Project Plan

Language Style Transfer on Non-Parallel Corpora

Supervisor: Professor Benjamin C.M. Kao

Zhijing Jin
maggie0@hku.hk

Jingran Zhou
jrchow@hku.hk

Edward Lui
edwardlw@hku.hk

Abstract

Language style transfer is a popular task in Natural Language Processing (NLP) which aims to modify the style of a sentence while keeping its content unchanged. Previous work mainly focuses on using adversarial methods, which are hard to train and struggled to produce high-quality outputs. In this project, we will first evaluate the performance of the state-of-the-art approaches, and then explore a novel approach, iterative semantic matching. Our model will be evaluated on three commonly used benchmarks: Yelp sentiment data set, formality style data set, and a democratic-versus-republican political slants data set. Our approach can be applied to a wide range of applications, including stylistic dialogue systems, text formality conversion, and news text generation.

Contents

1. Objectives

The ability to transfer styles of texts or images is an important measurement of the advancement of artificial intelligence (AI). We consider the task of text attribute transfer: altering a specific attribute (e.g., sentiment) of a sentence while preserving its attribute-independent content (e.g., changing "*screen is just the right size*" to "*screen is too small*") as *language style transfer* [7, 16, 3, 9]. The goal of this task is to transform an input sentence of one style (e.g. positive) to another style (e.g. negative) while preserving its style-independent content. In this context, language "*style*" refers to a range of linguistic phenomena including syntactic simplification, sentiment modification and others.

The goal of "controlled" language generation, which separates the semantic content of *what* is said from the stylistic dimensions of *how* it is said, has motivated a considerable number of research efforts. These include approaches relying on heuristic substitutions, deletions, and insertions to modulate demographic properties of a writer [13], integrating stylistic and demographic speaker traits in statistical machine translation [10, 8], and deep generative models controlling for a particular stylistic aspect, e.g., politeness [15], sentiment [16], or tense[6]. The latter approaches to style transfer, while more powerful and flexible than heuristic methods, are difficult to train [14, 1, 2] and have yet to show a high quality results, as most generated texts are fraught with grammar errors and shifted contents [7]. In a more recent work, Li et al. [7] demonstrate that directly implementing heuristic transformations such as modifying high polarity words in reviews outperforms the approaches based on Generative Adversarial Networks (GANs). This suggests the promise of developing style transformation algorithms without explicitly disentangling style and content at the representation level.

In this project, we propose to develop a novel approach to language style transfer. Instead of directly operating in the space of non-parallel data, we will incrementally construct a pseudo-parallel resource and feed it into a standard sequence-to-sequence machine translation system. The algorithm will start by identifying close sentence pairs across different corpora based on lexical similarity, and then utilize the translation models outputs to refine this data set iteratively. We will demonstrate our model on three benchmark data sets, including the Yelp review sentiment modification data set[1], political slant data set[17], and formality style data set[12]. The model outputs will be evaluated in terms of content preservation, fluency and sentiment correctness. We aim to outperform the state-of-the-art approaches by a significant margin in all three assessments.

[1] Data set available at https://www.yelp.com/dataset/challenge

2. Problem Statement and Benefits

The problem we aim to tackle is style transfer, namely given an input sentence x_0 with a style attribute s_0 and content c_0, we want to generate a new sentence x_1 with a different style attribute $s_1 \neq s_0$ but still with the same content $c_1 = c_0$, where style attributes s_* refers to a set of linguistic phenomena. We will develop a simple and easy-to-train model that outperforms the state-of-the-art approaches on three tasks, including sentiment modification, formality transfer, and political perspective transfer. The evaluation will be done with respect to content preservation, fluency and style correctness.

A multitude of applications can potentially benefit from solving this problem, for example, the design of digital assistants and chat bots with diverse tones. To illustrate, a simple reminder for running can be expressed in different styles. Customers who like a friendly assistant can receive a gentle reminder like "Would you like to go for a run with me?", whilst in contrast, individuals who prefer harsher prompts may get "Go running now, otherwise you will be a loser". Another application can be producing news texts of the same facts but different political tones (e.g. democratic and republican).

3. Related Work

Style transfer with non-parallel text corpus has become an active research area due to the recent advances in text generation tasks. Hu et al. [6] use variational auto-encoders with a discriminator to generate sentences with controllable attributes. The method learns a disentangled latent representation and generates a sentence from it using a code.

Shen et al. [16] propose a theoretical analysis of style transfer with non-parallel text corpus, presenting cross-alignment auto-encoders with discriminators architecture to generate sentences. Sentiment and word decipherment are the emphases for style transfer experiments in the study.

Fu et al. [3] suggest two models for style transfer. The first method uses multiple decoders for each category of style, whereas the second model uses style embeddings to augment the encoded representations, thus only one decoder is required to be learned to generate outputs in various styles. Scientific paper and newspaper titles are evaluated in style transfer, and reviews are evaluated in sentiment.

