Formality Style Transfer for Noisy Text: Leveraging Out-of-Domain Parallel Data for In-Domain Training via POS Masking

Isak Czeresnia Etinger **Alan W. Black**
Language Technologies Institute
Carnegie Mellon University
{ice,awb}@cs.cmu.edu

Abstract

Typical datasets used for style transfer in NLP contain aligned pairs of two opposite extremes of a style. As each existing dataset is sourced from a specific domain and context, most use cases will have a sizable mismatch from the vocabulary and sentence structures of any dataset available. This reduces the performance of the style transfer, and is particularly significant for noisy, user-generated text. To solve this problem, we show a technique to derive a dataset of aligned pairs (style-agnostic vs stylistic sentences) from an unlabeled corpus by using an auxiliary dataset, allowing for in-domain training. We test the technique with the Yahoo Formality Dataset and 6 novel datasets we produced, which consist of scripts from 5 popular TV-shows (*Friends, Futurama, Seinfeld, Southpark, Stargate SG-1*) and the *Slate Star Codex* online forum. We gather 1080 human evaluations, which show that our method produces a sizable change in formality while maintaining fluency and context; and that it considerably outperforms OpenNMT's Seq2Seq model directly trained on the Yahoo Formality Dataset. Additionally, we publish the full pipeline code and our novel datasets [1].

1 Introduction

Typical datasets used for style transfer in NLP contain aligned pairs of two opposite extremes of a style (Hughes et al., 2012; Xu et al., 2012; Jhamtani et al., 2017; Carlson et al., 2017; Xu, 2017; Rao and Tetreault, 2018). Those datasets are useful for training neural networks that perform style transfer on text that is similar (both in vocabulary and structure) to the text in the datasets. However, as each of those datasets is sourced from a specific domain and context, in most use cases there is not an available dataset of parallel data with vocabulary and structure similar to the one requested.

This is especially significant for style transfer with noisy/user-generated text, where a mismatch is common even when the training dataset is also noisy/user-generated. We explore formality transfer specifically for noisy/user-generated text. To the best of our knowledge, the best dataset for this is currently the Yahoo Formality Dataset (Rao and Tetreault, 2018). However, this dataset is limited to few domains and to the context of Yahoo answers instead of other websites or in-person chat.

To overcome this problem, we propose a technique to derive a dataset of aligned pairs from an unlabeled corpus by using an auxiliary dataset; and we apply this technique to the task of formality transfer on noisy/user-generated conversations.

2 Related Work

Textual style transfer has been a large topic of research in NLP. Early research directly fed labeled, parallel data to train generic Seq2Seq models. Jhamtani et al. (2017) employed this technique on Shakespeare and modern literature. Carlson et al. (2017) employed it on bible translations.

More recent methods have tackled the problem of training models with unlabeled corpora. They seek to obtain latent representations that would correspond to stylistics and semantics separately, then change the stylistic representation while maintaining the semantic one. This can be done by one of 3 ways (Tikhonov and Yamshchikov, 2018): employing back-translation; training a stylistic discriminator; or embedding words or sentences and segmenting embedding state-space into semantic and stylistic sections. Our method differs from those works in many aspects.

Artetxe et al. (2017) worked on unsupervised machine translation. It differs from our objective

[1] https://github.com/ICEtinger/StyleTransfer

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 11–16
Hong Kong, Nov 4, 2019.

because it is translation instead of style transfer. Our work employs POS tags as a latent shared representation of syntactic structures and style-free semantics across sentences of different styles. This is not possible (or much less direct) across different languages.

Han et al. (2017) presented a Seq2Seq model that uses two switches with tensor product to control the style transfer in the encoding and decoding processes. Fu et al. (2018) proposed adversarial networks for the task of textual style transfer. Yang et al. (2018) presented a new technique that uses a target domain language model as the discriminator to improve training. Our method is modular with respect to the main Seq2Seq neural model, so it can more easily leverage state-of-the-art (Merity et al., 2017) new models, e.g. most recent versions of OpenNMT (Klein et al., 2017).

