Formality Style Transfer with Shared Latent Space

Yunli Wang[†], **Yu Wu**[◇], **Lili Mou**[‡], **Zhoujun Li**[†] , **Wenhan Chao**[†]
[†]State Key Lab of Software Development Environment, Beihang University, Beijing, China
[◇]Microsoft Research, Beijing, China
[‡]Dept. Computing Science, University of Alberta, Edmonton, Canada
Alberta Machine Intelligence Institute (Amii)
{wangyunli,lizj,chaowenhan}@buaa.edu.cn
Wu.Yu@microsoft.com doublepower.mou@gmail.com

Abstract

Conventional approaches for formality style transfer borrow models from neural machine translation, which typically requires massive parallel data for training. However, the dataset for formality style transfer is considerably smaller than translation corpora. Moreover, we observe that informal and formal sentences closely resemble each other, which is different from the translation task where two languages have different vocabularies and grammars. In this paper, we present a new approach, Sequence-to-Sequence with Shared Latent Space (S2S-SLS), for formality style transfer, where we propose two auxiliary losses and adopt joint training of bi-directional transfer and auto-encoding. Experimental results show that S2S-SLS (with either RNN or Transformer architectures) consistently outperforms baselines in various settings, especially when we have limited data.[1]

1 Introduction

The formality analysis of text plays an important role in natural language processing (NLP)-related applications, such as style analysis and human-computer interaction (Pavlick and Tetreault, 2016). With the advance of neural networks for sentence generation, it is possible to synthesize a new sentence, changing the formality style of an input sentence while retaining its meaning. This is referred to as *formality style transfer* in Rao and Tetreault (2018).

Rao and Tetreault (2018) in the meantime constructed the GYAFC[2] dataset, where they make crowdsourcing efforts to manually write 119K formal sentences from informal ones (examples shown in Table 1). It fertilizes the research of formality style transfer with parallel data, as most other style-transfer datasets (e.g., sentiment) are non-parallel, with only a style label available for each sentence.

In previous work, researchers have applied machine translation frameworks to style transfer when parallel data are available. For example, Xu et al. (2012) adopt phrase-based statistical machine translation, and Rao and Tetreault (2018) employ sequence-to-sequence (Seq2Seq) neural networks. Formality style transfer, however, is different from rich-resource machine translation mainly in two aspects.

First, machine translation systems usually require massive parallel data for training, whereas the parallel corpus for formality style transfer is much smaller. The WMT-17 English-Chinese corpus, for example, has 25M pairs (Bojar et al., 2017), but GYAFC has only $\sim$50K for each domain. Second, a formality-transferred sentence closely resembles the original one. As shown in Table 1, only a few words and punctuations are changed. It differs from machine translation, where the source and target sides are of completely different languages.

As a result, directly applying a sequence-to-sequence (Seq2Seq) model not only has the risk of overfitting to the training set, but also may fail to fully utilize the nature of informal and formal sentences.

To tackle these problems, we propose a novel model S2S-SLS, a *Sequence-to-Sequence model with a Shared Latent Space*, for formality style transfer. In S2S-SLS, we have a single encoder for both

[1] Our code and outputs are available at: https://github.com/jimth001/formality_style_transfer_with_shared_latent_space

[2] GYAFC is the abbreviation of *Grammarly's Yahoo Answers Formality Corpus*.

Proceedings of the 28th International Conference on Computational Linguistics, pages 2236–2249
Barcelona, Spain (Online), December 8-13, 2020

Informal sentence:	**Formal sentence:**
I do not know are u ready for one ?	I do not know. Are you ready for one?
Sounds like a rhetorical question :)	It sounds like a rhetorical question.
what r ya talking abt	What are you talking about ?

Table 1: Examples of formality style transfer in the GYAFC dataset.

informal and formal styles, but two decoders responsible for each style, respectively. We design two auxiliary losses to ensure that the shared latent space indeed captures the semantics of input, while eliminating its style. Our shared encoded latent space with separate decoders also enables the joint training of informal-to-formal and formal-to-informal transfer (although informal-to-formal is the main focus in applications). It also allows auto-encoding training for regularization, which aims to decode a sentence (either formal or informal) from itself.

Compared with the traditional Seq2Seq model, our method has the following advantages: (1) The shared encoder allows the model to learn a style-independent representation for both informal and formal sentences. It is easier to generate a stylized output from a style-independent representation than from multiple style-specific representations, because the style-specific space is sparser. (2) Our auxiliary matching losses force the encoder to learn better style-independent representations that capture the semantics of the input. (3) The auto-encoding training serves as a further regularization and preventing over-fitting in the data-limited scenario during training. And (4) the joint training of bi-directional transfer takes advantage of the similarity of two styles in the shared encoding phase, and enables the two directions to boost each other.

To verify the effectiveness and generalization of our method, we conduct experiments in three different settings: Data Limited, Data Augmentation, and Pre-training. In the data-limited scenario, experimental results show that our method is significantly better than previous work (Rao and Tetreault, 2018) by 4 and 7 BLEU scores on the two domains (namely, F&R and E&M) of the GYAFC dataset. When we use large-scale non-parallel data to enhance our method in the data augmentation and pre-training settings, our method still consistently outperforms the baselines by 1 BLEU score. The ablation test further studies the effectiveness of the joint training, the auto-encoding training, and the auxiliary losses in different scenarios, showing the robustness of our method.

2 Related Work

Style transfer has drawn considerable attention in the past few years. It can be generally categorized into three settings: (1) with fully unlabeled data, (2) with style-labeled data, and (3) with parallel data.

