A Dataset for Low-Resource Stylized Sequence-to-Sequence Generation

Yu Wu,[1] **Yunli Wang,**[2] **Shujie Liu**[1]
[1]Microsoft Research, Beijing, China
[2]State Key Lab of Software Development Environment, Beihang University, Beijing, China
{yuwu1, shujliu}@microsoft.com {wangyunli}@buaa.edu.cn

Abstract

Low-resource stylized sequence-to-sequence (S2S) generation is in high demand. However, its development is hindered by the datasets which have limitations on scale and automatic evaluation methods. We construct two large-scale, multiple-reference datasets for low-resource stylized S2S, the Machine Translation Formality Corpus (MTFC) that is easy to evaluate and the Twitter Conversation Formality Corpus (TCFC) that tackles an important problem in chatbots. These datasets contain context to source style parallel data, source style to target parallel data, and non-parallel sentences in the target style to enable the semi-supervised learning. We provide three baselines, the pivot-based method, the teacher-student method, and the back-translation method. We find that the pivot-based method is the worst, and the other two methods achieve the best score on different metrics.

Introduction

The S2S framework (Sutskever, Vinyals, and Le 2014) has achieved great success in recent years. However, a surge of tasks require S2S models to generate texts in a specific style without abundant parallel data, such as formal response generation in chatbots, which is in high demand but does not perform very well (Shum, He, and Li 2018). Table 1 shows that replying formally is important for chatbots, especially in the domains of customer service.

We investigate the low-resource stylized sequence-to-sequence generation problem. Usually, context to target-style sentence pairs are unavailable, but adequate context to source-style sentence pairs are easy to collect. For instance, informal conversational data can be acquired easily on Twitter, but it is hard to find informal message and formal response text pairs (Li et al. 2016b). With the context to source-style sentence pairs, crowd-sourcing efforts can be made to construct source-style to target-style sentence pairs. In such a way, the context and target-style sentence is connected with the help of source style sentence, which is the main difference with the non-parallel style transfer task (Shen et al. 2017).

With such assumptions, we introduce two benchmark datasets, the Twitter Conversation Formality Corpus (TCFC) and the Machine Translation Formality Corpus (MTFC) by extending the Grammarly's Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault 2018). Both datasets focus on a specific style, formality, and consist of a large amount of training data, as well as human-annotated, multiple-reference test data. Specifically, the Twitter Conversation Formality Corpus aims to teach an agent to respond humans in a formal way. We prepare 1.7 million informal message-response pairs from Twitter, and 52,595 informal to formal English text pairs borrowed from the GYAFC for training. Regarding to model validation, we ask native speakers to rewrite 2000 informal responses into formal style.

Informal Input	My laptop appears a windows 10 issue
Informal Output	Hi Bro! receiving any error codes when trying to sync? any details with help us assist u.
Formal Output	Are you receiving any error codes when you try to sync? Any details would help us assist you better.

Table 1: An example of stylized chatbots. For ease description, we call the informal input as context.

Although the stylized conversation has many potential applications in the real world, it is hard to evaluate (Liu et al. 2016). Motivated by this, we further construct an easier evaluate task, the Machine Translation Formality Corpus. The MTFC consists of 15 million informal Chinese to informal English text pairs which are carefully filtered from the OpenSubtitle Dataset (Lison and Tiedemann 2016). The informal to formal English text pairs are borrowed from the GYAFC as well. For tuning and testing, we ask human annotators to create over 3,000 human-annotated informal Chinese to formal English pairs. For both datasets, we further prepare large-scaled non-parallel formal sentences to enable the training of semi-supervised methods (Sennrich, Haddow, and Birch 2016a).

Since this task can be treated as a specific multilingual machine translation problem, we employ three approaches from low-resource machine translation as baselines: 1) the pivot-based method (Cohn and Lapata 2007) that conducts stylized S2S generation in a pipeline manner; 2) the teacher-student model (Chen et al. 2017) that tackles the error propagation by knowledge distillation; and 3) the back-translation method (Sennrich, Haddow, and Birch 2016a)

	Source style and target style	# Avg. sentence per style	# Number of styles	Pivot Resource	#Ref Number
Twitter Persona (Li et al. 2016b)	Twitter Id ↔ Twitter Id	92.24	74,003	No	1
Kennedy's speech (Wang et al. 2017)	Twitter ↔ JFK speech	6474	2	No	1
Start Wars (Wang et al. 2017)	Twitter ↔ Start Wars movie script	495	2	No	1
PERSONA-CHAT (Zhang et al. 2018b)	Labeled persona ↔ Labeled persona	142	1155	No	1
TCFC (Ours)	Twitter ↔ Formal English	1,007,999	2	Yes	2
MTFC (Ours)	Informal Chinese ↔ Formal English	1,007,999	2	Yes	4

Table 2: Comparison with existing datasets. The third column is the averaged sentence number of each style (persona). The fourth column is whether the dataset provides parallel data between different styles.

that is able to leverage non-parallel data. Empirical results show that the pivot-based is the worst, indicating the problem cannot be handled perfectly by a simple combination of state-of-the-art sequence-to-sequence model and a style transfer model. The teacher-student method and the back-translation method obtain the top place on different metrics, showing that the knowledge distillation and data augmentation could mitigate some challenges of the task. Data and code are shared at `https://github.com/MarkWuNLP/Data4StylizedS2S`

Our contributions are summarized as follows: 1) a challenging dataset about conversation style transfer is created, which has many potential applications in the industry; 2) a machine translation formality corpus is introduced, which is easy to evaluate and consists of large parallel and non-parallel data; 3) typical methods borrowed from machine translation are tested on the datasets.

