(Male, Bachelor) and (Female, Ph.D) have different connotations: Parallelly Annotated Stylistic Language Dataset with Multiple Personas

Dongyeop Kang **Varun Gangal** **Eduard Hovy**
Carnegie Mellon University, Pittsburgh, PA, USA
{dongyeok,vgangal,hovy}@cs.cmu.edu

Abstract

Stylistic variation in text needs to be studied with different aspects including the writer's personal traits, interpersonal relations, rhetoric, and more. Despite recent attempts on computational modeling of the variation, the lack of parallel corpora of style language makes it difficult to systematically control the stylistic change as well as evaluate such models. We release PASTEL, the parallel and annotated stylistic language dataset, that contains $\approx$ 41K parallel sentences (8.3K parallel stories) annotated across different personas. Each persona has different styles in conjunction: gender, age, country, political view, education, ethnic, and time-of-writing. The dataset is collected from human annotators with solid control of input denotation: not only preserving original meaning between text, but promoting stylistic diversity to annotators. We test the dataset on two interesting applications of style language, where PASTEL helps design appropriate experiment and evaluation. First, in predicting a target style (e.g., male or female in gender) given a text, multiple styles of PASTEL make other external style variables controlled (or fixed), which is a more accurate experimental design. Second, a simple supervised model with our parallel text outperforms the unsupervised models using non-parallel text in style transfer. Our dataset is publicly available[1].

1 Introduction

Hovy (1987) claims that appropriately varying the style of text often conveys more information than is contained in the literal meaning of the words. He defines the roles of styles in text variation by pragmatics aspects (e.g., relationship between them) and rhetorical goals (e.g., formality), and provides example texts of how they are tightly coupled in practice. Similarly, Biber (1991) categorizes components of conversational situation by participants' characteristics such as their roles, personal characteristics, and group characteristics (e.g., social class). Despite the broad definition of style, this work mainly focuses on one specific aspect of style, *pragmatics aspects in group characteristics of speakers*, which is also called *persona*. Particularly, we look at multiple types of group characteristics in conjunction, such as gender, age, education level, and more.

Stylistic variation in text primarily manifest themselves at the different levels of textual features: lexical features (e.g., word choice), syntactic features (e.g., preference for the passive voice) and even pragmatics, while preserving the original meaning of given text (DiMarco and Hirst, 1990). Connecting such textual features to someone's persona is an important study to understand stylistic variation of language. For example, do highly educated people write longer sentences (Bloomfield, 1927)? Are Hispanic and East Asian people more likely to drop pronouns (White, 1985)? Are elder people likely to use lesser anaphora (Ulatowska et al., 1986)?

To computationally model a meaning-preserved variance of text across styles, many recent works have developed systems that transfer styles (Reddy and Knight, 2016; Hu et al., 2017; Prabhumoye et al., 2018) or profiles authorships from text (Verhoeven and Daelemans, 2014; Koppel et al., 2009; Stamatatos et al., 2018) without parallel corpus of stylistic text. However, the absence of such a parallel dataset makes it difficult both to systematically learn the textual variation of multiple styles as well as properly evaluate the models.

In this paper, we propose a *large scale, human-annotated, parallel stylistic dataset* called PASTEL, with focus on *multiple types of personas in conjunction*. Ideally, annotations for a parallel

[1] https://github.com/dykang/PASTEL

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1696–1706, Hong Kong, China, November 3–7, 2019.

style dataset should preserve the original meaning (i.e., *denotation*) between reference text and stylistically transformed text, while promoting diversity for annotators to allow their own styles of persona (i.e., *connotation*). However, if annotators are asked to write their own text given a reference sentence, they may simply produce arbitrarily paraphrased output which does not exhibit a stylistic diversity. To find such a proper input setting for data collection, we conduct a denotation experiment in §3. PASTEL is then collected by crowd workers based on the most effective input setting that balances both meaning preservation and diversity metrics (§4).

PASTEL includes stylistic variation of text at two levels of parallelism: ≈8.3K annotated, parallel stories and ≈41K annotated, parallel sentences, where each story has five sentences and has 2.63 annotators on average. Each sentence or story has the seven types of persona styles in conjunction: gender, age, ethnics, countries to live, education level, political view, and time of the day.

In §5, we introduce two interesting applications of style language using PASTEL: controlled style classification and supervised style transfer. The former application predicts a category (e.g., male or female) of target style (i.e., gender) given a text. Multiplicity of persona styles in PASTEL makes other style variables controlled (or fixed) except the target, which is a more accurate experimental design. In the latter, contrast to the unsupervised style transfer using non-parallel corpus, simple supervised models with our parallel text in PASTEL achieve better performance, being evaluated with the parallel, annotated text.

