

文本风格转换出了什么问题？

Alexey Tikhonov

altsoph@gmail.com

Ivan P. Yamshchikov

ivan@yamshchikov.info

摘要

许多最近的机器学习论文都探讨了文本自动风格转换，一个反直觉的事实是，它们对这种自然语言任务并无一致的表述。不同的研究人员提出不同的算法、数据集和度量指标来对任务进行描述。本文旨在讨论这一 NLP 任务的可能形式化，以深化本任务的研究。

引入

可以说，Gatys 等人(2016)引起了人们对文本风格转换算法的广泛关注，他们发展了效果惊艳的图像风格转换算法。Gatys 等人(2016)通过从特定层的网络输出中重构输入图像，在卷积神经网络(CNN)的不同处理阶段可视化了信息。这种方法使作者免于明确定义风格，并发现：网络的较高层抛弃了具体的像素信息，而图像的高层内容(high-level content)得以保留的。同时，丢弃关于场景的全局排布信息可以从不同的风格表示中重建输入图像的风格，这些风格表示都保存在 CNN 层的不同子集上。这种使高层图像信息变得明晰的高效信息解耦，似乎是针对对象识别而优化的 CNN 的内部特性，而在目前还不适用于文本。在 Zhang 和 LeCun (2015)的文章中，作者将 CNN 用于文本理解的任务中，这有望为未来的文本提供一个简单的语义-风格分解的机会，然而由于多种原因导致实现这一目标有相当难度。第一个困难是文字信息不具有图像特有的连续性；第二个是文本没有一个“典型”的指标可以单独考察文本的风格信息：不能说文本的风格是由字母决定的，而与单词或搭配没有任何关系；最后一个为人们通常考虑的是句子的风格化而非更长的文本模式。这意味着系统的每一个独立的输入中，可供系统使用的风格信息数量不可避免地要少得多。然而，这些困难以及缺乏对文本风格的明确一致定义，在某种程度上并没有影响对文本语境中风格转换问题的直观理解。这个问题的直观本质和其形式的复杂性之间的矛盾，催生出了许多研究项目。这些项目不仅从截然不同的角度看待这个问题，事实上，有的甚至完全是套着同一个名字在研究其他问题。(This contradiction between an intuitive nature of the problem and its' formal complexity results in a number of research projects that not only look at the problem from very different angles but, in fact, even look on different problems giving them the same name)此外，我们还列出了一些不同的贡献，大致可以分为三大类，它们对这个问题有着非常不同的理解。

相关工作

我们对当前文本风格转换的方法进行了系统化:

1. Ad-hoc defined style classes

尽管一个句子的情感并不等同于它的风格,但是有很多作品特别关注文本的情感。例如,在 Li 等人(2018)的研究中,作者使用二元情感分类器来评估风格转换的质量,分类器是在包含 Yelp 和 Amazon 的评论的同一个语料库上训练的,该语料库也被用于其他文本风格转换系统的训练。在该领域还有很多结果,参见 Kabbara 和 Cheung(2016),或者 Xu 等人(2018)。我们认为应该将这种对文本风格的理解称为临时的(Ad-hoc),因为这里的样式概念是从给定的训练数据集中严格逆向工程出来的。Ficler 和 Goldberg(2017),特别是 Fu 等人(2017)以一种清晰合理的方式概括了该方法。Fu 等人(2017)将风格定义为一组可测量的范畴(categorical)和/或连续参数。他们为每个包含文本风格的参数训练了一个分类器,然后用这些预训练好的分类器测试结果输出,最后使用分类正确句的百分占比来度量文本风格转换的成功率。主观上,Ad-hoc 方法主观上对于许多工业界任务来说极为有效。这一部分得益于清晰、实用的问题设置,另一部分是由于可用于 Ad-hoc 风格转换方法的自然语言人造(human-supervised)数据集日益增多。通过 Tikhonov 和 Yamshchikov(2018)等人的人类同行评审,该方法可以很好地解决问题。然而,由于这类方法并不涉及任何全局且一致的对风格的理解,因此将其称为“风格迁移”,不如称为情感转换或者说文本摘要会更合理。

2. NMT approaches

这些工作背后的思想是将两种不同的风格定义为两种不同的语言,从而将风格转换的问题转化为神经机器翻译的问题(NMT)。这类方法的典型例子是所谓的“时代风格”(Hughes 等人(2012), Xu 等(2012))或 Jhamtani 等(2017)用包含莎士比亚原文及白话翻译的平行语料训练的自动“莎士比亚生成器”。Carlson 等人(2017)使用圣经的翻译本作为平行语料,得出了内容自动简化“可以简单地被视为风格迁移任务”这一结论。由于缺乏具有语义等但风格相异的平行数据集,这种方法在实际的应用中有严重缺陷。因此 Jang 等人(2017)试图找到一种生成平行语料的自动化方法来解决这一问题。在 Rao 和 Tetreault(2018)的工作中,介绍了一个用于文本正式性风格迁移的数据集,以及在 NMT 语境下的评测机制和度量指标。对齐数据集的问题在 NMT 方法下依旧存在: Xu 等(2017)列举了七种语言风格,且给出了其论断:“世界上肯定有七种以上的语言风格,毕竟世界上都不只有七大奇迹”。每个进一步的“奇迹”都需要一个独立的对齐训练语料,所以 NMT 方法是一类相对劳动密集的方法。