Shen et al. (2017) proposed a model that assumes a shared latent content distribution across different text corpora, and leverages refined alignment of latent representations to perform style transfer. Our method does not assume such shared latent content distribution across different corpora. We instead leverage shared latent content distribution across different styles of a same corpus.

Zhang et al. (2018) presented a Seq2Seq model architecture using shared and private model parameters to better train a model from multiple corpora of different domains. Our method is modular with respect to the main Seq2Seq neural model, and is trained with a single corpus each time.

Li et al. (2018) proposed a method that uses retrieval of training sentences (after a deletion operation) during inference time to improve sentence generation. Our method uses a similar inspiration of selecting the "deleted" terms, but instead of being deleted, they are replaced by a latent shared representation of syntactic structures and style-free semantics in the form of POS tags. Additionally, we employ a modular Seq2Seq neural model with the replaced representation instead of retrieving training sentences.

Prabhumoye et al. (2018) presented a method that uses back-translation in French to obtain a latent representation of sentences with less stylistic characteristics. That technique requires that the French translation be trained on a dataset with similar vocabulary and structure as the data on which style transfer is applied. Our work does not have this requirement. Additionally, that work fixes the encoder and decoder in order to employ the back-translation, while our work employs a modular Seq2Seq neural model to leverage state-of-the-art Seq2Seq neural models.

3 Technique for Dataset Generation

Consider an unlabeled corpus A and a labeled, parallel dataset B. We show a technique that uses B to derive a dataset A' of aligned pairs from A.

If B contains aligned pairs of sentences with styles s_1 and s_2, then one technique to generate A' is to train a classifier between s_1 and s_2 on B, then to use the classifier to select subsets A_1 and A_2 from A following each style, i.e:

$$A_i = \{x \in A | P(class(x) = s_i) > t\}, \quad t \text{ constant}$$

Then, to create parallel data from $\{A_1, A_2\}$, use the classifier to select the terms that have the most weight in determining the style of sentences (e.g.: if Logistic Regression, use term coefficients, select term with coefficients above a certain threshold). Call the set of those terms T. For each sentence $x \in A_1 \cup A_2$, map x with an altered sentence x' which is equal to x when all terms in x that are in T are replaced by their POS tags in x. The set of pairs $\{(x, x')\} = A'$ is now parallel data.

POS tags are employed as a latent shared representation of syntactic structures and style-free semantics across sentences of different styles.

4 Neural Network Models

After obtaining the dataset in the format $\{(x, x')\}$ as described in Section 3, we train a typical Seq2Seq model to predict x from x'. Then, on inference time, we apply the same transformation described in Section 3 to the test set (that may have different styles from the training set), and apply the model on that transformed test set.

For example, consider we have a classifier of two styles: `formal` and `informal`. We use the classifier to produce datasets A_{formal} and $A_{informal}$ from an unlabeled corpus A. From A_{formal}, we produce $\{(x, x')\}$, and use it to train a model that predicts $\{x\}$ from $\{x'\}$. Recall that x' is equal to x when all terms in x that are the most characteristic of formality are replaced by their POS tags in x. During inference time, we want to transform a neutral or an informal sentence y to formal. We derive a y' from y at the same way we did for x', but now we replace the terms most characteristic of informality by their POS tags. We