Related Work

Text Style Transfer: Text style transfer with parallel data has been studied by Xu et al., who transfer modern English into Shakespeare style with a phrased-based machine translation model (PBMT). S2S models have been applied on this task (Jhamtani et al. 2017) and outperform PBMT on the same dataset. Similar techniques have been done on formality style transfer proposed by Rao and Tetreault.

As parallel data is hard to obtain, researchers begin to study text style transfer in an unsupervised manner. A popular approach is to learn a sentence disentangling representation with adversarial learning (Goodfellow et al. 2014). Hu et al. use variational auto-encoders with a discriminator to guide the decoder to generate sentences with desired attributes. Shen et al. propose a model building on distributional cross-alignment for style transfer and content preservation. Additionally, several studies (Fu et al. 2018; Zhao et al. 2018; John et al. 2018; Logeswaran, Lee, and Bengio 2018) design their methods under the adversarial learning or reinforcement learning framework. Recently, several works (Zhang et al. 2018b; Lample et al. 2019) explore how to create pseudo-parallel data by leveraging unsupervised machine translation methods (Lample et al. 2018). Text editing technique has been applied on this task in (Li et al. 2018). Multi-task learning is applied to the task (Niu, Rao, and Carpuat 2018) as well. Both style transfer and stylized S2S aim to generate text in a desired style. The difference is that stylized S2S generates a new content based on the task demands, whereas the text style transfer aims to preserve original content of the input. The most relevant work is Niu et al. (2017), which proposes to rerank generated sentences to satisfy the formality demand.

Datasets: Existing style transfer datasets focus on a "paraphrase setting", which aims to change the source sentence style while keeps its content. Popular datasets concerning sentiment modification include Yelp and Amazon (He and McAuley 2016). Gigaword (Graff and Cieri 2003) enables news style transfer, and GYAFC (Rao and Tetreault 2018) focuses on formality transfer. In contrast, our datasets require generate a specific style sentence under a conditional context, so a desired model should generate correct content along with a specific style, which significantly increases the task difficulty.

Before us, there are several works create personalized dialog datasets (Li et al. 2016b; Wang et al. 2017; Niu and Bansal 2018; Zhang et al. 2018a). Datasets of personalized dialogue systems define styles by movie character scripts (Li et al. 2016b) or Twitter Ids (Wang et al. 2017), resulting in limited sentences for a specific style compared to our datasets. As shown in 2, we can collect millions of formal/informal sentences with a high-confidence classifier, but it is very hard to collect Kennedy's style data. Furthermore, our datasets provide pivot resources (paraphrase sentences in different styles), enabling the training of pivot method. To guarantee the quality of automatic evaluation, we provide more references for testing.

Dataset Creation Process

We create two datasets for stylized S2S generation, referred to as the Twitter Conversation Formality Corpus (TCFC) and the Machine Translation Formality Corpus (MTFC). In this section, we elaborate on how to construct a parallel corpus $\mathcal{D} = \{(\mathbf{x_i}, \mathbf{y}_{i,s})\}_{i=0}^{N}$ comprising of context to source-style sentence pairs, a parallel corpus $\mathcal{S} = \{(\mathbf{y}_{j,s}, \mathbf{y}_{j,t})\}_{j=0}^{M}$ comprising of source-style to target-style sentence pairs, and a non-parallel corpus $\mathcal{M}_t$ containing formal sentences. $\mathbf{x}$, $\mathbf{y_s}$ and $\mathbf{y_t}$ refer to a context, a source-style sentence and a target-style sentence respectively.

Background: GYAFC Dataset

Since the construction of $\mathcal{S}$ and $\mathcal{M}_t$ is based on the GYAFC, we first give a brief introduction of the dataset. The GYAFC is the largest human labeled informal ↔ formal dataset. Firstly, the authors extract informal sentences from Entertainment&Music (E&M) and Family&Relationship (F&R) domains of the Yahoo Answers L6 corpus[1] with an in-house classifier. Sentences that are questions, contain URLs and are shorter than 5 words or longer than 25 are removed. Crowd-sourcing efforts are made to construct training, validation, and test sets, in which a worker is asked to rewrite the informal sentences to formal sentences with detailed instructions. Finally, there are around 50k text pairs for training, 3k for validation and 1.5k for testing for each domain.