With fully unlabeled data. In this setting, the data are unlabeled, which could be raw text or images *per se*, and style transfer is accomplished in an unsupervised manner. For example, auxiliary losses of orthogonality and mutual information could help to learn independent features (Kumar et al., 2017; Chen et al., 2016); they have been shown to successfully disentangle features of color, rotation, etc. in image processing. In NLP applications, Xu et al. (2019a) capture the most salient feature by detecting the global variance, and perform unsupervised style transfer of sentiment. However, such approach hardly works for the intriguing formality style.

With non-parallel labeled data. By non-parallel labeled data, we mean that each data sample is annotated with its style label only. The content of different styles is generally the same in the corpus level, but for an individual data point, we do not have a style-transferred sentence with the same content. In sentiment-transfer sentence generation (Li et al., 2018), for example, a sentence is labeled with its sentiment tag (positive vs. negative). However, there is no alignment between a positive sentence and a negative one for the same content (subject).

With supervision signals of style labels, it is possible to learn style and/or content spaces. Adversarial training (Goodfellow et al., 2014) ensures that a certain latent space is indistinguishable in different styles, facilitating style-transfer generation (Hu et al., 2017). Fu et al. (2018) use real-valued embedding to represent styles explicitly. John et al. (2019) and Bao et al. (2019) extend such approach by encoding both style and content spaces in a disentangled way. Li et al. (2018) propose an editing-based approach

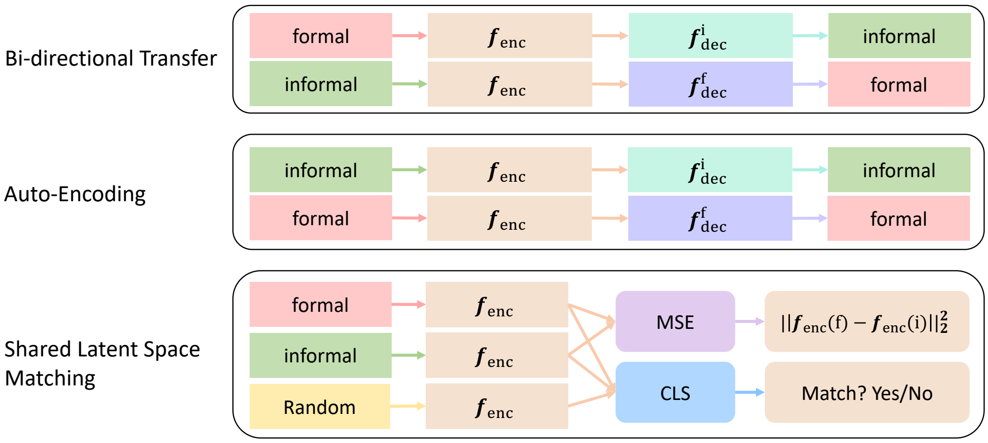
Bi-directional Transfer
formal
$\boldsymbol{f}_{\text{enc}}$
$\boldsymbol{f}^{\text{i}}_{\text{dec}}$
informal
informal
$\boldsymbol{f}_{\text{enc}}$
$\boldsymbol{f}^{\text{f}}_{\text{dec}}$
formal
Auto-Encoding
informal
$\boldsymbol{f}_{\text{enc}}$
$\boldsymbol{f}^{\text{i}}_{\text{dec}}$
informal
formal
$\boldsymbol{f}_{\text{enc}}$
$\boldsymbol{f}^{\text{f}}_{\text{dec}}$
formal
Shared Latent Space Matching
formal
informal
Random
$\boldsymbol{f}_{\text{enc}}$
$\boldsymbol{f}_{\text{enc}}$
$\boldsymbol{f}_{\text{enc}}$
MSE
CLS
$\|\boldsymbol{f}_{\text{enc}}(\text{f}) - \boldsymbol{f}_{\text{enc}}(\text{i})\|_2^2$
Match? Yes/No

Figure 1: Overview of our approach. Our S2S-SLS model has one encoder for all sentences, but two decoders for informal and formal styles, respectively. As also seen, our model has auxiliary matching losses in the shared latent space, and can be trained jointly with bi-directional tranfer and auto-encoding.

to style transfer. In Xu et al. (2018), a cycled reinforcement learning method is used to balance fluency and sentiment on unpaired data. Our joint training is inspired by cycled training, but works in a different way with a paralleled corpus.

Recently, researchers propose to synthesize pseudo-parallel data for style transfer. Zhang et al. (2018) build a word translation table by cross-domain word embeddings, and then use a phrased-based machine translation (PBMT) model to translate from one style to another. Subramanian et al. (2018) propose a denoising auto-encoding loss with online back-translation to generate pseudo-parallel data.

With parallel data. If we have sentences of different styles for the same content, we call it a parallel dataset. In this case, the machine translation framework (phrase-based or neural method) can be adopted to transfer one style to another. Xu et al. (2012) transfer modern English to Shakespeare's style. Rao and Tetreault (2018) create the GYAFC formality style transfer dataset, and have introduced several strong baselines by adopting phrase-based and neural machine translation with data augmentation techniques. Niu et al. (2018) and Xu et al. (2019b) tackle the formality style transfer problem with a multitask learning framework. Our paper follows the setting with parallel data, and extend the translation framework with a shared latent space for different styles and several matching losses.

3 Problem Formulation

Suppose we have a parallel dataset $\mathcal{D}_p = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ for formality transfer, where $\mathbf{x}_i = (x_{i,1}, \cdots, x_{i,s_i})$ is an informal sentence, and $\mathbf{y}_i = (y_{i,1}, \cdots, y_{i,t_i})$ is a formal sentence expressing the same meaning as $\mathbf{x}_i$. N is the number of pairs; s_i and t_i are the lengths of the respective sentence.