Our research will differ from the previous models which heavily depend on the untractable GANs. We propose a simple and effective approach finding matches between two corpora and training translation models to iteratively refine it. Different from operating in the space of given non-parallel data, our algorithm starts by identifying close sentence pairs across a variety of corpora based on lexical similarity and then utilizes the translation model's outputs to refine this data set iteratively.

4. Scope

We focus on three tasks in style transfer: sentiment, formality and political slants. There are three benchmark data sets corresponding to each task. We aim at developing an algorithm with comparable results to the start-of-the-art approaches proposed by Li et. al [7] with respect to three criteria, content preservation, fluency, and style correctness.

5. Theoretical Background

5.1. LSTM Model

Long Short-Term Memory (LSTM) cells [5] are a powerful means to capture semantic information of long sequences. They can approximate any time-dependent function $f(t)$ given by a number of function values. They have been used in a wide range of NLP applications, including machine translation and entailment recognition.

Specifically, an LSTM network successively reads the input token x_t, internal state c_{t-1} and the visible state h_{t-1}, and generates the new states c_t and h_t:

$$\begin{aligned}
\vec{i}_t &= \sigma(W^i \vec{x}_t + U^i \vec{h}_{t-1} + b^i) \\
\vec{f}_t &= \sigma(W^f \vec{x}_t + U^f \vec{h}_{t-1} + b^f) \\
\vec{o}_t &= \sigma(W^o \vec{x}_t + U^o \vec{h}_{t-1} + b^o) \\
\vec{z}_t &= \tanh(W^z \vec{x}_t + U^z \vec{h}_{t-1} + b^z) \\
\vec{c}_t &= \vec{i}_t \odot \vec{z}_t + f_t \odot \vec{c}_{t-1} \\
\vec{h}_t &= \vec{o}_t \odot \tanh(\vec{c}_t)
\end{aligned}$$

where i, f and o are input, forget and output gates, respectively. Therefore, they can adaptively read or forget information from internal memory states. The last hidden state $\vec{h}_T$ is usually taken as the representation of a sentence x.

5.2. General Adversarial Networks (GANs)

A Generative Adversarial Network (GAN) involves Generator (G) and Discriminator (D) networks, whose purposes are to map random noise to samples and discriminate real and generated samples respectively [4]. Formally speaking, the GAN objective involves finding a Nash equilibrium to the following two-player min-max problem:

$$\min_G \max_D \mathbb{E}_{x \sim q_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(x)))], \tag{1}$$

where $\boldsymbol{z} \in \mathbb{R}^{d_z}$ is a latent variable drawn from distribution $p(\boldsymbol{z})$ such as $\mathcal{N}(0, I)$ or $\mathcal{U}[-1, 1]$. G and D are usually convolutional neural networks for images [11], or instead LSTMs for texts.

6. Methodology

Given two corpora $X_1 = \{x_1^{(1)}, \cdots, x_1^{(n)}\}$ of style s_1 and $X_2 = \{x_2^{(1)}, \cdots, x_2^{(m)}\}$ of style s_2, our task is to learn a model that takes a sentence x_1 from style s_1 and transfers it into style s_2 while preserving its content. Note that X_1 and X_2 are not parallel, so we do not see any example rewrites in training data and have to learn them in an unsupervised way.

We have devised the following method to iteratively align the two corpora as shown in Algorithm 1. Between every two mono-style corpora, there are potentially many sentence pairs from different styles (x_1, x_2) that have similar content. For example, in the sentiment modification task, where X_1 and X_2 are negative and positive reviews respectively, there are a number of existing sentences that comment on the same topic with a different sentiment, e.g. "The sandwich is awful" and "The sandwich is great"; "I'll never come back" and "I would definitely come back". The first step is to construct pseudo pairs by some unsupervised similarity measurement. A simple approach to try first is match sentence $x_1 \in X_1$ with $x_2 \in X_2$ by TF-IDF lexical distance. For more complicated semantic matching of sentences, more neural networks-based methods need to explored. Based on the pseudo pairs, a machine translation model will be applied to these aligned texts. This trained model can be iteratively utilized to construct more accurate pseudo pairs. Specifically, for each x_1 we find the nearest neighbor of $M^{(t)}(x_1)$ from X_2, denoted as $\hat{x}_2$, and pair x_1 with $\hat{x}_2$ to be the training data for the next iteration. We can set $M^{(0)}$ as an identity function so that in the beginning we are using the original x_1 for matching.

Algorithm 1 Iterative Matching and Translation for Style Transfer

Input: Two corpora of different styles X_1, X_2
  $M^{(0)} \leftarrow$ identity function, $D \leftarrow \{\}$
  for $t = 1, \cdots, T$ **do**
    for $x_1 \in X_1$ **do**
      $\hat{x}_2 \leftarrow \arg\max_{x_2} \text{Sim}(M^{(t-1)}(x_1), x_2)$
      if $\text{Sim}(M^{(t-1)}(x_1), \hat{x}_2) >$ threshold **then**
        $D[x_1] \leftarrow \hat{x}_2$
      end if
    end for
    Train a machine translation model $M^{(t)}$ on D
  end for
Output: Style transfer model $M^{(T)} : \mathcal{X}_1 \rightarrow \mathcal{X}_2$ =0

7. Feasibility assessment

Pilot experiment with a rudimentary version our algorithm has already been conducted on one of the benchmark data sets, Yelp reviews. The results are comparable to the

state-of-the-art methods, indicating the high viability of this project. More careful design of the algorithm will be the main technical focus of this project.