We hope PASTEL sheds light on the study of stylistic language variation in developing a solid model as well as evaluating the system properly.

2 Related Work

Transferring styles between text has been studied with and without parallel corpus:

Style transfer without parallel corpus: Prior works transfer style between text on single type of style aspect such as sentiment (Fu et al., 2018; Shen et al., 2017; Hu et al., 2017), gender (Reddy and Knight, 2016), political orientation (Prabhumoye et al., 2018), and two conflicting corpora (e.g, paper and news (Han et al., 2017), or real and synthetic reviews (Lipton et al., 2015)). They use different types of generative models in the same way as style transfer in images, where meaning preservation is not controlled systematically. Prabhumoye et al. (2018) proposes back-translation to get a style-agnostic sentence representation. However, they lack parallel ground truth for evaluation and present limited evaluation for meaning preservation.

Style transfer with parallel corpus: Few recent works use parallel text for style transfer between modern and Shakespearean text (Jhamtani et al., 2017), sarcastic and literal tweets (Peled and Reichart, 2017), and formal and informal text (Heylighen and Dewaele, 1999; Rao and Tetreault, 2018). Compared to these, we aim to understand and demonstrate style variation owing to multiple demographic attributes.

Besides the style transfer, other applications using stylistic features have been studied such as poetry generation (Ghazvininejad et al., 2017), stylometry with demographic information (Verhoeven and Daelemans, 2014), modeling style bias (Vogel and Jurafsky, 2012) and modeling biographic attributes (Garera and Yarowsky, 2009). A series of works by (Koppel et al., 2011, 2009; Argamon et al., 2009; Koppel and Winter, 2014) and their shared tasks (Stamatatos et al., 2018) show huge progress on author profiling and attribute classification tasks. However, none of the prior works have collected a stylistic language dataset to have multiple styles in conjunction, parallely annotated by a human. The multiple styles in conjunction in PASTEL enable an appropriate experiment setting for controlled style classification task in Section 5.1.

3 Denotation Experiment

We first provide a preliminary study to find the best input setting (or denotation) for data collection to balance between two trade-off metrics: meaning preservation and style diversity.

3.1 Preliminary Study

Table 1 shows output texts produced by annotators given different input denotation settings. The basic task is to provide an input denotation (e.g., a sentence only, a sequence of images) and then ask them to reproduce text maintaining the meaning of the input but with their own persona.

For instance, if we provide a *single reference sentence*, annotators mostly repeat the input text with a little changes of the lexical terms. This

Denotation:	**Produced sentences**:
single ref. sentence	the old door with wood was the only direction to the courtyard
story(imgs)	The old wooden door in the stonewall looks like a portal to a fairy tale.
story(imgs.+keyw words)	Equally so, he is intrigued by the heavy wooden door in the courtyard.
Reference sentence: the old wooden door was only one way into the courtyard.	

Table 1: Textual variation across different denotation settings. Each sentence is produced by a same annotator. Note that providing reference sentence increases fidelity to the reference while decreases diversity.

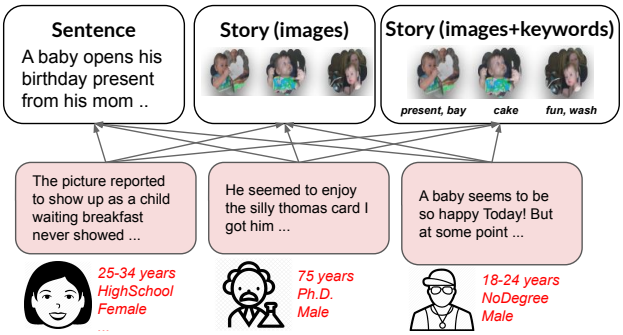


Figure 1: Denotation experiment finds the best input setting for data collection, that preserves meaning but diversifies styles among annotators with different personas.

setup mostly preserves the meaning by simply paraphrasing the sentence, but annotators' personal style does not reflect the variation. With a *single image*, on the other hand, the outputs produced by annotators tend to be diverse. However, the image can be explained with a variety of contents, so the output meaning can drift away from the reference sentence.

If a series of consistent images (i.e., a story) is given, we expect a stylistic diversity can be more narrowed down, by grounding it to a specific event or a story. In addition to that, some keywords added to each image of a story help deliver more concrete meaning of content as well as the style diversity.