The goal of formality style transfer is to train a generation model, transferring a sentence in one style to another. This includes transferring an unseen informal sentence $\mathbf{x}_*$ to its formal expression $\mathbf{y}_*$, and *vice versa*. In real NLP tools such as writing assistants, it appears that informal-to-formal transfer makes more sense than formal-to-informal transfer. We follow Rao and Tetreault (2018) and regard the informal-to-formal transfer as the major goal of formality style transfer.

4 Our Approach

Our method generally follows the sequence-to-sequence (Seq2Seq) framework, but explores the shared latent space for both formal and informal sentences. We call our method Seq2Seq with Shared Latent Space (**S2S-SLS**). Figure 1 depicts the overall framework of S2S-SLS. It has a shared encoder $\boldsymbol{f}_{\text{enc}}$, but two decoders $\boldsymbol{f}^{\text{i}}_{\text{dec}}$ and $\boldsymbol{f}^{\text{f}}_{\text{dec}}$ for the informal and formal styles, respectively.

We will then describe our matching losses in detail. They ensure that the shared latent space of formal and informal styles does capture semantic information while eliminating the style.

We also introduce the joint training of informal-to-formal and formal-to-informal transfer, as well as auto-encoding training; they serve as a regularization for the model, which is important to formality style transfer when the data are limited.

Our method is applicable to different neural architectures, for example, recurrent neural networks (RNNs) with the attention mechanism and Transformer architectures, as used in our experiments. Due to space limit, we will describe the details in Appendix A.

4.1 Matching Losses

We would like all sentences to be encoded to a shared latent space regardless of the style, and thus we propose two simple yet effective auxiliary losses in our work.

Specifically, let $\mathbf{x}_i$ and $\mathbf{y}_i$ be a pair of informal and formal sentences with the same semantics. We hope the encoded representations of a pair of style-transferred sentences are close in the vector space. Therefore, we penalize their Euclidean distance

$$\mathcal{L}_{\text{dist}} = \sum_{i=1}^{N} \|\boldsymbol{f}_{\text{enc}}(\mathbf{x}_i) - \boldsymbol{f}_{\text{enc}}(\mathbf{y}_i)\|_2^2 \tag{1}$$

where $\boldsymbol{f}_{\text{enc}}(\mathbf{x}_i)$ represents the sentence-level representation of $\boldsymbol{x}_i$ produced by the encoder.

Such matching loss ensures that the latent space is shared between formal and informal styles, since optimizing (1) would ideally give the same encoded representation for both $\mathbf{x}_i$ and $\mathbf{y}_i$.

However, (1) alone may learn a trivial function, e.g., $\boldsymbol{f}_{\text{enc}} \equiv 0$, which obtains the minimum matching loss. Even jointly trained with the sequence-to-sequence loss, (1) would discourage the encoding of semantics in the latent space, which may be bypassed through the attention mechanism (Bahuleyan et al., 2018).

We therefore design another auxiliary loss, which aims to classify if the semantics of two sentences is the same. Concretely, let $\mathbf{s}_1$ and $\mathbf{s}_2$ be two sentences. For positive samples, they are a pair in the parallel corpus, i.e., $(\mathbf{s}_1, \mathbf{s}_2) = (\mathbf{x}_i, \mathbf{y}_i)$ for some i. We randomly pick a sentence $\mathbf{u}$ in the corpus $\mathcal{D}_p$, and form two negative samples $(\mathbf{x}_i, \mathbf{u})$ and $(\mathbf{y}_i, \mathbf{u})$.

We use the Siamese architecture (Bromley et al., 1994) for classification, where a logistic regression unit is applied to $[\boldsymbol{f}_{\text{enc}}(\mathbf{s}_1) \circ \boldsymbol{f}_{\text{enc}}(\mathbf{s}_2); \text{abs}(\boldsymbol{f}_{\text{enc}}(\mathbf{s}_1) - \boldsymbol{f}_{\text{enc}}(\mathbf{s}_2))]$, where $\circ$ represents element-wise product and $\text{abs}(\cdot)$ represents element-wise absolute value. These two operations ensure that the Siamese architecture is symmetric in $\mathbf{s}_1$ and $\mathbf{s}_2$, because they can be either informal or formal.

Let $\hat{t}_i \in (0, 1)$ be the output of the logistic regression and $t_i \in \{0, 1\}$ be the ground truth, indicating if a particular pair $\mathbf{s}_1$ and $\mathbf{s}_2$ have the same semantic. We impose a semantic classification loss as

$$\mathcal{L}_{\text{classification}} = \sum_{i=1}^{N} [-t_i \log \hat{t}_i - (1 - t_i) \log(1 - \hat{t}_i)] \tag{2}$$

In this way, the shared latent space has to capture the semantics of a sentence.

It should be pointed out that our matching loss works in a different way from the adversarial loss. In the non-parallel labeled setting (see Related Work), an adversarial loss is applied to classify the style of a sentence but the encoder is trained in an adversarial fashion, so that the encoded vectors are indistinguishable of styles in the population level. In our scenario, however, we have paired samples, and thus our matching losses are more suited for formality style transfer.

4.2 Learning Method

Despite the matching losses proposed in the previous subsection, the style-transfer generator is trained by sequence-aggregated cross-entropy loss. Suppose we are transferring an informal sentence $\mathbf{x}_i$ to its formal expression $\mathbf{y}_i$. The loss is

$$\mathcal{L}_{\text{i2f}} = -\sum_{i=1}^{N} \sum_{j=1}^{t_i} \log p(y_{i,j} | \boldsymbol{f}_{\text{enc}}(\mathbf{x}_i), y_{i,<j}) \tag{3}$$

where $\boldsymbol{f}_{\text{enc}}(\mathbf{x}_i)$ represents the encoder outputs for $\mathbf{x}_i$, and $y_{i,<j}$ represents $y_{i,1}, \cdots, y_{i,j-1}$. t_i is the length of $\mathbf{y}_i$ and N is the total number of samples.