In this paper, we utilize the dataset of the E&M domain as $\mathcal{S}$. As the in-house formality classifier in Rao and Tetreault is not released, we train a formality classifier on the human labeled 50k text pairs by regarding formal sentences as positive instances and informal ones as negative. The classifier achieves 92% accuracy on GYAFC data. We also test its performance on out-domain data (Tweets and subtitles). The accuracy on Tweets and subtitles are 83% and 78% respectively. Then we apply this classifier on sentences in E&M domain of Yahoo Answer L6 corpus, and select 1,007,999 sentences with high confidence scores as formal sentences to construct $\mathcal{M}_t$.

Twitter Conversation Formality Corpus

For the TCFC, we construct dataset $\mathcal{D}$ by crawling message-response pairs from Twitter. To minimize noise, we remove messages or responses that are shorter than 5 words or longer than 25 words. In the pre-processing, we remove hashtags, emoticons, and @mentions. Finally, we got 1,727,251 message-response pairs. The message-response pairs, parallel data borrowed from the GYAFC, and the non-parallel corpus mined from Yahoo Answers are all training data of the task, the statistic of which is shown in Table 3.

We ask two native speakers [2] to transfer 2000 responses [3] to formal responses for testing (1000 for tuning and 1000 for testing), where messages are visible as well. We teach annotators with detailed instructions and examples sampled from the GYAFC dataset to ensure the rewriting quality. The annotators are permitted to abandon samples, if she cannot understand the conversation clearly. Finally, we obtain 980 and 978 messages for tuning and testing.

The average char-level edit distance between the original informal responses and the formal rewrite responses is 27.33 and the distribution of the edit distance is plotted in Figure 1, indicating that the formality transfer cannot be completed if we only conduct a minor change. According to our statistics of 100 sampled pairs, we see specific percentages of sentence-level paraphrases (33%), phrasal paraphrases (42%), edits to punctuations (50%), expansion of contractions (22%), capitalization (53%), and normalization (9%). Definitions and examples are illusrated in Table 4. These numbers demonstrate that the task cannot be solved by generating informal responses at first and then rewrite with rules. Sentence structures are different across different styles. We discuss the performance between pivot-based method and end-to-end method in the experiment.

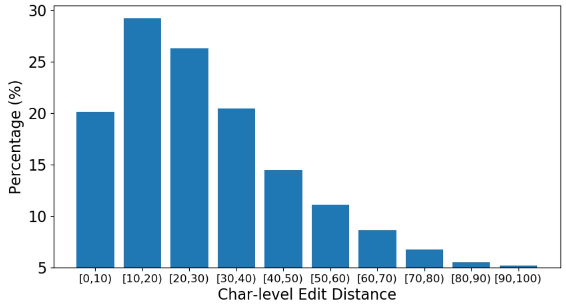

Figure 1: Distribution of character-level edit distance.

Machine Translation Formality Corpus

The goal of the MTFC is to translate an informal Chinese sentence into formal English, which is easy to evaluate and facilitates the development of spoken language translation. Ideally, $\mathcal{D}$ should be constructed by collecting human labeled Chinese ↔ English parallel data from Yahoo Answers. However, it is very exhausted to annotate millions of parallel data for training. We select bilingual subtitle parallel data to build the dataset $\mathcal{D}$. We collect a huge amount of Chinese ↔ English pairs by mining dual subtitles from the OpenSubtitle. To ensure data quality, we carefully detect and extract dual Chinese-English pairs following method in (Zhang, Ling, and Dyer 2014). Bad sentences are filtered out according to the alignment score obtained by the fast-align toolkit[4]. Furthermore, the formality classifier is employed to select the informal subtitles with a high confidence. All subtitles in $\mathcal{D}$ have over 70% probability predicted by the classifier to be an informal sentence. We remove subtitles that are shorter than 5 words or longer than 25 words to control the data length distribution. We end up with 14 million Chinese-English pairs.

We extend the GYAFC dataset to create a validation and a test set. The GYAFC provides 2877 and 1416 informal-formal English sentence pairs in the Entertainment & Music domain for tuning and test, in which each pair contains one informal sentence and four formal sentences as references. For each text pair, we asked a Chinese annotator to translate the informal English sentence into informal Chinese, since a Chinese person is capable of writing fluent sentences in

[1] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l

[2] The annotators are hired by a data labeling company. Their payment is 0.5$ per sentence.

[3] The averaged formality score of Tweet responses is 0.21

[4] https://github.com/clab/fast_align

	Training			Evaluation		
	Dataset $\mathcal{S}$	Dataset $\mathcal{D}$	Dataset $\mathcal{M}_t$	Validation	Test	Avg. Words
MTFC	52,595/12.61/12.68	14,280,494/8.72/10.39	1,007,999	2865	1412	12.71/12.65/12.74
TCFC	52,595/12.61/12.68	1,727,251/12.73/11.46	1,007,999	980	978	14.27/15.68/16.08

Table 3: Corpus statistics. In the column of dataset $\mathcal{D}$, three numbers are the number of sentence pairs, the average word count of $\mathbf{x}$, and the average word count of $\mathbf{y_s}$. Similarly, three numbers of dataset $\mathcal{S}$ are the number of the sentence pairs, the word count of a source-style sentence $\mathbf{y_s}$, and the average word count of a target-style sentence $\mathbf{y_t}$.