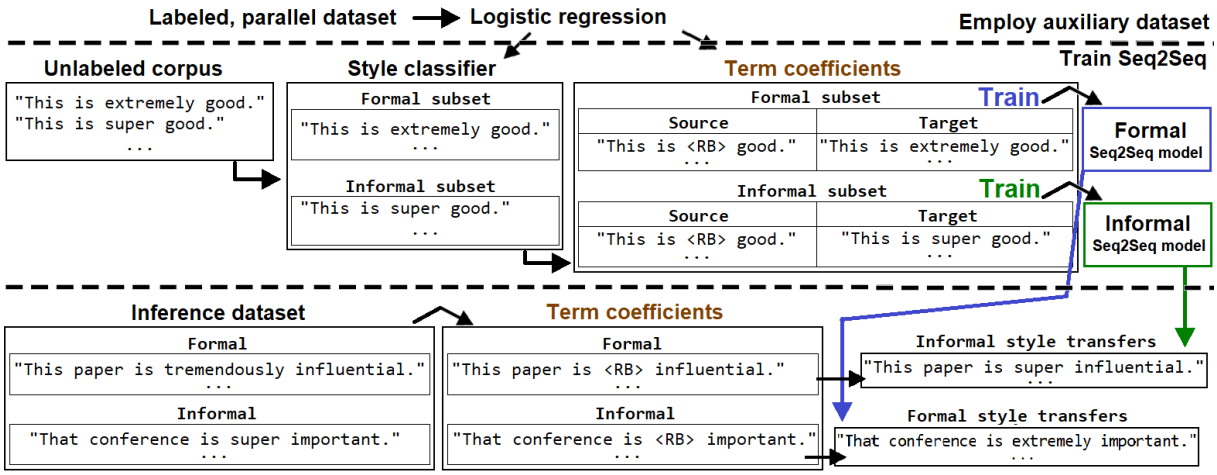

Figure 1: Pipeline for generating data, training Seq2Seq models, and applying style transfer.

feed this transformed y' to the model, and it predicts $\hat{y}$, which should be formal because the model learned to replace POS tags by words that are formal and are suited to the other words in the sentence. The full pipeline is shown in Figure 1.

5 Datasets

We used multiple datasets, existing and novel.

The **Yahoo Formality Dataset** was obtained from (Rao and Tetreault, 2018), and it contains 106k formal-informal pairs of sentences. Informal sentences were extracted from Yahoo Answers ("Entertainment & Music" and "Family & Relationships" categories). Formal (parallel) sentences were produced with mechanical turks.

The **TV-Shows Datasets** are the scripts of 5 popular TV-shows from the 1990's and 2000's (*Friends*, *Futurama*, *Seinfeld*, *Southpark*, *Stargate SG-1*), with 420k sentences in total. The datasets are novel: we produced them by crawling a website that contains scripts of TV-shows and movies (IMS); except for *Friends*, obtained from (Fri).

The **Slate Star Codex** is a novel dataset we produced in this work. It is comprised of 3.2 million sentences from comments in the online forum *Slate Star Codex*(SSC), which contains very formal language in the areas of science and philosophy. It was obtained by crawling the website, and contains posts from 2013 to 2019.

6 Experimental Setup

We applied the techniques explained in Sections 3 and 4. We used the Yahoo Formality Dataset as labeled dataset B and either a TV-show dataset, all TV-shows together, or the Slate Star Codex dataset as unlabeled corpus A. A Logistic Regression model was employed as the classifier [2] , and OpenNMT as the Seq2Seq models [3].

The hyperparameters of the Seq2Seq models are shown in Table 1.

Hyper-parameter	Value
Encoder	
type	LSTM
rnn hidden size	100
layers	1
Decoder	
type	LSTM
rnn hidden size	100
layers	1
General	
word vec size	200
optimizer	Adam
learning rate	$1e^{-3}$
train/validation split	90/10
vocabulary size	30k for SSC, TV merged 10k for single TV-shows

Table 1: Hyperparameters.

[2] Scikit-learn's model was used. Terms were stemmed with Porter Stemming before being fed to the model, and only terms with frequency $\geq$2 in the dataset were fed.

[3] To derive formal and informal datasets from each of our original unlabeled corpora, we applied our logistic regression model on each sentence in each corpus. Sentences with informality scores ≤ 0.6 were considered formal, scores ≥ 0.65 were considered informal, and others were ignored for being neutral. Terms were replaced by POS tags in the following manner: the N terms in each sentence with the highest absolute weight (from the Log-Reg model) are replaced by POS tags, provided they pass a certain threshold (-0.001 for formal terms, and 0.2 for informal terms). N is the floor of the number of terms in the sentence divided by 5.