Likewise, the loss of formal-to-informal transfer is

$$\mathcal{L}_{\text{f2i}} = -\sum_{i=1}^{N}\sum_{j=1}^{s_i} \log p(x_{i,j}|\boldsymbol{f}_{\text{enc}}(\mathbf{y}_i), x_{i,<j}) \tag{4}$$

where s_i is the length of $\mathbf{x}_i$.

As mentioned, the encoder $\boldsymbol{f}_{\text{enc}}$ is shared in (3) and (4). Such sharing of the encoder makes sense because the encoded latent space is assumed to capture semantics but no style information, so it can be used for style transfer in both directions.

Furthermore, auto-encoding loss can be applied to reconstruct a sentence from itself (either formal or informal), with losses $\mathcal{L}_{\text{i2i}}$ and $\mathcal{L}_{\text{f2f}}$ (details are not repeated). This prevents over-fitting when the dataset is small.

The joint training of style transfer in both directions as well as auto-encoding training reduces the model parameters by pooling things together. It also serves as a regularization, which is important in small-data training of formality style transfer. In summary, our training objective is

$$\mathcal{L} = \mathcal{L}_{\text{i2f}} + \mathcal{L}_{\text{f2i}} + \mathcal{L}_{\text{i2i}} + \mathcal{L}_{\text{f2f}} + \alpha\mathcal{L}_{\text{dist}} + \beta\mathcal{L}_{\text{classification}} \tag{5}$$

where α and β are hyper-parameters (set to 10 and 1 in our experiments) balancing the sequence-aggregated cross-entropy losses and matching losses. In the data augmentation and pre-training scenarios (see below), we observe that auto-encoding loss may have a negative effect (analyzed in Table 5); thus, $\mathcal{L}_{\text{i2i}}$ and $\mathcal{L}_{\text{f2f}}$ are not applied in these cases.

5 Experiments

5.1 Experimental Setup

We evaluate our model on the GYAFC benchmark dataset (Rao and Tetreault, 2018), which consists of human written informal-formal text pairs in two domains (Entertainment & Music and Family & Relationship) of the Yahoo Answer corpus. As mentioned, informal-to-formal is a more realistic application than formal-to-informal. Thus, we follow Rao and Tetreault (2018) and mainly address the informal-to-formal problem in our experiment.

Table 2 shows the statistics of the training, development, and test sets. In GYAFC, each sentence in the test set has four references, against which BLEU scores are computed.

Domain	Train	Dev	Test
Entertainment & Music (E&M)	52,595	2,877	1,416
Family & Relationship (F&R)	51,967	2,788	1,332

Table 2: Corpus statistics.

Our experiments are conducted on the F&R and E&M domains separately. To fully examine our approach, we have three settings with different data: (1) Data limited. We only use the parallel training set to train our model. (2) Data augmentation. We further use additional non-parallel data to enhance our method by creating pseudo-parallel data. And, (3) Pre-training. We use pre-trained models to initialize our encoder and decoders. In here, the treatment is different from our previous work, which focuses on combining the pretraining model with rule-based systems (Wang et al., 2019). The pretraining setting can also be considered as a way of augmenting models with large amounts of non-parallel data.

We implement our model with Tensorflow 1.12.0. All hyper-parameters are tuned on the validation dataset. In the data-augmentation scenario, we follow Rao and Tetreault (2018) to create pseudo-parallel data for data augmentation. In the pre-training scenario, we use the pre-trained GPT-2 model (117M) released by OpenAI[3] to initialize the shared encoder and the two decoders. More details of experimental

[3] https://github.com/openai/gpt-2

settings are introduced in the Appendix C. These variants are denoted by *S2S-SLS(RNN)*, *S2S-SLS(RNN)-Combined*, and *S2S-SLS(GPT)*, respectively.

5.2 Competing Methods

We compare our model with the following state-of-the-art methods in previous studies.

Rule-Based Approach: Rao and Tetreault (2018) use manually written rules to transfer an informal sentence to a formal one. Examples of the rules include capitalizing the first word and proper nouns, removing repeated punctuations, and handcrafting a list of expansion for abbreviations.

NMT-Baseline: Rao and Tetreault (2018) adopt a neural machine translation (NMT) system with the Seq2Seq network and the attention mechanism (Bahdanau et al., 2014) for formality style transfer.

NMT-Copy: Rao and Tetreault (2018) further use the copy mechanism (Gu et al., 2016) for enhancing the NMT baseline.

PBMT-Combined: Rao and Tetreault (2018) report the result given by phrase-based statistical machine translation (PBMT) with a self-training method (Ueffing, 2006). We further reproduce the PBMT-Combined results with our code and non-parallel language modeling data, denoted as *PBMT-Combined**.

NMT-Combined: Rao and Tetreault (2018) propose to synthesize a pseudo-parallel corpus by back-translation (Sennrich et al., 2016) with the PBMT-Combined system. Then, the NMT model is trained on the combination of the parallel and pseudo-parallel corpora. The method is thus called *NMT-Combined*. As our augmented data and pre-processing are different with *NMT-Combined*, we further report the result with our code base, denoted as *NMT-Combined**.

JTHTA: Xu et al. (2019b) propose a method that uses one Seq2Seq model to do bi-directional transfer with formality annotations. They design formality classifier-guided loss and two reconstruction losses for jointly training. Notice that they combine the two domains for training, which is considered as leveraging additional parallel data.