3. Post-NMT approaches

这些方法和 NMT 方法的逻辑一致,但研究者们一致认为,风格标记不应该像 Ad-hoc 方法那样被割裂开,而应该从语料库中自动地抽取出来。不同风格即是不同语言这一类比在这里也成立。然而,由于缺乏平行语料库这一限制,研究者们试图从近期的无样本(zero-shot)NMT 技术、GAN 等方法中迂回地解决问题。近几年来,post-NMT 方法的观念开始形成势头,比如 Artetxe 等人(2017), Han 等人(2017), Shen 等人(2017), Zhang 等人(2017), Prabhumoye 等人(2018), Zhang 等人(2018)的工作。所有这类方法都试图单独获得语义和风格的隐含表示(类似于我们对图像的信息解耦)。一般以以下多种方式实现:

- 获取单词或句子的正则化或某种对齐嵌入,并将嵌入状态空间分割为语义和风格
- 使用两次迁移(来回两次)作为风格迁移方法的质量目标
- 训练一个风格判别器

文本风格迁移

当前风格迁移的方法综述引发了许多讨论。其中最有趣的一个问题是：文本风格和文本语义的隐含直观理解之间是否有联系。事实上，Ad-hoc 类的观点认为风格并不意味着在风格转换后句子的语义部分不会改变。例如，如果假设情感是一种风格特性——就像上述的一些作者的观点那样——这可能会导致下面的问题。若在数据集 X 上单独优化语义和风格对应的损失函数，损失函数将会是这样的形式：

$$L_{total}(X) = A_{semantic}(X) + A_{stylistic}(X)$$

其中 A 表示某种风格迁移操作的准度指标。

一般来说，人们会认为“这里的糖果很棒”和“这里的糖果很糟”是两个风格相同，语义相反的句子。但若要把情感当作一种风格特征，那么就要将 $A_{stylistic}(X)$ 看作是 $A_{semantic}(X)$ 的一个函数（这意味着谈论风格迁移毫无意义），要么就得承认存在一种信息解耦方式，使得 $A_{stylistic}(X)$ 和 $A_{semantic}(X)$ 是 X 上相互独立的函数。事实上在这个例子中，若假设“这里的糖果(this place has candy)”包含了句子中全部有效的语义，而“很棒(great)”是句子中“风格”的部分，就能得到一种相互独立的信息解耦。不过这样的信息解耦方式意味着最终损失必然不稳定，句子的语义保留度和风格准确性会遭受不同程度的损失。此外，由于 $A_{semantic}(X)$ 和 $A_{stylistic}(X)$ 度量方法上的不同，也可能使最终损失 $L_{total}(X)$ 很不稳定。例如，“这里的糖果很一般(this place has average candy)”，“那家伙的焦糖很糟(that fella sells awful caramels)”，“那些餐厅的小糖果恶心爆了(those restaurants serve horrific bonbons)”可能最终损失 $L_{total}(X)$ 很接近。人们直观上会认为“那家伙的焦糖很糟”和“那些餐厅的小糖果恶心爆了”是两个语义等价且风格不同的句子，但这却并不那么满足“情感是一种风格属性”的假设。还必须提到一点，在这种可解耦的假设下，由于语言的多模态性，“这里有棉花糖(this place has cotton candy)”这句话在语义上等价于“这里的糖果很棒(this place has great candy)”，且风格上会被标为情感中性。所以如果使用等式 1 中的损失 $L_{total}(X)$ ，可能更适合这样的句子：“那家伙卖的焦糖很糟糕(that fella sells awful caramels)”（后一句的情感部分不会去补偿它与“这里的糖果很棒”的语义部分上的差异）。这也会导致“这里的糖果很一般(this place has average candy)”与“这里有棉花糖(this place has cotton candy)”的语义等价，即使常识上它们是风格相似但语义不同的句子。这个简单的例子说明了 Ad-hoc 方法的一个常见缺陷，即这种方法必须在特定的预定义信息解耦假设下确定损失，而做出这个假设并非易事。虽然它并不妨碍 Ad-hoc 在特定 NLP 任务上的应用，但是人们必须留意并控制好这一缺陷。另一方面，NMT 方法虽然不会遇到损失解耦难题，但也有些其他问题，主要有以下三点：