3.2 Experimental Setup

In order to find the best input setting that preserves meaning as well as promotes a stylistic diversity, we conduct a denotation experiment as described in Figure 1. The experiment is a subset of our original dataset, which have only 100 samples of annotations.

A basic idea behind this setup is to provide (1) a perceptually common denotation via sentences or images so people share the same context (i.e., denotation) given, (2) a series of them as a "story" to limit them into a specific event context, and (3) two modalities (i.e., text and image) for better disambiguation of the context by grounding them to each other.

We test five different input settings[2]: *Single reference sentence*, ***Story*** *(images)*, ***Story*** *(images)* + *global keywords*, ***Story*** *(images + local keywords)*, and ***Story*** *(images + local keywords + ref. sentence)*.

For the keyword selection, we use RAKE algorithm (Rose et al., 2010) to extract keywords and rank them for each sentence by the output score. Top five uni/bigram keywords are chosen at each story, which are called *global keywords*. On the other hand, another top three uni/bigram keywords are chosen at each image/sentence in a story, which are called *local keywords*. Local keywords for each image/sentence help annotators not deviate too much. For example, *local keywords* look like *(restaurant, hearing, friends)* $\rightarrow$ *(pictures, menu, difficult)* $\rightarrow$ *(salad, corn, chose)* for three sentences/images, while *global keywords* look like *(wait, salad, restaurant)* for a story of the three sentences/images.

We use Visual Story Telling (ViST) (Huang et al., 2016) dataset as our input source. The dataset contains stories, and each story has five pairs of images and sentences. We filter out stories that are not temporally ordered using the timestamps of images. The final number of stories after filtering the non-temporally-ordered stories is 28,130. For the denotation experiment, we only use randomly chosen 100 stories. The detailed pre-processing steps are described in Appendix.

3.3 Measuring Meaning Preservation & Style Diversity across Different Denotations

For each denotation setting, we conduct a quantitative experiment to measure the two metrics: meaning preservation and style diversity. The two metrics pose a trade-off to each other. The best input setting then is one that can capture both in appropriate amounts. For example, we want meaning of the input preserved, while lexical or syn-

[2]Other settings like *Single reference image* are tested as well, but they didn't preserve the meaning well.

	denotation settings	**Style Diversity**	**Meaning Preservation**	
		E(GM)	METEOR	VectorExtrema
sentence-level	*single ref. sentence*	<u>2.98</u>	**0.37**	**0.70**
	***story**(images)*	2.86	0.07	0.38
	***story**(images) + global keywords*	2.85	0.07	0.39
	***story**(images + local keywords)*	**3.07**	0.17	<u>0.53</u>
	***story**(images + local keywords + ref. sentence)*	2.91	<u>0.21</u>	0.43
story-level	***story**(images)*	4.43	0.1	0.4
	***story**(images) + global keywords*	4.43	0.1	0.42
	***story**(images + local keywords)*	**4.58**	<u>0.19</u>	**0.55**
	***story**(images + local keywords + ref. sentence)*	<u>4.48</u>	**0.22**	<u>0.44</u>

Table 2: Denotation experiment to find the best input setting (i.e., meaning preserved but stylistically diverse). **story-level** measures the metrics for five sentences as a story, and **sentence-level** per individual sentence. Note that *single reference sentence* setting only has sentence level. For every metrics in both meaning preservation and style diversity, the higher the better. The **bold** number is the highest, and the <u>underlined</u> is the second highest.

Figure 2: Final denotation setting for data collection: an event that consists of a series of five images with a handful number of keywords. We ask annotators to produce text about the event for each image.

tactic features (e.g., POS tags) can vary depending on annotator's persona. We use the following automatic measures for the two metrics:

Style Diversity measures how much produced sentences (or stories) differ amongst themselves. Higher the diversity, better the stylistic variation in language it contains. We use an entropy measure to capture the variance of n-gram features between annotated sentences: Entropy (Gaussian-Mixture) that combines the N-Gram entropies (Shannon, 1951) using Gaussian mixture model (N=3).

Meaning Preservation measures semantic similarity of the produced sentence (or story) with the reference sentence (or story). Higher the similarity, better the meaning preserved. We use a hard-measure, METEOR (Banerjee and Lavie, 2005), that calculates F-score of word overlaps between the output and reference sentences[3]. Since the hard measures do not take into account all semantic similarities [4], we also use a soft measure, VectorExtrema (VecExt) (Liu et al., 2016). It computes cosine similarity of averaged word embeddings (i.e., GloVe (Pennington et al., 2014)) between the output and reference sentences.

