020-12-07]. <http://arxiv.org/abs/1508.06576>.
- [2] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2223-2232.
- [3] FU Z, TAN X, PENG N, et al. Style transfer in text: exploration and evaluation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2018, 32(1): 663-670.
- [4] PRABHUMOYE S, TSVETKOV Y, SALAKHUTDINOV R, et al. Style transfer through back-translation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2018: 866-876.
- [5] SHEN T, LEI T, BARZILAY R, et al. Style transfer from non-parallel text by cross-alignment[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2017: 6833-6844.
- [6] DAI N, LIANG J, QIU X, et al. Style transformer: unpaired text style transfer without disentangled latent representation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 5997-6007.
- [7] JOHN V, MOU L, BAHULEYAN H, et al. Disentangled representation learning for non-parallel text style transfer[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 424-434.
- [8] LI J, JIA R, HE H, et al. Delete, retrieve, generate: a simple approach to sentiment and style transfer[C]//2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2018: 1865-1874.
- [9] XU J, SUN X, ZENG Q, et al. Unpaired sentiment-to-sentiment translation: a cycled reinforcement learning approach[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Computational Linguistics, 2018: 979-988.
- [10] WU X, ZHANG T. "Mask and Infill": applying masked language model to sentiment transfer[C]//International Joint Conferences on Artificial Intelligence. Menlo Park: AAAI Press, 2019: 5271-5277.
- [11] SUDHAKAR A, UPADHYAY B, MAHESWARAN A. "Transforming" delete, retrieve, generate approach for controlled text style transfer[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Menlo Park: AAAI Press, 2019: 3260-3270.
- [12] ZHANG Z, REN S, LIU S, et al. Style transfer as unsupervised machine translation[EB/OL]. [2020-12-07]. <http://arxiv.org/abs/1808.07894>.
- [13] JIN Z, JIN D, MUELLER J, et al. IMaT: unsupervised text attribute transfer via iterative matching and translation[C/OL]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Computational Linguistics, 2019 [2020-12-07]. <https://www.aclweb.org/anthology/D19-1306.pdf>.
- [14] LUO F, LI P, ZHOU J, et al. A dual reinforcement learning framework for unsupervised text style transfer[C/OL]. International Joint Conferences on Artificial Intelligence. Menlo Park: AAAI Press, 2019 [2020-12-07]. <https://www.ijcai.org/Proceedings/2019/0711.pdf>.
- [15] LIU D, FU J, ZHANG Y, et al. Revision in continuous space: unsupervised text style transfer without adversarial learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2020, 34(5): 8376-8383.
- [16] WANG Y, LEE H Y. Learning to encode text as human-readable summaries using generative adversarial networks[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 4187-4195.
- [17] DOS SANTOS C N, MELNYK I, PADHI I. Fighting offensive language on social media with unsupervised text

- style transfer [C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg: Association for Computational Linguistics, 2018: 189-194.
- [18] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.
- [19] SHU L, PAPANGELIS A, WANG Y C, et al. Controllable text generation with focused variation[C] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. Stroudsburg: Association for Computational Linguistics, 2020: 3805-3817.
- [20] LEEFTINK W, SPANAKIS G. Towards controlled transformation of sentiment in sentences [EB/OL]. [2020-12-07]. <http://arxiv.org/abs/1901.11467>.
- [21] LI P. An empirical investigation of pre-trained transformer language models for open-domain dialogue generation[EB/OL]. [2020-12-07]. <http://arxiv.org/abs/2003.04195>.
- [22] ZHOU H, YOUNG T, HUANG M, et al. Commonsense knowledge aware conversation generation with graph attention[C] // Proceedings of the 27th International Joint Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2018: 4623-4629.
- [23] BAHDANAU D, CHO K H, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2020-12-07]. <https://arxiv.org/pdf/1409.0473.pdf>.
- [24] RUSH A M, CHOPRA S, WESTON J. A neural attention model for abstractive sentence summarization [C] // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 379-389.
- [25] JANG E, GU S, POOLE B. Categorical reparameterization with gumbel-softmax [EB/OL]. [2020-12-07]. <https://arxiv.org/abs/1611.01144>.
- [26] ZHANG T, KISHORE V, WU F, et al. BERTScore: evaluating text generation with BERT[EB/OL]. [2020-12-07]. <https://arxiv.org/pdf/1904.09675.pdf>.
- [27] HU Z, YANG Z, LIANG X, et al. Toward controlled generation of text [C] // International Conference on Machine Learning. Brookline: PMLR/Microtome Publishing, 2017: 1587-1596.
- [28] MAAS A, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis [C] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011: 142-150.
- [29] DOU Q, KNIGHT K. Large scale decipherment for out-of-domain machine translation [C] // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: Association for Computational Linguistics, 2012: 266-275.
- [30] XU W, RITTER A, DOLAN W B, et al. Paraphrasing for style [C] // 24th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2012: 2899-2914.
- [31] RAO S, TETREAULT J. Dear Sir or Madam, may I introduce the GYAFC dataset: corpus, benchmarks and metrics for formality style transfer [C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg: Association for Computational Linguistics, 2018: 129-140.
- [32] WU Y, WANG Y, LIU S. A dataset for low-resource stylized sequence-to-sequence generation [C] // Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2020, 34(5): 9290-9297.
- [33] CHENG Y, GAN Z, ZHANG Y, et al. Contextual text style transfer [C] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. Stroudsburg: Association for Computational Linguistics, 2020: 2915-2924.
- [34] CARLSON K, RIDDELL A, ROCKMORE D. Evaluating prose style transfer with the Bible [J]. Royal Society Open Science, 2018, 5(10): 171920.
- [35] JHAMTANI H, GANGAL V, HOVY E, et al. Shakespearizing modern language using copy-enriched sequence-to-sequence models [C] // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2017, 6: 10-19.
- [36] VINYALS O, FORTUNATO M, JAITLEY N. Pointer networks[EB/OL]. [2020-12-07]. <http://arxiv.org/abs/1506.03134>.
- [37] GU J, LU Z, LI H, et al. Incorporating copying