Dataset	Target formality	Avg. formality score (1–5)	Avg. suitability score (1–5)	Total # of sentences
Friends	formal	$2.25 \rightarrow 3.43$ $(+1.2)$	$3.44 \rightarrow 3.43$ (-0.0)	105k
	informal	$3.91 \rightarrow 1.63$ (-2.3)	$3.18 \rightarrow 3.53$ $(+0.4)$	
Futurama	formal	$2.04 \rightarrow 3.43$ $(+1.4)$	$3.39 \rightarrow 2.29$ (-1.1)	27k
	informal	$4.41 \rightarrow 1.85$ (-2.6)	$3.71 \rightarrow 3.00$ (-0.7)	
Seinfeld	formal	$1.84 \rightarrow 3.18$ $(+1.3)$	$3.58 \rightarrow 2.82$ (-0.8)	94k
	informal	$3.62 \rightarrow 1.71$ (-1.9)	$4.00 \rightarrow 3.41$ (-0.6)	
Southpark	formal	$1.92 \rightarrow 3.47$ $(+1.6)$	$3.00 \rightarrow 3.18$ $(+0.2)$	77k
	informal	$3.92 \rightarrow 1.59$ (-2.3)	$3.69 \rightarrow 3.06$ (-0.6)	
Stargate-SG1	formal	$2.17 \rightarrow 4.06$ $(+1.9)$	$3.50 \rightarrow 3.17$ (-0.3)	117k
	informal	$4.59 \rightarrow 1.77$ (-2.8)	$3.41 \rightarrow 3.30$ (-0.1)	
All TV-Shows	formal	$2.38 \rightarrow 4.18$ $(+1.8)$	$3.77 \rightarrow 3.76$ (-0.0)	420k
	informal	$3.94 \rightarrow 1.92$ (-2.0)	$4.24 \rightarrow 3.92$ (-0.3)	
Slate Star Codex	formal	$3.53 \rightarrow 4.40$ $(+0.9)$	$3.67 \rightarrow 3.93$ $(+0.3)$	3.2M
	informal	$4.75 \rightarrow 2.86$ (-1.9)	$4.19 \rightarrow 3.93$ (-0.3)	
Yahoo (baseline)	formal	$2.45 \rightarrow 2.80$ $(+0.4)$	$3.85 \rightarrow 3.05$ (-0.8)	218k
	informal	$3.89 \rightarrow 3.33$ (-0.6)	$4.33 \rightarrow 2.79$ (-1.5)	

Table 2: Results of experiments on formality and sentence suitability.

Numbers and proper names were replaced by symbols <NUMBER> and <NAME> respectively, in order to greatly reduce data sparcity.

After splitting each corpus in formal and informal sentences (according to our logistic regression model), we randomly selected 60 sentences from each corpus (30 formal and 30 informal) as held-out test sets, and transformed them to opposite styles. Sentences were assigned evenly split to 3 human evaluators. To avoid bias, each sentence was randomly shown either original or transformed with equal probabilities (without evaluators' knowledge). Each sentence was shown accompanied with a *context*: preceding sentence in the TV-show (or SSC post), character speaking and TV-show name. Evaluators rated each sentence formality and *suitability* (how fluent and appropriate it is for the context) in a 1–5 scale[4].

Additionally, to serve as baseline, we trained two Seq2Seq models (formal-to-informal and informal-to-formal) on OpenNMT directly on the pairs of parallel sentences of the Yahoo Formality Dataset. We used the same hyper-parameters as the other experiments. Then we applied the model on the All TV-Shows corpus and performed the same human evaluation as described above, but we doubled the number of sentences analyzed to 120.

7 Results

Results are presented in Table 2. The average scores show the differences between the scores of the original and transformed sentences.

The technique produced a sizable change in formality while maintaining fluency and context. When transforming informal sentences to formal, the average formality score increased by $\sim$1.5 points (in a 5-point scale) for TV shows, and 0.9 point for SSC. In the formal-to-informal transformation, the formality score decreased by $\sim$2.2. The absolute changes in formality seem to correlate with the formality scores of the original sentences. They do not seem to correlate with the total number of sentences in each dataset.

Average suitability scores suffered a small decrease for corpora with a low number of sentences. The biggest decrease was for Futurama, whose training datasets contained only $\sim$10k sentences (after splitting the 27k total in the corpus). Other datasets contained smaller decreases in suitability, or even small improvements over the original sen-

[4] **1:** The sentence does not form any grammatical structure, or the evaluator cannot understand its meaning. **2:** The sentence forms segments of grammatical structures, and the evaluator can barely understand the intended meaning. **3:** The sentence is a few words away from perfect English, and the evaluator probably understands its meaning; or meaning is clear, but not appropriate for the context. **4:** The sentence is in almost perfect English (usually only missing a word or a comma, which is common in informal oral speech) and the meaning is clear; or the English is perfect but the meaning or words used are not perfectly appropriate for the context. **5:** The sentence is in perfect English and perfectly appropriate for the context.

tences. The largest corpora (All TV-Shows and SSC) maintained suitability scores approximately unchanged ($\in [-0.3, +0.3]$).