Table 2 shows results of the two metrics across different input settings we define. For the sentence level, as expected, *single reference sentence* has the highest meaning preservation across all the metrics because it is basically paraphrasing the reference sentence. In general, ***Story*** (*images + local keywords*) shows a great performance with the highest diversity regardless of the levels, as well as the highest preservation at the soft measure on the story-level. Thus, we use ***Story***(*images+local keywords*) as the input setting for our final data collection, which has the most balanced performance on both metrics. Figure 2 shows an example of our input setting for crowd workers.

4 PASTEL: A Parallelly Annotated Dataset for Stylistic Language Dataset

We describe how we collect the dataset with human annotations and provide some analysis on it.

[3]Other measures (e.g., BLEU (Papineni et al., 2002), ROUGE (Lin and Hovy, 2003)) show relatively similar performance.

[4]METEOR does consider synonymy and paraphrasing but is limited by its predefined model/dictionaries/resources for the respective language, such as Wordnet

4.1 Annotation Schemes

Our crowd workers are recruited from the Amazon Mechanical Turk (AMT) platform. Our annotation scheme consists of two steps: (1) ask annotator's demographic information (e.g., gender, age) and (2) given an input denotation like Figure 2, ask them to produce text about the denotation with their own style of persona (i.e., connotation).

In the first step, we use seven different types of persona styles; *gender*, *age*, *ethnic*, *country*, *education level*, and *political orientation*, and one additional context style *time-of-day* (tod). For each type of persona, we provide several categories for annotators to choose. For example, *political orientation* has three categories: Centrist, Left Wing, and Right Wing. Categories in other styles are described in the next sub-section.

In the second step, we ask annotators to produce text that describes the given input of denotation. We again use the pre-processed ViST (Huang et al., 2016) data in §3 for our input denotations. To reflect annotators' persona, we explicitly ask annotators to reflect their own persona in the stylistic writing, instead of pretending others' persona. We attach detailed annotation schemes at Figure 5 in Appendix.

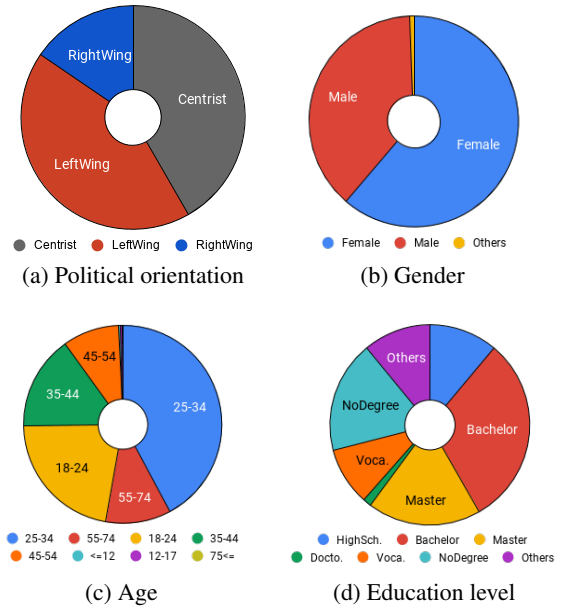
To amortize both costs and annotators' effort at answering questions, each HIT requires the participants to annotate three stories after answering demographic questions. One annotator was paid $0.11 per HIT. For English proficiency, the annotators were restricted to be from USA or UK. A total 501 unique annotators participated in the study. The average number of HIT per annotator was 9.97.

	Number of Sentences	Number of Stories
Train	33,240	6,648
Valid	4,155	831
Test	4,155	831
total	41,550	8,310

Table 3: Data statistics of the PASTEL.

Once we complete our annotations, we filter out noisy responses such as stories with missing images and overtly short sentences (i.e., minimum sentence length is 5). The dataset is then randomly split into train, valid, and test set by 0.8, 0.1, and 0.1 ratios, respectively. Table 3 shows the final number of stories and sentences in our dataset.

4.2 Analysis and Examples



(a) Political orientation (b) Gender

(c) Age (d) Education level

Figure 3: Distribution of annotators for each personal style in PASTEL. Best viewed in color.

Figure 3 shows demographic distributions of the annotators. Education-level of annotators is well-balanced, while gender and political view are somewhat biased (e.g., 68% of annotators are Female, only 18.6% represent themselves as right-wing). Table 8 in Appendix includes the categories in other styles and their distributions.