A more tangible deliverable of the project is the application of style transfer algorithms to more AI platforms, for example stylish dialogue systems. This goal can be achieved by implementing existing approaches in language style transfer. Adapting these approaches is a problem of engineering and refinement, which is highly attainable.

8. Division of work

Jin	Design algorithms Further literature review Implement our model
Zhou	Implement baseline methods Conduct error analysis Design application-oriented adaptations
Lui	Implement baseline methods Collect and clean data sets Set up evaluation on Amazon Mechanical Turk Implement the adaptations

9. Schedule

Oct, 2018	Collect and preprocess data sets (formality, political slants); implement three baseline methods from literature reviews; configure Amazon Mechanical Turk account
Nov, 2018	Refine algorithm; conduct error type analysis; review literature on other domains work (including computer vision, reinforcement learning); Complete human intelligence tasks with Amazon Mechanical Turk
Dec, 2018	**Milestone: Deliver an unsupervised learning algorithm with style transfer in formality, political leaning or sentiment**; Write up the first report
Feb, 2019	Collect dialogue data set (e.g. Semantic Web Interest Group IRC Chat Logs); explore unsupervised methods on large language corpus e.g. different translations of books
Mar, 2019	**Milestone: Make broader application-oriented adaptations in three contexts, including human writing, understanding politics and sentiments in digital assistants**
Apr, 2019	Prepare presentations and exhibition

10. Deliverables

We will deliver an novel algorithm that takes a sentence x_0 with content c_0 and style s_0 as input and outputs another sentence x_1 with the same content $c_1 = c_0$ but of a different style $s_1 \neq s_0$, where styles s_0, s_1 may refer to positive and negative comment styles, formal and informal tones, or democratic and republican political slants. Ideally, the algorithm will outperform most state-of-the-art models, in terms of three criteria: content preservation, fluency and style correctness.

Another deliverable is the application to three real-life cases. Firstly, our algorithm's ability to transfer formality style can be used to develop an application that converts a casual writing into formal texts. Secondly, with the political tone transfer functionality, we will be able to generate news texts of distinctive political stances. This would help providing diverse perspectives to news readers. Thirdly, the algorithm can provide existing digital assistants or chat bots with more human-like dialogue styles (e.g. humorous, romantic).

11. Prerequisites

Item	Budget	Description
GPU Server	N/A	For intensive computation
Mechanical Turk	HKD 3000	For human evaluation

12. Risks and mitigation

The main technical difficulty of this project is the design of a novel algorithm. Adopting an agile development practice, we assign two months' time for the development phase (Oct - Nov 2018) and divide the development into multiple phases. Each phase will be around two weeks, encompassing planning, designing, implementation, testing and error analysis. Fewer iterations will be done in case of blockers.

To be in accordance with the FYP schedule, we limit our scope to three tasks, which are sentiment modification, formality switching and political slant transfer. These three tasks are well-defined and of high feasibility as evidenced by related works.

13. Summary

Language style transfer has been previously trained using adversarial methods. In this project, we will propose and evaluate a new method, iterative semantic matching, and translation for style transfer. Our compelling algorithm aims at outperforming the state-of-the-art approaches and applied into practical cases.

14. Acknowledgements

We thank our supervisor Professor Benjamin Kao for guiding us through our thinking process in the planning phase, which kept us on the right track. *His helpful advice and generous support in providing us with directions as well as necessary hardware* have given our group a promising start of this final year project.

References

[1] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017. 3

[2] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 7, 2017. 3

[3] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan. Style transfer in text: Exploration and evaluation. *arXiv preprint arXiv:1711.06861*, 2017. 3

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 4

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[6] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596, 2017. 3

[7] J. Li, R. Jia, H. He, and P. Liang. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*, 2018. 3, 4

[8] X. Niu, M. Martindale, and M. Carpuat. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, 2017. 3

[9] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*, 2018. 3

[10] E. Rabinovich, S. Mirkin, R. N. Patel, L. Specia, and S. Wintner. Personalized machine translation: Preserving original author traits. *arXiv preprint arXiv:1610.05461*, 2016. 3

[11] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 4

[12] S. Rao and J. Tetreault. Dear sir or madam, may i introduce the yafc corpus: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*, 2018. 3

[13] S. Reddy and K. Knight. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, 2016. 3

[14] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. 3

[15] R. Sennrich, B. Haddow, and A. Birch. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, 2016. 3

[16] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6833–6844, 2017. 3

[17] R. Voigt, D. Jurgens, V. Prabhakaran, D. Jurafsky, and Y. Tsvetkov. Rtgender: A corpus for studying differential responses to gender. In *LREC*, 2018. 3