Editing type	Definition	Message	Informal Response	Formal Response
Sentence Paraphrase	The sentence structure is paraphrase	I've followed so many people recently	remember when you had your bio as ' i follow back all swifties ' or something like that ? ?	Do you recall the time when you have your profile set as I follow back all swifties or something as such?
Phrasal Paraphrase	Only a phrase is changed in rewriting	After 4 years I'll have 960K tweets	After next four months I will have **25k for sure** .	After the next four months I will **certainly have 25,000**.
Contractions	Contractions are expanded.	Phone is not working , reach me via Facebook or Twitter if you need me	I wondered why you **didn't** text me back .	I wondered why you **did not** text me back .
Punctuation	Rewrite fixes punctuation errors.	ITS BEEN ALMOST 5 YEARS . WHY ARE U STILL RELEVANT IN MY LIFE !	but we only met a few months ago **.........**	but we only met a few months ago**.**
Normalization	Informal expressions are normalized.	OMG THANK YOU I LOVE U SM ALEX IM SO HAPPY WITH HOW IT TURNED OUT	I love **u** more. And **u** should be happy	I love **you** more. And **you** should be happy.
Capitalization	Correct word capitalization form.	cheer up ! You'll find a way to get the grades you always do !	**thank** you Jess ! Hope you're doing ok in college.	**Thank** you JESS ! I hope that you are doing well in your college .

Table 4: Examples in the TCFC dataset. Our dataset provides both informal and formal responses. We analyzes editing types between informal and formal responses.

Chinese. The annotator is permitted to discard the instance he does not understand clearly. By this means, we can get 2865 and 1412 <informal Chinese, informal English, formal English> sentence triples for tuning and test. In evaluation, we use <informal Chinese, formal English> text pairs to test performance.

Approaches

Pivot-Based Method

The most straight-forward method to tackle this problem is the pipeline-based method, also referred to as the pivot-based method (Cohn and Lapata 2007), where $\mathbf{y_s}$ is utilized as a pivot language to "bridge" $\mathbf{x}$ and $\mathbf{y_t}$. Formally, the generative model ($\mathbf{x} \rightarrow \mathbf{y_t}$) can be decomposed into two sub-models, where $\hat{\mathbf{y}}_t$ is computed by

$$\underset{\mathbf{y_t}}{\text{argmax}}\left(\sum_{\mathbf{y_s}}\Big(P(\mathbf{y_t}|\mathbf{y_s};\theta_{y_s\to y_t})P(\mathbf{y_s}|\mathbf{x};\theta_{x\to\mathbf{y_s}})\Big)\right) \tag{1}$$

where $\theta_{y_s\to y_t}$ and $\theta_{x\to y_s}$ are two parameters learned by maximum likelihood estimation on $\mathcal{D}$ and $\mathcal{S}$. Due to the exponential search space, the decoding process is usually approximated with two steps. The first step aims to generate y_s conditioned on context $\mathbf{x}$, formulated as

$$\hat{\mathbf{y_s}} = \underset{\mathbf{y_s}}{\text{argmax}} P(\mathbf{y_s}|\mathbf{x};\theta_{x\to y_s}). \tag{2}$$

After that, a target style sentence is obtained by

$$\hat{\mathbf{y_t}} = \underset{\mathbf{y_s}}{\text{argmax}} P(\mathbf{y_t}|\mathbf{y_s};\theta_{y_s\to y_t}). \tag{3}$$

Although the pivot-based method is a reasonable solution for this task, it suffers from two problems: error propagation and model discrepancy. In practice, we cannot obtain a perfect model to translate $\mathbf{x}$ into $\mathbf{y_s}$, therefore, errors made in the first step will propagate to the second step, which may hurt the quality of outputs. More seriously, the topics and vocabulary of $\mathcal{D}$ and $\mathcal{S}$ are loose-related, decreasing the performance of the method.

Teacher-Student Framework

To deal with the error propagation problem, the teacher-student framework first learns a teacher model $P(\mathbf{y_t}|\mathbf{y_s}, \theta_{y_s\to y_t})$ with the use of $\mathcal{S}$. Then, the teacher model teaches the student model $P(\mathbf{y_t}|\mathbf{x};\theta_{x\to y_t})$ by minimizing the KL divergence

$$\text{KL}\left(P(\mathbf{y_t}|\mathbf{y_s};\theta_{y_s\to y_t})||P(\mathbf{y_t}|\mathbf{x};\theta_{x\to y_t})\right). \tag{4}$$

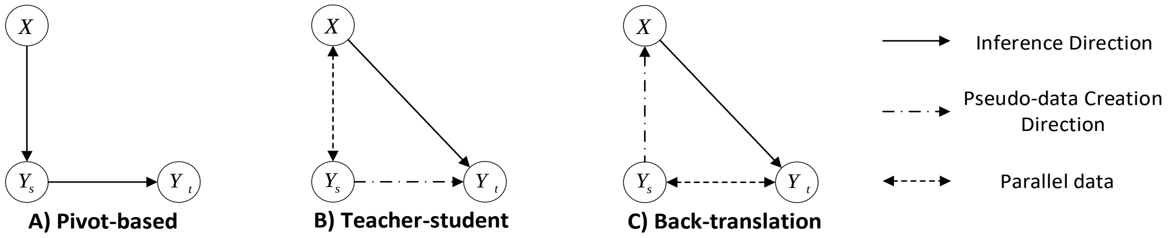

Figure 2: The picture shows how to create pseudo-parallel data and how to generate style-specific sentences.