- mechanism in sequence-to-sequence learning [C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg; Association for Computational Linguistics, 2016; 1631-1640.
- [38] NIU X, RAO S, CARPUAT M. Multi-task neural models for translating between styles within and across languages [C] // Proceedings of the 27th International Conference on Computational Linguistics. Stroudsburg; Association for Computational Linguistics, 2018; 1008-1021.
- [39] XU R, GE T, WEI F. Formality style transfer with hybrid textual annotations [EB/OL]. [2020-12-07]. <http://arxiv.org/abs/1903.06353>.
- [40] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[EB/OL]. [2020-12-07]. <http://www.persagen.com/files/misc/radford2019language.pdf>.
- [41] WANG Y, WU Y, MOU L, et al. Harnessing pre-trained neural networks with rules for formality style transfer[C] // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg; Association for Computational Linguistics. Menlo Park; AAAI Press, 2019; 3564-3569.
- [42] ZHANG Y, GE T, SUN X. Parallel data augmentation for formality style transfer[C] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg; Association for Computational Linguistics, 2020; 3221-3228.
- [43] KINGMA D P, WELING M. Auto-encoding variational bayes[EB/OL]. [2020-12-07]. <http://arxiv.org/abs/1312.6114>.
- [44] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C] // Proceedings of the 31st International Conference on Neural Information Processing Systems. Cambridge; MIT Press, 2017; 6000-6010.
- [45] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation[C] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg; Association for Computational Linguistics, 2014; 1532-1543.
- [46] KUSNER M, SUN Y, KOLKIN N, et al. From word embeddings to document distances [C] // International Conference on Machine Learning. Brookline; Microtome Publishing, 2015; 957-966.
- [47] MIR R, FELBO B, OBRADOVICH N, et al. Evaluating style transfer for text [C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg; Association for Computational Linguistics, 2019; 495-504.
- [48] PANG R Y, GIMPEL K. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer [C] // Proceedings of the 3rd Workshop on Neural Generation and Translation. Stroudsburg; Association for Computational Linguistics, 2019; 138-147.
- [49] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. [2020-12-07]. <https://arxiv.org/pdf/1408.5882.pdf>.
- [50] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation [C] // Proceedings of the 40th annual meeting of the Association for Computational Linguistics. Stroudsburg; Association for Computational Linguistics, 2002; 311-318.
- [51] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C] // Proceedings of the 26th International Conference on Neural Information Processing Systems. Cambridge; MIT Press, 2013; 3111-3119.
- [52] STENT A, MARGE M, SINGHAI M. Evaluating evaluation methods for generation in the presence of variation [C] // International Conference on Intelligent Text Processing and Computational Linguistics. Berlin; Springer, 2005; 341-351.
- [53] HELBIG D, TROIANO E, KLINGER R. Challenges in emotion style transfer; an exploration with a lexical substitution pipeline [C] // Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media. Stroudsburg; Association for Computational Linguistics, 2020; 41-50.
- [54] YANG P, LI L, LUO F, et al. Enhancing topic-to-essay generation with external commonsense knowledge [C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg; Association for Computational Linguistics, 2019; 2002-2012.
- [55] LUO F, LI P, YANG P, et al. Towards fine-grained text sentiment transfer[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg; Association for Computational Linguistics, 2019; 2013-2022.

(责任编辑 梁 洁)