Bi-directional FT: Niu et al. (2018) propose a multi-task learning framework, which does bi-directional formality transfer, borrows parallel data of machine translation, and merges the two domains of GYAFC for training. We notice that they use four randomly seeded models as an ensemble in the decoding stage, which improves the performance by 1–2 BLEU scores in E&M and F&R domains. However, we focus on a single model.

GPT-Finetuning: For a fair comparison in the pre-training scenario, we fine-tune a Seq2Seq model implemented with GPT-2 as a baseline.

Unsupervised Approaches: It is also curious to see the performance of formality style transfer without using the alignment in the parallel corpus. We notice that Li et al. (2020) report formality style transfer results by learning from search towards a heuristic objective function. The BLEU scores are around or less than 60, significantly lower than using the alignment information. Therefore, they are not listed as competing methods in our experiment (Table 3).

5.3 Evaluation Metrics

We follow Rao et al. (2018) and evaluate the model by both automatic metrics and human judgements.

Formality: The automatic metric for formality is a machine learning classifier, which assesses the chance of success in formality transfer. Rao and Tetreault (2018) develop a feature engineering approach, requiring an extra labeled corpus for training, which is unfortunately not released. As a replacement, we train a GRU-based classifier using the training data of GYAFC for each domain (E&M and F&R). It achieves 92% accuracy, being a reasonable style classifier.

Meaning Preservation: We evaluate whether the meaning of the source sentence is preserved by a model trained on the Semantic Textual Similarity (STS) dataset, also following Rao and Tetreault (2018). In STS, the similarity of two sentences' meaning is on a scale of 1 to 6, where 6 means two sentences expressing the same meaning. We adopt the BERT-Base[4] model (Devlin et al., 2019) to extract sentence features. Then, the two sentences' representations (predicted by BERT) are concatenated and fed to a regression model to predict the similarity of meaning.

[4] `https://github.com/google-research/bert`

	Family & Relationships				Entertainment & Music			
	Formality	Meaning	BLEU	PINC	Formality	Meaning	BLEU	PINC
Original Informal	21.31	4.76	51.66	0	20.05	4.85	50.30	0
Formal Reference	81.53	3.20	100.00	65.59	79.61	3.78	100.00	66.93
Limited Data (Only Training Set)								
Rule-based†	57.50	**4.24**	66.36	27.75	48.69	**4.37**	60.35	28.26
NMT-Baseline†	79.31	3.40	68.26	49.35	77.38	3.24	58.26	54.94
NMT Copy†	**80.33**	3.39	68.09	49.68	**78.32**	3.22	58.66	54.61
S2S-SLS(RNN)	77.15	3.74	**72.58**	39.79	75.60	3.63	**65.85**	43.42
Data Augmentation with Unparallel Data								
PBMT-Combined†	77.45	3.82	72.40	44.02	73.50	3.90	66.87	45.26
PBMT-Combined*	72.70	**3.88**	71.75	40.99	66.94	**4.00**	64.91	43.27
NMT-Combined†	77.94	3.82	73.78	41.76	73.81	3.88	67.55	43.45
NMT-Combined*	76.75	3.77	73.31	41.40	69.70	3.96	67.66	39.69
S2S-SLS(RNN)-Combined	**79.35**	3.78	**74.62**	42.02	**74.33**	3.86	**68.41**	42.59
Additional Parallel Data								
JTHTA	-	-	74.43	-	-	-	69.63	-
Bi-directional FT†	74.54	3.97	75.33	39.39	70.61	3.98	72.01	41.74
Pre-training with Unparallel Data								
GPT-Finetuning*	**77.78**	3.74	75.61	43.75	76.03	**3.77**	70.33	46.62
S2S-SLS(GPT)	76.71	**3.80**	**76.61**	42.71	**78.62**	3.73	**71.10**	48.68

Table 3: Results for informal-to-formal transfer on F&R and E&R domains. * indicates that the baseline is implemented by ourselves. Numbers with † are obtained by evaluating outputs released by the respective paper. Otherwise, we quote the BLEU scores from previous papers. PINC reflects the dissimilarity to the original informal sentences, which does not correlate to the quality of style transfer well.

Overall: The overall quality of style-transferred sentences is evaluated by BLEU (Papineni et al., 2002) and PINC (Chen and Dolan, 2011). BLEU evaluates the n-gram overlap against references, and correlates the most with human annotation according to Rao and Tetreault (2018). PINC evaluates the dissimilarity between an output sentence and an input. A PINC score of 0 indicates that the input and output sentences are the same.

Human Evaluation: We further ask four human annotators to evaluate 200 randomly sampled results of different models. Specifically, an informal sentence and a formal sentence generated by a model are given to a human annotator for a 4-scale evaluation as follows. 4: Perfect. The output is very formal and preserves all content in the source sentence. 3: Good. The output is formal and most of the content is preserved. 2: Fair. The output is informal, or some important content is missed. 1: Bad. The output is very informal, or most of the content is missed. Notably, if all content is preserved but it is an informal sentence, it obtains 2.

Our human evaluation is conducted in a strictly blind fashion, meaning that all samples are shuffled so that the annotator does not know which model generated a particular sentence.

5.4 Experimental Results

Table 3 shows the automatic metrics of informal-to-formal transfer on the F&R and E&M domains. As seen, there is usually a trade-off among formality, fluency, and meaning. This is in fact expected because copying the input verbatim leads to perfect meaning preservation but no formality transfer. Our S2S-SLS model has a good balance among all these metrics and achieves the best overall scores of BLEU (the higher, the better) in the three scenarios.