Because $\theta_{y_s \to y_t}$ is fixed in the teaching process, Equation 4 is rewritten as

$$\mathcal{J} = -\sum_{(\mathbf{x},\mathbf{y_s})\in\mathcal{D}} \sum_{\mathbf{y'_t}} q(\mathbf{y'_t}|\mathbf{y_s}) \log P(\mathbf{y'_t}|\mathbf{x};\theta_{x\to y_t}), \qquad (5)$$

where $q(\mathbf{y'_t}|\mathbf{y_s})$ represents the teacher's sequence distribution over the sample space of all possible sequences. Due to the exponential search space, we consider an approximation of the objective by replacing the teacher distribution q with

$$q(\mathbf{y_t}|\mathbf{y_s}) = \mathbb{1}\{\hat{\mathbf{y_t}} = \underset{\mathbf{y'_t}}{\text{argmax}} q(\mathbf{y'_t}|\mathbf{y_s})\} \qquad (6)$$

where $\mathbb{1}(\cdot)$ is an indicator function, and $\hat{\mathbf{y_t}}$ is obtained by beam search. Finally, the objective function is formulated as

$$\mathcal{J} = -\sum_{(\mathbf{x},\mathbf{y_s})\in\mathcal{D}} \mathbb{1}\{\mathbf{y_t} = \hat{\mathbf{y_t}}\} \log P(\mathbf{y_t}|\mathbf{x};\theta_{x\to y_t}). \qquad (7)$$

Equation 7 gives to a simple training procedure, in which the student network is trained on the data generated by a teacher network. The procedure allows the parameter estimation in one model, avoiding error propagation problem.

Back-Translation

Back-translation (Sennrich, Haddow, and Birch 2016a) has proven effective on data augmentation. It has been widely applied to various tasks, such as unsupervised machine translation (Lample et al. 2018) and text style transfer (Rao and Tetreault 2018). We also test the performance of back-translation on stylized S2S generation.

Specifically, we first train two back-directional models, including a target-style to source-style model parameterized by $P(\mathbf{y_s}|\mathbf{y_t}, \theta_{y_t \to y_s})$, and a source-style to context model parameterized by $P(\mathbf{x}|\mathbf{y_s}, \theta_{y_s \to x})$. The pseudo-parallel data are created by two ways, using a limited parallel corpus $\mathcal{S}$ and a large scale non-parallel corpus $\mathcal{M}_t$ respectively. $\forall(\mathbf{y_s}, \mathbf{y_t}) \in \mathcal{S}$, we translate $\mathbf{y_s}$ to $\hat{\mathbf{x}}$ with

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\text{argmax}} P(\mathbf{x}|\mathbf{y_s}, \theta_{y_s \to x}). \qquad (8)$$

forming a pseudo text pair $(\hat{\mathbf{x}}, \mathbf{y_t})$. Similarly, $\forall \bar{\mathbf{y}} \in \mathcal{M}_t$, we translate $\bar{\mathbf{y}}$ to $\hat{\mathbf{x}}$ with

$$\underset{x}{\text{argmax}} \left(\sum_{\mathbf{y_s}} \big(P(\mathbf{x}|\mathbf{y_s}, \theta_{y_s \to x}) P(\mathbf{y_s}|\bar{\mathbf{y}}_\mathbf{t}, \theta_{y_t \to y_s}) \big) \right). \qquad (9)$$

where the decoding process is also decomposed to two discrete steps as described in Equation 2 and 3, forming a pseudo-parallel data $(\hat{\mathbf{x}}, \bar{\mathbf{y}}_\mathbf{t})$. By merging the data generated by Equation 8 and 9, a large pseudo-parallel dataset $\mathcal{P} = \{(\hat{\mathbf{x}}_\mathbf{l}, \mathbf{y'}_{\mathbf{t},\mathbf{l}})\}$ is obtained. Finally, we use $\mathcal{P}$ to train a generative model by maximizing the log-likelihood of

$$\mathcal{J} = \sum_{l} \log P(\mathbf{y'}_{t,l}|\hat{\mathbf{x}}_l, \theta_{x \to \mathbf{y_t}}). \qquad (10)$$

Both teacher-student and back-translation methods create pseudo-parallel data for model training, with the difference that the target side of back-translation generated data is human written, while it is model generated for teacher-student methods, whose source side is human written.