Table 4 shows few examples randomly chosen from our dataset: two at sentence level (top, middle) and one at story level (bottom). Due to paucity of space, we only show a few types of persona styles. For example, we observe that Education level (e.g., NoDegree vs. Graduate) actually reflects a certain degree of formality in their writing at both sentence and story levels. In §5.1, we conduct an in-depth analysis of textual variation with respect to the persona styles in PASTEL.

5 Applications with PASTEL

PASTEL can be used in many style related applications including style classification, stylometry (Verhoeven and Daelemans, 2014), style transfer (Fu et al., 2018), visually-grounded style transfer, and more. Particularly, we chose two applications, where PASTEL helps design appropriate experiment and evaluation: controlled style classification (§5.1) and supervised style transfer (§5.2).

Reference Sentence: went to an art museum with a group of friends.	
edu:HighSchoolOrNoDiploma	My friends and I went to a art museum yesterday .
edu:Bachelor	I went to the museum with a bunch of friends.

Reference Sentence: the living room of our new home is nice and bright with natural light.	
edu:NoDegree, *gender*:Male	The natural lightning made the apartment look quite nice for the upcoming tour .
edu:Graduate, *gender*:Female	The house tour began in the living room which had a sufficient amount of natural lighting.

Reference Story: Went to an art museum with a group of friends . We were looking for some artwork to purchase, as sometimes artist allow the sales of their items . There were pictures of all sorts , but in front of them were sculptures or arrangements of some sort . Some were far out there or just far fetched . then there were others that were more down to earth and stylish. this set was by far my favorite.very beautiful to me .

edu:HighSchool, *ethnic*:Caucasian, *gender*:Female	My friends and I went to a art museum yesterday . There were lots of puchases and sales of items going on all day . I loved the way the glass sort of brightened the art so much that I got all sorts of excited . After a few we fetched some grub . My favorite set was all the art that was made out of stylish trash .
edu:Bachelor, *ethnic*:Caucasian, *gender*:Female	I went to the museum with a bunch of friends . There was some cool art for sale . We spent a lot of time looking at the sculptures . This was one of my favorite pieces that I saw . We looked at some very stylish pieces of artwork .

Table 4: Two sentence-level (top, middle) and one story-level (bottom) annotations in PASTEL. Each text produced by an annotator has their own persona values (underline) for different types of styles (italic). Note that the reference sentence (or story) is given for comparison with the annotated text. Note that misspellings of the text are made by annotators.

5.1 Controlled Style Classification

A common mistake in style classification datasets is not controlling external style variables when predicting the category of the target style. For example, when predicting a gender type given a text $P(\textit{gender}=\underline{Male}|\texttt{text})$, the training data is only labeled by the target style *gender*. However, the `text` is actually produced by a person with not only *gender*=$\underline{Male}$ but also other persona styles such as *age*=55-74 or *education*=HighSchool. Without controlling the other external styles, the classifier is easily biased against the training data.

We define a task called *controlled style classification* where all other style variables are fixed[5], except one to classify. Here we evaluate (1) which style variables are relatively difficult or easy to predict from the text given, and (2) what types of textual features are salient for each type of style classification.

Features. Stylistic language has a variety of features at different levels such as lexical choices, syntactic structure and more. Thus, we use following features:

- **lexical** features: ngram's frequency (n=3), number of named entities, number of stop-words
- **syntax** features: sentence length, number of each Part-of-Speech (POS) tag, number of out-of-vocabulary, number of named entities
- **deep** features: pre-trained sentence encoder using BERT (Devlin et al., 2019)
- **semantic** feature: sentiment score

where named entities, POS tags, and sentiment scores are obtained using the off-the-shelf tools such as Spacy[6] library. We use $70K$ n-gram lexical features, 300 dimensional embeddings, and 14 hand-written features.

[5]The distribution of number of training instances per variable is given in Appendix

[6]https://spacy.io/

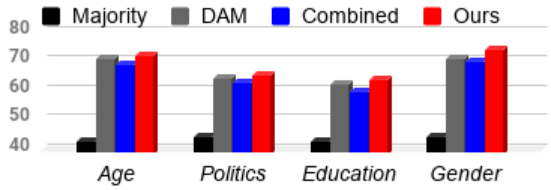
Models. We train a binary classifier for each personal style with different models: logistic regression, SVM with linear/RBF kernels, Random Forest, Nearest Neighbors, Multi-layer Perceptron, AdaBoost, and Naive Bayes. For each style, we choose the best classifiers on the validation. Their F-scores are reported in Figure 4. We use sklearn's implementation of all models (Pedregosa et al., 2011).[7] We consider various regularization parameters for SVM and logistic regression (e.g., c=[0.01, 0.1, 0.25, 0.5, 0.75, 1.0].