1. 可以说任何两个语义对齐的语料库就是一个有效的风格对吗？比如说，从“narcissistic”到“academic”的风格转换是一个风格迁移任务吗？
2. 一个语料库中的风格会有交叠吗？我们如何控制它？
3. 语料库中的每个句子都有明确的风格成分吗？在语料库中，如果没有明确的风格成分，我们应该删除或混合句子吗？这样的操作会改变最后的风格迁移算法吗？

这些问题 post-NMT 方法也有，因为两者的主要区别在于两个语料库是否语义对齐，而风格在 NMT 和 post-NMT 方法中均为语料库的隐含内容分布。

除此之外我们为风格迁移任务提出了两个准则：

1. 风格是文本的一个整体属性，它可以被标记为一个特定类别或一个子语料。这种类别或子语料的选择是任务相关的。
2. 风格必须通过以下方式与语义“正交”：任何与语义相关的信息都可以任何风格表示。在这两个准则下，风格转换任务就变成了相对于某个风格参考系的平移，且对齐了隐含表示

以便进行风格转换。词汇和句子结构是进行这种转换的两个主要工具。令句子 X 和其变体 $\tilde{X}$ 在状态空间 $\mathbb{R}^n \otimes \mathbb{S}$ 中表为两个点，其中 $\mathbb{S}$ 是风格维度而 $\mathbb{R}^n$ 是语义嵌入空间。风格迁移问题的损失函数就可以定义为：

$$L_{total}(X, \tilde{X}) = D_{\mathbb{R}^n}(X, \tilde{X}) + D_{\mathbb{S}}(X, \tilde{X})$$

其中 $D_{\mathbb{R}^n}$ 和 $D_{\mathbb{S}}$ 为给定语料上的距离指标。这样形式化的要点是我们不在 $\mathbb{S}$ 上强加任何结构。这种结构是通过给定语料库的探索获得的，且从一种风格转换过来的句子与从另一种风格转换过来的句子相合。

结论

在这篇短文中，我们列出了处理风格迁移问题的不同方法。我们将它们大致分为三大类：1) 使用预定义度量集处理风格的 Ad-hoc 方法，2) 使用语义对齐的语料库，将风格视为文本的整体属性，并认为不应将其明确形式化的 NMT 方法，3) 以及基于未明确形式化的相同风格表示，且试图避免构建对齐平行语料库的 post-NMT 方法。我们主张将 Ad-hoc 任务与 NMT、post-NMT 风格迁移任务区分看待。我们还认为在风格和语义之间特定正交性的假设下进行风格迁移，在 NMT 和 post-NMT 的方法中是可以实现的，但使用 Ad-hoc 的方法无法达成这一目的

参考文献

- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. In arXiv preprint.
- Keith Carlson, Allen Riddell, and Daniel Rockmore. 2017. Zero-shot style transfer in text using recurrent neural networks. In arXiv preprint.
- Jessica Fidler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In arXiv preprint.
- Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2017. Style transfer in text: Exploration and evaluation. In arXiv preprint.
- Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural network. Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference, pages 2414–2423.
- Mengqiao Han, Ou Wu, and Zhendong Niu. 2017. Un-supervised automatic text style transfer using LSTM. National CCF Conference on Natural Language Processing and Chinese Computing, pages 281–292.
- James M. Hughes, Nicholas J. Foti, David C. Krakauer, and Daniel N. Rockmore. 2012. Quantitative patterns of stylistic influence in the evolution of literature. Proceedings of the National Academy of Sciences, 109(20):7682–7686.
- Se Won Jang, Jesik Min, and Mark Kwon. 2017. Writing style conversion using neural machine translation. In preprint.
- Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. In arXiv preprint.
- Jad Kabbara and Jackie Chi Kit Cheung. 2016. Stylistic transfer in natural language

generation systems using recurrent neural networks. Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods, pages 43–47.

- Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In arXiv preprint.
- Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Styletransfer through back-translation. In arXiv preprint.
- Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In arXiv preprint.
- Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. 31st Conference on Neural Information Processing Systems, pages 6833–6844.
- Alexey Tikhonov and Ivan P. Yamshchikov. 2018. Guess who? multilingual approach for the auto-mated generation of author-stylized poetry. In arXiv preprint.
- Jingjing Xu, Xu Sun, Qi Zeng, Xuancheng Ren, Xiaodong Zhang, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In arXiv preprint.
- Wei Xu. 2017. From shakespeare to twitter: What are language styles all about? Proceedings of the Workshop on Stylistic Variation, pages 1–9.
- Wei Xu, Alan Ritter, William B. Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. Proceedings of COLING, pages 2899–2914.
- Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Style transfer from non-parallel text by cross-alignment. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 1:1959–1970.
- Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. In arXiv preprint.
- Ye Zhang, Nan Ding, and Radu Soricut. 2018. Shaped: Shared-private encoder-decoder for text style adaptation. In arXiv preprint.