Data Augmentation

As 50k text pairs are not large enough for a NMT model, we employ a data augmentation technique for above three three methods, which will increase the accuracy of the estimation of $P(\mathbf{y_t}|\mathbf{y_s}; \theta_{y_s \to y_t})$ and $P(\mathbf{y_s}|\bar{\mathbf{y}}_\mathbf{t}, \theta_{y_t \to y_s})$. Inspired by (Lample et al. 2018), we train a formal → informal model by employing the PBMT model, where the language model of PBMT is trained on the E&M and F&R domains of Yahoo Answers L6. Then we utilize the PBMT to translate sentences in $\mathcal{M}_t$ to informal style. After removing poor quality back-translation results (word repetition or too long), we merge the back-translation results with original text pairs in $\mathcal{S}$. It should be noted that text pairs in $\mathcal{S}$ are duplicated 10 times to ensure the quality of the final pseudo-parallel data.

Experiments

Implementation Details

We describe implementation details on the MTFC, and the situation on the TCFC is similar. In the pivot-based model, the Transformer model (Vaswani et al. 2017) is adopted to approximate the conditional sequence generation probability $P(\mathbf{y_s}|\mathbf{x}, \theta_{x \to y_s})$. The transformer model consists of a 6-layer encoder and decoder, whose model size is 512. The multi-head attention quantity is 8. All models are trained on 4 Tesla Titan X GPUs for a total of 200K steps using the Adam algorithm (Kingma and Ba 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We employ the byte-pair encoding (BPE) approach (Sennrich, Haddow, and Birch 2016b) to handle the open vocabulary problem, whose size is 25,000. The initial learning rate is set to 0.2 and decayed according to the schedule in (Vaswani et al. 2017). During training, the batch size is 4096 words and checkpoints are created every 5000

	Formality		Fluency		Overall	
	Auto	Human	Auto	Human	BLEU	Human
Base Model	0.555	-0.31	3.61	3.65	33.47	2.65
Pivot$_{rule}$	0.679	0.14	3.58	3.68	37.83	3.14
Pivot$_{model}$	0.757	0.25	3.64	3.57	38.75	2.76
Teacher-student	**0.768**	**0.57**	3.60	3.78	40.07	3.22
Back-translation	0.707	0.46	3.59	3.75	**40.68**	**3.28**
Correlation	0.396		0.211		0.435	

Table 5: Results on the MTFC. The last row shows the Spearman rank correlation (Spearman 1987) between an automatic metric and the human annotation. Content preservation, a typical metric of style transfer, is not employed in this paper, because this is a language translation task.

	Formality		Fluency		Overall					Diversity	
	Auto	Human	Auto	Human	BLEU-2	Avg	Ext	Grd	human	Dis-1	Dis-2
Base Model	0.506	0.95	3.25	4.63	1.93	0.641	0.414	0.484	2.21	0.086	0.247
Pivot$_{rule}$	0.583	1.13	3.69	4.71	2.86	0.662	0.425	0.500	2.41	0.071	0.216
Pivot$_{model}$	0.807	1.42	3.72	**4.75**	3.47	0.681	0.427	0.527	2.51	0.065	0.205
Teacher-student	**0.862**	1.77	**3.73**	4.68	**4.10**	0.684	0.425	0.538	2.87	0.052	0.155
Back-translation	0.844	**2.02**	3.72	4.62	3.97	0.684	0.423	0.530	**3.16**	**0.113**	**0.320**
Correlation	0.355		0.189		0.08	0.177	0.169	0.172	-	-	

Table 6: Results on the TCFC.

batches. The generative probability of $P(\mathbf{y_t}|\mathbf{y_s}, \theta_{y_s \rightarrow y_t})$ is estimated by a sequence to sequence (S2S) model, in which the encoder and decoder are 1-layer GRU with 512 units.

Regarding the teacher-student model, we employ the GRU in the pivot-based model as the teacher model, which translates informal English sentences to formal English. A Transformer is used as the student model, which is trained from scratch on these text pairs using Equation 7. In terms of back-translation, we fine-tune the Transformer used in the pivot-based model with pseudo-parallel data generated by the back translation. For all models, the beam size is 4 and the length penalty is 1.2. We further report results of the base model and the Pivot$_{rule}$. The base model means that we directly evaluate the generated $\mathbf{y_s}$ of the Pivot$_{model}$. Pivot$_{rule}$ represents that we rewrite the generation result $\mathbf{y_s}$ with several effective rules. [5]

Evaluation Metrics

Formality: We train a GRU based classifier using the training data of the GYAFC by regarding formal/informal sentences as positive/negative instances respectively. The formality classifier achieves 92% accuracy on the validation and test data of the GYAFC. We utilize the classifier to assign a formality score for a generated sentence as an automatic evaluation metric.

Fluency: We employ the method proposed in (Heilman et al. 2014) to evaluate the output fluency. It is a statistical model, which is able to assign a score from 0 to 4 for the grammar correctness of a sentence.

Overall Evaluation: We evaluate machine translation results with case-sensitive BLEU (Papineni et al. 2002). In terms of conversation, following suggestions in (Liu et al. 2016), we evaluate results with Embedding Average (Avg), Embedding Extrema (Ext), Embedding Greedy (Grd) and BLEU-2. Following Li et al., we evaluate the response diversity based on the ratios of distinct unigrams and bigrams, denoted as Distinct-1 and Distinct-2.