We use neural network based baseline models: deep averaging networks (DAN, Iyyer et al., 2015) of GloVe word embeddings (Pennington et al., 2014)[8]. We also compare with the non-controlled model (Combined) which uses a combined set of samples across all other variables except for one to classify using the same features we used.

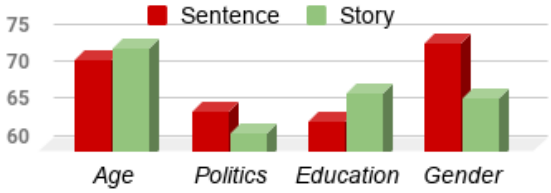
Setup. We tune hyperparameters using 5-fold cross validation. If a style has more than two categories, we choose the most conflicting two: *gender*:{Male, Female}, *age*: {18-24, 35-44}, *education*: {Bachelor, No Degree}, and *politics*: {LeftWing, RightWing}. To classify one style, all possible combinations of other styles ($2 * 2 * 2$=8) are separately trained by different models. We use the macro-averaged F-scores among the separately trained models on the same test set for every models.

Results. Figure 4 shows F-scores (a) among different styles and (b) between sentences and stories. In most cases, multilayer perceptron (MLP) outperforms the majority classifier and other models by large margins. Compared to the neural baselines and the combined classifier, our models show better performance. In comparison between controlled and combined settings, controlled setting achieves higher improvements, indicating that fixing external variables helps control irrelevant features that come from other variables. Among different styles, gender is easier to predict from the text than ages or education levels. Interestingly, a longer context (i.e., story) is helpful in predicting age or education, whereas not for political view and gender.

In our ablation test among the feature types, the combination of different features (e.g., lexical, syntax, deep, semantic) is very complementary and effective. Lexical and deep features are two most significant features across all style classifiers, while syntactic features are not.



(a) different style types



(b) sentences vs stories

Figure 4: Controlled style classification: F-scores on (a) different types of styles on sentences and on (b) our best models between sentences and stories. Best viewed in color.

Gender:Male	*Gender*:Female
PROPN, ADJ, #_ENTITY, went, party, SENT_LEN	happy, day, end, group, just, snow, NOUN
Politics:LeftWing	*Politics*:RightWing
female, time, NOUN, ADP, VERB, porch, day, loved	SENT_LENGTH, PROPN, #_ENTITY, n't, ADJ, NUM
Education:Bachelor	*Education*:NoDegree
food, went, #_STOPWORDS, race, ADP	!, just, came, love, lots, male, fun, n't, friends, happy
Age:18-24	*Age*:35-44
ADP, come, PROPN, day, ride, playing, sunset	ADV, did, town, went, NOUN, #_STOPWORDS

Table 5: Most salient lexical (lower cased) and syntactic (upper cased) features on story-level classification. Each feature is chosen by the highest coefficients in the logistic regression classifier.

Table 5 shows the most salient features for classification of each style. Since we can't interpret deep features, we only show lexical and syntactic features. The salience of features are ranked by coefficients of a logistic regression classifier. Interestingly, female annotators likely write more nouns and lexicons like 'happy', while male annotators likely use pronouns, adjectives, and named

[7] http://scikit-learn.org/stable/

[8] Other architectures such as convolutional neural networks (CNN, Zhang et al., 2015) and recurrent neural networks (LSTM, Hochreiter and Schmidhuber, 1997) show comparable performance as DAN.

entities. Annotators on left wing prefer to use 'female', nouns and adposition, while annotators on right wing prefer shorter sentences and negative verbs like 'n't'. Not many syntactic features are observed from annotators without degrees compared to with bachelor degree.

5.2 Supervised Style Transfer

The style transfer is defined as $(S, \alpha) \rightarrow \hat{S}$: We attempt to alter a given source sentence S to a given target style α. The model generates a candidate target sentence $\hat{S}$ which preserves the meaning of S but is more faithful to the target style α so being similar to the target annotated sentence $\bar{S}_\alpha$. We evaluate the model by comparing the predicted sentence $\hat{S}$ and target annotated sentence $\bar{S}_\alpha$. The sources are from the original reference sentences, while the targets are from our annotations.