Another observation is that the improvement of our S2S-SLS model is much larger in the data-limited setting, as it outperforms a plain NMT model by 4–7 BLEU scores in the two domains. The result verifies that it is difficult to train a sequence-to-sequence model only on the small parallel dataset; that our matching losses, joint training, and auto-encoding improve the performance to a large extent. In the data augmentation and pre-training scenarios, the gap is smaller: our S2S-SLS outperforms the baselines by about 1 BLEU point for both F&R and E&M. Nevertheless, it shows that our approach can be combined

	F&R	E&M
PBMT-Combined	3.28	3.23
NMT-Combined	3.34	3.32
S2S-SLS(RNN)-Combined	**3.43**	**3.41**
GPT-Finetuning	3.55	3.49
S2S-SLS(GPT)	**3.64**	**3.61**

Table 4: Human evaluation results. The numbers in bold are statistically significant compared with both baselines (t-test and bootstrap with $p \leq 0.05$).

	Famlity&Relationships				Entertainment&Music			
	FM	MN	BLEU	PINC	FM	MN	BLEU	PINC
Limited Data								
S2S-SLS	77.15	3.74	72.58	39.79	75.60	3.67	65.85	43.42
−Auto-encoding	82.48	3.31	66.62	51.86	81.87	3.07	54.71	58.18
−Joint training	76.72	3.69	71.90	40.12	74.91	3.60	64.08	44.78
−Matching loss	75.35	3.72	71.07	38.88	72.97	3.43	59.84	44.33
Data Augmentation								
S2S-SLS	79.35	3.78	74.62	42.02	74.33	3.86	68.41	42.59
+Auto-encoding	70.63	3.94	71.75	35.61	61.62	4.09	66.64	34.30
−Joint training	80.50	3.66	73.17	43.90	68.96	3.97	67.04	40.10
−Matching loss	78.56	3.76	73.78	41.57	75.55	3.88	67.47	42.13
With Pre-Training								
S2S-SLS	76.71	3.80	76.61	42.71	78.62	3.73	71.10	48.68
+Auto-encoding	72.94	3.84	73.02	38.10	68.43	3.97	67.91	39.56
−Joint training	77.12	3.76	75.98	43.41	74.51	3.75	69.54	46.30
−Matching loss	78.95	3.73	75.77	45.25	77.89	3.73	70.67	48.46

Table 5: Ablation test on the F&R and E&M domains. FM: Formality. MN: Meaning.

with other data augmentation techniques for further improvement. The result of S2S-SLS(GPT2) further confirms that our model consistently performs well with different neural architectures.

We also see that S2S-SLS(GPT) outperforms JTHTA in the BLEU score, where JTHTA uses additional parallel data. The performance of our single S2S-SLS(GPT) model is also close to Bi-directional FT with model ensembles.

Table 4 presents the results of human evaluation. It shows consistent evidence that S2S-SLS outperforms the baselines in both F&R and E&M domains. For the two domains, the Spearman's rank correlation between human annotators are 0.83 and 0.75, and the Pearson Correlation between human annotations and BLEU scores are 0.53 and 0.51. This confirms that 1) humans reach a relatively high agreement; that 2) the automatic metrics in Table 3 are indeed correlated to human judgement.

5.5 Model Analysis

We conduct an ablation test on the F&R and E&M domains in Table 5, where we analyze the effect of matching losses, joint training (informal-to-informal and formal-to-informal), as well as auto-encoding losses. We see that the auto-encoding loss is particularly effective in the data-limited scenario, but it has a negative effect when a large non-parallel corpus is used. A plausible explanation is that the auto-encoding loss can be thought of as a regularization term, which prevents over-fitting when the size of training data is small, but it restricts the model capacity if data are adequate. Also, the PINC score increases if we have auto-encoding training, indicating that the output is more similar to the input. Therefore, we do not apply the auto-encoding loss in the data augmentation and pre-training scenarios.

For the other two components—namely, the matching losses and joint training of bi-directional style transfer—we see that both of them play a role in our model.

We further conduct a qualitative analysis of the encoded latent space effectiveness in Figure 2, where the encoded representations of nine pairs of informal and formal sentences are shown by t-SNE

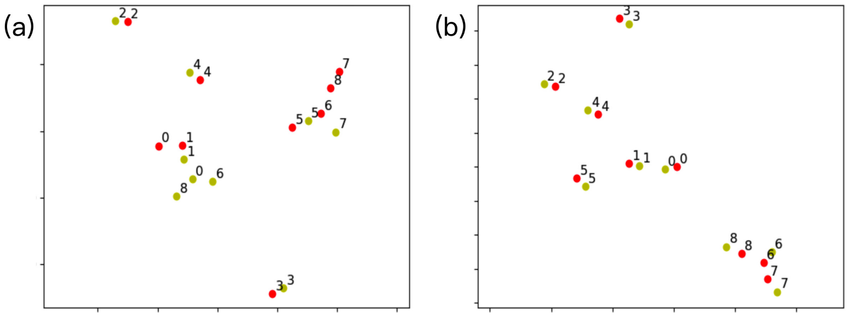

Figure 2: t-SNE plot of the encoded sentence representations of informal (red) and formal (yellow) sentences, detailed in Appendix B. (a) NMT-Combined*; (b) S2S-SLS(RNN)-Combined.

plot (Rauber et al., 2016). We select nine examples on which S2S-SLS(RNN) performs better than NMT-Combined* to verify whether a shared latent space helps formality style transfer. We see clearly that, for the better transferred samples, the latent space is indeed more related to meaning but less to style. The informal and formal sentences of these data points (0–9 in Figure 2) and model outputs are shown in the Appendix B. Randomly selected examples are also presented in Appendix B as a case study.