Human Evaluation: The outputs of 300 randomly sampled contexts are chosen for human evaluation. Three human annotators are required to label the formality score of all model outputs from -3 to 3, denoted as: -3: Very Informal, -2: Informal, -1: Somewhat Informal, 0: Neutral, 1: Somewhat Formal, 2: Formal and 3: Very Formal. We also ask humans to evaluate output fluency on a scale of 5: 5: Perfect, 4: Comprehensible, 3: Somewhat Comprehensible, 2: Incomprehensible 1 Other. We further assign a relative rank for each model output by considering its overall quality. Specifically, given a context sentences, we collect and shuffle outputs of different models, and then ask humans to rank them in a descending order, where the top one gets a score of 4, the second one gets a score of 3, etc[6]. For each output, its appropriateness to the context has a higher priority than its formality in labeling. The overall score is

$$overall(m_i) = \frac{1}{|C|} \sum_{c \in C} rank(y_{c,m_i}) \qquad (11)$$

where m_i denotes the i-th model, C is a set of contexts, and y_{c,m_i} is the output of m_i for c.

Evaluation Results

Tables 5 and 6 show the evaluation results.

Human Assessment: Regarding the MTFC, the back-translation and teacher-student method are top-2 in terms of overall quality, since both of them avoid error propagation

[5] Rules include capitalization, lowercase words with all upper characters, remove repetitive words, expand contractions, and remove slang and swear words.

[6] If a relative rank is $y_{c,m_1} > y_{c,m_2} = y_{c,m_3} > y_{c,m_4}$, they will receive 4,3,3,1 respectively.

Editing type	% in dataset and output	Direct Translation (Base model)	Teacher-student (TS)	Ground Truth
Big Paraphrase	42%/19%	For a guy who doesn't like romantic movies, this is really nice.	This is really good for a man who does not like romantic movies.	As a man that generally does not like romantic movies, this was really nice.
Small Paraphrase	26%/29%	This is the stupidest thing I seen in years.	That is the most stupid thing I have seen in years.	It is the biggest load of foolishness that I have seen for ages.
Rule	32%/52%	**It's** time to move on.	It is time to move on .	It is time to move on .

Table 7: Editing Analysis on the MTFC. We distinguish big/small paraphrase by whether the sentence structure is changed. The first number in the second column represents the ratio in the dataset.

by completing this task with an end-to-end model. Teacher-student method achieves the best score on formality, because 1) the formality patterns given by the teacher model are easier to learn for a neural model, and 2) pseudo-data of back-translation may contain noise. As expected, the pivot$_{model}$ cannot handle the task well, even worse than pivot$_{rule}$ on the overall quality. After observing the outputs, pivot$_{model}$ sometimes misses important words to increase output formality in style transfer that drastically hurts the translation quality. Another possible reason for a bad BLEU score is that the topic of $\mathcal{D}$ (i.g. training data of the pivot$_{model}$) may slightly differ from the topic of dataset $\mathcal{S}$. All models show comparable results on fluency, because their decoders are all based on neural models which are able to generate plausible sentences. The trend on the TCFC is similar to the MTFC. Formality and fluency scores on the TCFC dataset are slightly higher than the MTFC. A possible explanation for this might be that output sentences in conversation are shorter and more generic, so they are easier to transfer.

Automatic Assessment: For the MTFC, automatic metric of formality and BLEU correlate with human moderately well, but fluency score correlates with human poorly. This is mainly because the statistical model is trained on essays which may slight differ from sentences of Yahoo Answers. On the TCFC, all overall metrics show weak correlation with human judgments, which is consistent with conclusion in (Liu et al. 2016). By comparison, MTFC is a better choice for automatic evaluation. Responses generated by the back-translation method are more diverse. It mainly because non-parallel sentences in $\mathcal{M}_t$, sampled from Yahoo Answers, are more diverse and informative than sentences in dialogue.

Discussion

Quality across sentence length: We break up the testset of the MTFC into buckets based on source sentence length (0-5 subword tokens, 5-10 subword tokens, etc.) and compute corpus-level BLEU scores for each. Figure 3 shows that the performance of all models decreases as the input sentence becomes longer. The back-translation and teacher-student model outperform the pivot-based model when a source sentence is long. This is mainly because error propagation becomes more serious as inputs become longer. The undesirable translation quality of context to source style may severely hurt the performance the subsequent style transfer model, but end-to-end models can alleviate this problem.

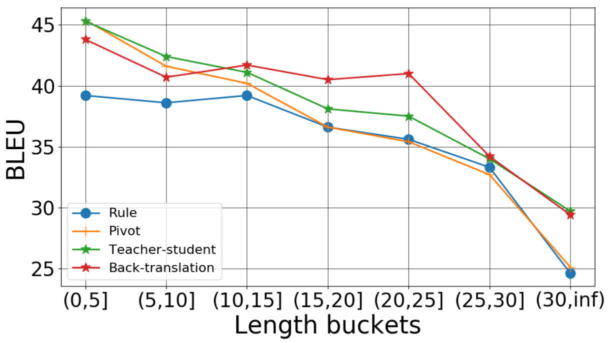

Figure 3: BLEU scores across sentence length.