Models. We compare five different models:

- AsItIs: copies over the source sentence to the target, without any alterations.
- WordDistRetrieve: retrieves a training source-target pair that has the same target style as the test pair and is closest to the test source in terms of word edit distance (Navarro, 2001). It then returns the target of that pair.
- EmbDistRetrieve: Similar to WordDistRetrieve, except that a continuous bag-of-words (CBOW) is used to retrieve closest source sentence instead of edit distance.
- Unsupervised: use unsupervised style transfer models using Variational Autoencoder (Shen et al., 2017) and using additional objectives such as cross-domain and adversarial losses (Lample et al., 2017)[9]. Since unsupervised models can't train multiple styles at the same time, we train separate models for each style and macro-average their scores at the end. In order not to use the parallel text in PASTEL, we shuffle the training text of each style.
- Supervised: uses a simple attentional sequence-to-sequence (S2S) model (Bahdanau et al., 2014) extracting the parallel text from PASTEL. The model jointly trains different styles in conjunction by concatenating them to the source sentence at the beginning.

We avoid more complex architectural choices for Supervised models like adding a pointer component or an adversarial loss, since we seek to establish a minimum level of performance on this dataset.

[9]We can't directly compare with Hu et al. (2017); Prabhumoye et al. (2018) since their performance highly depends on the pre-trained classifier that often shows poor performance.

Setup. We experiment with both Softmax and Sigmoid non-linearities to normalize attention scores in the sequence-to-sequence attention. Adam (Kingma and Ba, 2014) is used as the optimizer. Word-level cross entropy of the target is used as the loss. The batch size is set to 32. We pick the model with lowest validation loss after 15 training epochs. All models are implemented in PyTorch (Paszke et al., 2017).

For an evaluation, in addition to the same hard and soft metrics used for measuring the meaning preservation in §3, we also use BLEU_2 (Papineni et al., 2002) for unigrams and bigrams, and ROUGE (Lin and Hovy, 2003) for hard metric and Embedding Averaging (EA) similarity (Liu et al., 2016) for soft metric.

	Hard $(\hat{S},\bar{S}_\alpha)$			Soft $(\hat{S},\bar{S}_\alpha)$	
Models: $(S, \alpha) \rightarrow \hat{S}$	$\mathbf{B_2}$	**M**	**R**	**EA**	**VE**
AsItIs	**35.41**	**12.38**	21.08	0.649	0.393
WordDistRetrieve	30.64	7.27	22.52	0.771	0.433
EmbDistRetrieve	33.00	8.29	24.11	0.792	0.461
Unsupervised					
· Shen et al. (2017)	23.78	7.23	21.22	0.795	0.353
· Lample et al. (2017)	24.52	6.27	19.79	0.702	0.369
Supervised					
· S2S	26.78	7.36	25.57	0.773	0.455
· S2S+GloVe	31.80	10.18	29.18	0.797	0.524
· S2S+GloVe+PreTr.	31.21	10.29	**29.52**	**0.804**	**0.529**

Table 6: Supervised style transfer. GloVe initializes with pre-trained word embeddings. PreTr. denotes pre-training on YAFC. Hard measures are $\mathbf{BLEU_2}$, **M**ETEOR, and **R**OUGE, and soft measures are **E**mbeding**A**veraging and **V**ector**E**xtrema.

Results. Table 6 shows our results on style tranfer. We observe that initializing both en/decoder's word embeddings with GloVe (Pennington et al., 2014) improves model performance on most metrics. Pretraining (PreTr.) on the formality style transfer data YAFC (Rao and Tetreault, 2018) further helps performance. All supervised S2S approaches outperform both retrieval-based baselines on all measures. This illustrates that the performance scores achieved are not simply a result of memorizing the training set. S2S methods surpass AsItIs on both soft measures and ROUGE. The significant gap that remains on BLEU remains

Source (S): I'd never seen so many beautiful flowers.
Style (α): (Morning, HighSchool) $S + \alpha \rightarrow \hat{S}$: the beautiful flowers were beautiful. $\bar{S}_\alpha$: the flowers were in full bloom. **Style** (α): (Afternoon, NoDegree) $S + \alpha \rightarrow \hat{S}$: The flowers were very beautiful. $\bar{S}_\alpha$: Tulips are one of the magnificent varieties of flowers.
Source (S): she changed dresses for the reception and shared food with her new husband.
Style (α): (Master, Centrist) $S + \alpha \rightarrow \hat{S}$: The woman had a great time with her husband $\bar{S}_\alpha$: Her husband shared a cake with her during reception
Style (α): (Vocational, Right) $S + \alpha \rightarrow \hat{S}$: The food is ready for the reception $\bar{S}_\alpha$: The new husband shared the cake at the reception

Table 7: Examples of style transferred text by our supervised model (S2S+GLOVE+PRETR.) on PASTEL. Given source text (S) and style (α), the model predicts a target sentence $\hat{S}$ compared to annotated target sentence $\bar{S}_\alpha$.

a point of exploration for future work. The significant improvement against the unsupervised methods (Shen et al., 2017; Lample et al., 2017) indicates the usefulness of the parallel text in PASTEL.