In general, all datasets showed sizable differences of formality when the formal or informal transformation was applied, and showed small decreases in suitability for small datasets (e.g. 10k training sentences for Futurama) and approximately no changes in suitability for larger datasets. Note that the suitability scores for the original sentences were not 5, because many sentences in the conversations employed in the datasets are in oral ("wrong") English, had small typos, or do not seem appropriate for the context.

The baseline (directly training the OpenNMT model with the Yahoo Formality Dataset) only showed small absolute changes in formality ($\sim$0.5) and lost a sizable amount of average suitability score (-0.8 or -1.5). We suspect the main reason for the loss of average suitability is the mismatch of the data used to train the model with the data on which the style transfer was applied, both in terms of vocabulary and in structure. The main reason for the smaller absolute change in formality scores, we suspect, is the model being conservative on making changes when it encountered sentences with many new terms. For many sentences generated by the model, the generated sentence was equal to the original sentence, which did not occurred as frequently in the other models (probably because of a greater match between training data and inference data).

On the All TV-Shows dataset, our method outperforms the baseline by 1.4 points in absolute formality change (both formal and informal transfers), and by 0.8 and 1.2 in average suitability.

8 Conclusion

In this work we presented a technique to derive a dataset of aligned pairs from an unlabeled corpus by using an auxiliary dataset. The technique is particularly important for noisy/user-generated text, which often lack datasets of matching vocabulary and structure. We tested the technique with the Yahoo Formality Dataset and 7 novel datasets we produced by web-crawling, which consists of scripts from 5 TV-shows, all TV-shows together, and the SSC online forum. We gathered 1080 human evaluations on the formality and suitability of sentences, and showed that our method produced a sizable change in formality while maintaining fluency and context; and that it considerably outperformed OpenNMT's Seq2Seq model trained directly on the Yahoo Formality Dataset.

A possible application of this technique in industry is to use large standard datasets as auxiliary to build style transformers based on specific corpora relevant to the industry. For example, a company wishing to change the formality of comments in its website could use the Yahoo Formality Dataset as the auxiliary dataset and use the logs of comments in its own website as the main corpus. This would enable them to create style transfers that are suited to the vocabulary and structures they use, improving style-transfer and fluency.

For future work, we plan to research different models for selecting the words most characteristic of formality instead of the logistic regression model used, such as neural models.

We make available the full pipeline code (ready-to-run) and our novel datasets: https:/github.com/ICEtinger/StyleTransfer

References

Friends dataset. https://github.com/npow/friends-chatbot/tree/master/data.

IMSDB. https://www.imsdb.com/.

Slate Star Codex online forum. https://slatestarcodex.com/.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2017. Zero-shot style transfer in text using recurrent neural networks. *arXiv preprint arXiv:1711.04731*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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.

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.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *CoRR*, abs/1804.06437.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. *CoRR*, abs/1804.09000.

Sudha Rao and Joel R. Tetreault. 2018. Dear sir or madam, may I introduce the YAFC corpus: Corpus, benchmarks and metrics for formality style transfer. *CoRR*, abs/1803.06535.

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.

Alexey Tikhonov and Ivan P Yamshchikov. 2018. What is wrong with style transfer for texts? *arXiv preprint arXiv:1808.04365*.

Wei Xu. 2017. From shakespeare to twitter: What are language styles all about? In *Proceedings of the Workshop on Stylistic Variation*, pages 1–9.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. *Proceedings of COLING 2012*, pages 2899–2914.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, pages 7287–7298.

Ye Zhang, Nan Ding, and Radu Soricut. 2018. Shaped: Shared-private encoder-decoder for text style adaptation. *arXiv preprint arXiv:1804.04093*.