Editing Analysis: We evaluate the editing types of the constructed datasets and test whether the baselines can accomplish these. We randomly select 100 informal-formal English text pairs from the MTFC, and find that the constructed datasets are super rich in abstractive transformations (42% big paraphrase and 26% small paraphrase), which demonstrates the difficulty of the datasets. However, when we compare the base model and the teacher-student model, we find that the ratio of big/small paraphrase drops to 19%/29%. It indicates that even the best performance model does not learn the paraphrase very well. Hence, there is a substantial room for future work to investigate the abstractive transformations.

Conclusion and Future Work

We focus on low-resource stylized sequence-to-sequence generation and construct two large-scale datasets. The MTFC is easy to evaluate, and the TCFC is beneficial to the dialogue system. We further test the performance of three methods, and find that current models cannot learn paraphrase well. In the future, we will investigate how to migrate this problem with limited parallel data.

References

Chen, Y.; Liu, Y.; Cheng, Y.; and Li, V. O. K. 2017. A teacher-student framework for zero-resource neural machine translation. In *ACL 2017*, 1925–1935.

Cohn, T., and Lapata, M. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *ACL 2007*.

Fu, Z.; Tan, X.; Peng, N.; Zhao, D.; and Yan, R. 2018. Style transfer in text: Exploration and evaluation. In *AAAI 2018*, 663–670.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS 2014*, 2672–2680.

Graff, D., and Cieri, C. 2003. English gigaword corpus. *Linguistic Data Consortium*.

He, R., and McAuley, J. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW 2016*, 507–517.

Heilman, M.; Cahill, A.; Madnani, N.; Lopez, M.; Mulholland, M.; and Tetreault, J. R. 2014. Predicting grammaticality on an ordinal scale. In *ACL 2014*, 174–180.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward controlled generation of text. In *ICML 2017*, 1587–1596.

Jhamtani, H.; Gangal, V.; Hovy, E.; and Nyberg, E. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

John, V.; Mou, L.; Bahuleyan, H.; and Vechtomova, O. 2018. Disentangled representation learning for text style transfer. *arXiv preprint arXiv:1808.04339*.

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

Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018. Phrase-based & neural unsupervised machine translation. In *EMNLP 2018*, 5039–5049.

Lample, G.; Sandeep, S.; Eric Michael, S.; Ludovic, D.; MarcAurelio, R.; and Y-Lan, B. 2019. Mulitiple-attribute text rewriting. In *ICLR*.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL 2016*, 110–119.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G. P.; Gao, J.; and Dolan, W. B. 2016b. A persona-based neural conversation model. In *ACL 2016*.

Li, J.; Jia, R.; He, H.; and Liang, P. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *NAACL 2018*, 1865–1874.

Lison, P., and Tiedemann, J. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Liu, C.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP 2016*, 2122–2132.

Logeswaran, L.; Lee, H.; and Bengio, S. 2018. Content preserving text generation with attribute controls. In *NeurIPS 2018*, 5108–5118.

Niu, T., and Bansal, M. 2018. Polite dialogue generation without parallel data. *TACL* 6:373–389.

Niu, X.; Martindale, M.; and Carpuat, M. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *EMNLP*.

Niu, X.; Rao, S.; and Carpuat, M. 2018. Multi-task neural models for translating between styles within and across languages. *arXiv preprint arXiv:1806.04357*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

Rao, S., and Tetreault, J. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *NAACL 2018*, 129–140.

Sennrich, R.; Haddow, B.; and Birch, A. 2016a. Edinburgh neural machine translation systems for WMT 16. In *WMT 2016*, 371–376.

Sennrich, R.; Haddow, B.; and Birch, A. 2016b. Neural machine translation of rare words with subword units. In *ACL 2016*.

Shen, T.; Lei, T.; Barzilay, R.; and Jaakkola, T. S. 2017. Style transfer from non-parallel text by cross-alignment. In *NIPS 2017*, 6833–6844.

Shum, H.; He, X.; and Li, D. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of IT & EE* 19(1):10–26.

Spearman, C. 1987. The proof and measurement of association between two things. *The American journal of psychology* 100(3/4):441–471.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS 2014*, 3104–3112.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*, 5998–6008.

Wang, D.; Jojic, N.; Brockett, C.; and Nyberg, E. 2017. Steering output style and topic in neural response generation. In *EMNLP*.

Xu, W.; Ritter, A.; Dolan, B.; Grishman, R.; and Cherry, C. 2012. Paraphrasing for style. In *COLING 2012*, 2899–2914.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Zhang, Z.; Ren, S.; Liu, S.; Wang, J.; Chen, P.; Li, M.; Zhou, M.; and Chen, E. 2018b. Style transfer as unsupervised machine translation. *arXiv preprint arXiv:1808.07894*.

Zhang, S.; Ling, W.; and Dyer, C. 2014. Dual subtitles as parallel corpora. In *LREC 2014.*, 1869–1874.

Zhao, Y.; Bi, W.; Cai, D.; Liu, X.; Tu, K.; and Shi, S. 2018. Language style transfer from sentences with arbitrary unknown styles. *arXiv preprint arXiv:1808.04071*.