Table 7 shows output text $\hat{S}$ produced by our model given a source text S and a style α. We observe that the output text changes according to the set of styles.

6 Conclusion and Future Directions

We present PASTEL, a parallelly annotated stylistic language dataset. Our dataset is collected by human annotation using a proper denotation setting that preserves the meaning as well as maximizes the diversity of styles. Multiplicity of persona styles in PASTEL makes other style variables controlled (or fixed) except the target style for classification, which is a more accurate experimental design. Our simple supervised models with our parallel text in PASTEL outperforms the unsupervised style transfer models using non-parallel corpus. We hope PASTEL can be a useful benchmark to both train and evaluate models for style transfer and other related problems in text generation field.

We summarize some directions for future style researches:

- In our ablation study, salient features for style classification are not only syntactic or lexical features but also content words (e.g., love, food). This is a counterexample to the hypothesis implicit in much of recent style research: *style* needs to be *separately modeled* from *content*. We also observe that some texts remain similar across different annotator personas or across outputs from our transfer models, indicating that some content is stylistically invariant. Studying these and other aspects of the content-style relationship in PASTEL could be an interesting direction.
- Does any external variable co-varying with the text qualify to be a style variable/facet? What are the categories of style variables/facets? Do architectures which transfer well across one style variable (e.g gender) generalize to other style variables (e.g age)? We opine that these questions are largely overlooked by current style transfer work. We hope that our consideration of some of these questions in our work, though admittedly rudimentary, will lead to them being addressed extensively in future style transfer work.

Acknowledgements

This work would not have been possible without the ViST dataset and helpful suggestions with Ting-Hao Huang. We also thank Alan W Black, Dan Jurafsky, Wei Xu, Taehee Jung, and anonymous reviewers for their helpful comments.

References

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

Leonard Bloomfield. 1927. Literate and illiterate speech. *American speech*, 2(10):432–439.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Chrysanne DiMarco and Graeme Hirst. 1990. Accounting for style in machine translation. In *Third International Conference on Theoretical Issues in Machine Translation, Austin*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. *CoRR*, abs/1711.06861.

Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 710–718. Association for Computational Linguistics.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an Interactive Poetry Generation System. *Proceedings of ACL 2017, System Demonstrations*, pages 43–48.

Mengqiao Han, Ou Wu, and Zhendong Niu. 2017. Unsupervised Automatic Text Style Transfer using LSTM. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 281–292. Springer.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center Leo Apostel, Vrije Universiteit Brüssel*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596.

Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *NAACL*.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proc. of ACL-IJCNLP*, volume 1, pages 1681–1691.

Harsh Jhamtani, Varun Gangal, Eduard H. Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *CoRR*, abs/1707.01161.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45:83–94.

Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics.

Zachary C Lipton, Sharad Vikram, and Julian McAuley. 2015. Generative concatenative nets jointly learn to Write and Classify reviews. *arXiv preprint arXiv:1511.03683*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. 2017. Pytorch.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Lotem Peled and Roi Reichart. 2017. Sarcasm sign: Interpreting sarcasm with sentiment based monolingual machine translation. In *ACL*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. In *ACL*.

Sudha Rao and Joel R. Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *NAACL-HLT*.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, pages 1–20.

Claude E Shannon. 1951. Prediction and entropy of printed English. *Bell Labs Technical Journal*, 30(1):50–64.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6830–6841.

Efstathios Stamatatos, Francisco Rangel, Michael Tschuggnall, Benno Stein, Mike Kestemont, Paolo Rosso, and Martin Potthast. 2018. Overview of pan 2018. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 267–285. Springer.

Hanna K Ulatowska, Mari M Hayashi, Michael P Cannito, and Susan G Fleming. 1986. Disruption of reference in aging. *Brain and language*, 28(1):24–41.

Ben Verhoeven and Walter Daelemans. 2014. CLiPS Stylometry Investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *LREC 2014-NINTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION*, pages 3081–3085.

Adam Vogel and Dan Jurafsky. 2012. He said, she said: Gender in the ACL anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41. Association for Computational Linguistics.

Lydia White. 1985. The pro-drop parameter in adult second language acquisition. *Language learning*, 35(1):47–61.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proc. of NIPS*.