6 Conclusion

In this paper, we propose a novel sequence-to-sequence model with shared latent space for formality style transfer, called S2S-SLS. We observe that the formality style transfer task is different from machine translation in the size of data and the relationship between source and target sequences. To address these issues, our S2S-SLS model uses a single encoder to capture sentences of different formality styles, and our matching losses ensure that the latent space captures semantic information while eliminating style.

Experimental results show that our approach significantly improves baseline models in the data-limited setting. In the data-augmentation and pre-training scenarios where a large scale non-parallel corpus is used to augment the model, our approach still outperforms the baseline methods. The experiments also show that our method can be adapted to different neural architectures. The ablation test and case study provide further analysis of our S2S-SLS model, showing the effect of our auxiliary matching losses, joint training, and auto-encoding training.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (Grant Nos. U1636211, 61672081, and 61370126), the Beijing Advanced Innovation Center for Imaging Technology (Grant No. BAICIT-2016001), the Fund of the State Key Laboratory of Software Development Environment (Grant No. SKLSDE-2019ZX-17), the Alberta Machine Intelligence Institute (Amii) Fellow Program, the Canadian CIFAR AI Chair Program, and the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant No. RGPIN-2020-04465. This research was also enabled in part by the support of Compute Canada (`www.computecanada.ca`). Zhoujun Li is the corresponding author.

References

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

Hareesh Bahuleyan, Lili Mou, Olga Vechtomova, and Pascal Poupart. 2018. Variational attention for sequence-to-sequence models. In *COLING*, pages 1672–1682.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xin-yu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. In *ACL*, pages 6008–6019.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

Ondrej Bojar, Jindrich Helcl, Tom Kocmi, Jindrich Libovický, and Tomás Musil. 2017. Results of the WMT17 neural MT training task. In *WMT*, pages 525–533.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using A "siamese" time delay neural network. In *NIPS*, pages 737–744.

David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pages 190–200.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, pages 2172–2180.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *AAAI*, pages 663–670.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*, pages 2672–2680.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL*, pages 1631–1640.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *ACL*, pages 690–696.

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

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *ACL*, pages 424–434.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. 2017. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *NAACL-HLT*, pages 1865–1874.

Jingjing Li, Zichao Li, Lili Mou, Xin Jiang, Michael R Lyu, and Irwin King. 2020. Unsupervised text generation by learning from search. In *NeurIPS*.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *COLING*, pages 1008–1021.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *TACL*, 4:61–74.

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*, pages 129–140.

Paulo E. Rauber, Alexandre X. Falcão, and Alexandru C. Telea. 2016. Visualizing time-dependent data using dynamic t-sne. In *EuroVis*, pages 73–77.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *WMT*, pages 371–376.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. *arXiv preprint arXiv:1811.00552*, abs/1811.00552.

Nicola Ueffing. 2006. Self-training for machine translation. In *NIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In *EMNLP-IJCNLP*, pages 3573–3578.

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

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *ACL*, pages 979–988.

Peng Xu, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019a. Unsupervised controllable text generation with global variation discovery and disentanglement. *CoRR*, abs/1905.11975.

Ruochen Xu, Tao Ge, and Furu Wei. 2019b. Formality style transfer with hybrid textual annotations. *arXiv preprint arXiv:1903.06353*.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *arXiv preprint arXiv:1808.07894*, abs/1808.07894.

A Model Architectures

Here, we describe the architecture details of RNN-Based S2S-SLS and GPT-Based S2S-SLS.

A.1 RNN-Based Architecture

Given a sentence, we first tokenize it by the byte-pair encoding (Sennrich et al., 2015, BPE) and obtain a sequence of word pieces. They are represented by embeddings $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K$, pretrained on the Yahoo Answer L6 Corpus[5] with the FastText model (Bojanowski et al., 2017).

We implement the encoder $\boldsymbol{f}_{\text{enc}}$ as a bi-directional recurrent neural network (RNN) with gated recurrent units (Cho et al., 2014, GRU). Formally, let $\overrightarrow{\boldsymbol{h}}_i$ and $\overleftarrow{\boldsymbol{h}}_i$ be the encoded vectors for each direction. They are given by

$$\overrightarrow{\boldsymbol{h}}_i = f_{\overrightarrow{\text{GRU}}}(\overrightarrow{\boldsymbol{h}}_{i-1}, \boldsymbol{w}_i), \ \ \overleftarrow{\boldsymbol{h}}_i = f_{\overleftarrow{\text{GRU}}}(\overleftarrow{\boldsymbol{h}}_{i+1}, \boldsymbol{w}_i) \tag{6}$$

The output of the encoder is the concatenation of both GRU states, i.e., $\boldsymbol{h}_i = \left[\overrightarrow{\boldsymbol{h}}_i; \overleftarrow{\boldsymbol{h}}_i\right]$, for $i = 1, \cdots, K$. For this example, $\boldsymbol{h}_k$ is the output of $\boldsymbol{f}_{\text{enc}}$.

On the target side, we have two decoders $f_{\text{dec}}^{\text{i}}$ and $f_{\text{dec}}^{\text{f}}$ for each style, working in the same way but parametrized differently. Consider the decoder for formal sentences. It is yet another GRU-RNN, enhanced by an attention mechanism. Let $\boldsymbol{h}'_j$ be the GRU state at the jth step of decoding. We compute an attention vector c_i, which is a convex combination of $\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_t\}$

$$\boldsymbol{c}_j = \sum_{i=1}^{K} \alpha_{j,i} \boldsymbol{h}_i, \tag{7}$$

where $\alpha_{j,i}$ is the attention weight, computed by

$$\alpha_{j,i} = \frac{\exp\{e_{j,i}\}}{\sum_{i'=1}^{K} \exp\{e_{j,i'}\}}, \quad e_{j,i} = [\boldsymbol{h}'_{j-1}; \boldsymbol{w}'_{j-1}]\mathbf{W}_\alpha \boldsymbol{h}_i \tag{8}$$

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

where $\boldsymbol{v}$ and $\mathbf{W}_\alpha$ are attention parameters, and $\boldsymbol{w}'_j$ is the embedding of jth BPE word piece in the target.

The attention vector is fed to the GRU update as

$$\boldsymbol{h}'_j = f^{(\text{dec})}_{\text{GRU}}(\boldsymbol{h}'_{j-1}, [\boldsymbol{w}'_{j-1}; \boldsymbol{c}_j]) \tag{9}$$

At each step of the decoder, it predicts the next word piece with a probabilistic distribution over the entire vocabulary (word pieces) by a softmax layer:

$$\boldsymbol{p}_j = \text{softmax}(\mathbf{W}_p \boldsymbol{h}'_j + \boldsymbol{b}_p), \tag{10}$$

where $\mathbf{W}_p$ and $\boldsymbol{b}_p$ are the parameters.

Likewise, the decoder for the informal style, denoted by $f^{\text{i}}_{\text{dec}}$, is constructed in similar way and details are not repeated.

A.2 Transformer-Based Architecture

Our transformer-based architecture is slightly different from the classic Transformer (Vaswani et al., 2017). We implement the encoder and decoder with GPT-2 blocks for using the pre-trained GPT-2 model to initialize our model conveniently. So our encoder uses masked multi-head attention mechanism, and the decoder does not have the cross-attention to the encoder outputs. Instead, we use a single masked multi-head attention mechanism for the concatenation of encoder outputs and decoder inputs.

Formally, for a sentence with K word pieces $(\boldsymbol{wp}_1, \cdots, \boldsymbol{wp}_K)$, the encoder outputs of layer l are $(\boldsymbol{h}^l_1, \cdots, \boldsymbol{h}^l_K)$, denoted as Enc^l. Let $\boldsymbol{h}'^l_j$ be the l-th block's output of the decoder for the j-th step of decoding and Block'_l be the l-th block of the decoder. Then $\boldsymbol{h}'^l_j$ can be calculated as:

$$\boldsymbol{h}'^l_j = \text{Block}'_l([\text{Enc}^{l-1}; \text{Dec}^{l-1}_{1:j}]) \tag{11}$$

where $\text{Dec}^{l-1}_{1:j}$ represents $(\boldsymbol{h}'^{l-1}_1, \cdots, \boldsymbol{h}'^{l-1}_j)$.

To produce a sentence-level representation, we append a global flag [REP] to the word pieces, and use the encoder output of [REP] as the encoded feature $\boldsymbol{f}_{\text{enc}}$, which is then fed to the decoder.

B Case Study

As mentioned, we select several examples on which S2S-SLS(RNN) performs better to verify whether the representation in the shared latent space is indeed more style-independent. Table 6 shows the nine groups of sentences we selected, including informal sentences, formal references, the output of *NMT-Combined**, and the output of *S2S-SLS(RNN)-Combined*.

We further random sample 10 groups of examples which are showed in Table 7. Although S2S-SLS(RNN)-Combined produces some same results with NMT-Combined*, we can see that it does better at grammar and content in more sentences.

C Experimental Details

C.1 Formality Classifier

We tokenize the GYAFC data by the byte pairwise encoding. The word embeddings are pre-trained using FastText with in-domain data of Yahoo Answers. The dimension of the pre-trained embedding is 300. We use a one-layer bidirectional GRU for modeling sentence-level representation. The number of hidden units for each direction is 32. Based on the final hidden state of the bidirectional GRU, we adopt a multilayer perceptron for predicting the categories (informal or formal). We use the Adam algorithm to train our model with a batch size 256. We set the initial learning rate as 0.001. We employ early stopping as a regularization strategy and we find that the best result is always achieved within 10 epochs in our experiments.

#	Original Informal	Formal Reference	S2S-SLS(RNN)-Combined	NMT-Combined*
1	the guy i like is mad at me.	The man I like is angry with me.	The man I like is mad at me .	The guy I am mad at me .
2	what the hell r ya talking abt?	What are you talking about?	What are you talking about ?	What the hell are you talking about ?
3	and she still lets me play with my trucks and watch cartoons!!!!!	She still lets me play with my trucks and watch cartoons!	She still lets me play with my trucks and watch cartoons .	She still let me play with my trucks and watch watching ons !
4	Sounds like a rhetorical question :)	It sounds like a rhetorical question.	It sounds like a rhetorical question .	It sounds like a rhecal question .
5	you can join a community site with live chat and webcam chat	You can join a community site with live chat and webcam chat.	You can join a community site with live chat and webcam chat .	You can join a church site with live chat and webcam chat chat .
6	or just you wanna say that!?	Do you want to say that?	Do you want to say that ?	Or just you want to say that !
7	he then asked me, can i come with you?	He then asked me "Could I come with you?".	He asked me , " Can I come with you ? "	He asked me , I can come with you ?
8	u'll find a man who really deserves u one day).	You will find a man who really deserves you one day.	You will find a man who really deserves you one day .	You will find a man who deserves you one day .
9	i am waiting till im married but when is it too far?	I am waiting until I am married, but when is it too far?	I am waiting until I am married , but when is it too far ?	I am waiting until I am married , but when it is too far .

Table 6: Examples of informal sentences, formal references, and formal outputs produced by S2S-SLS(RNN)-Combined and NMT-Combined*.

C.2 PBMT

We use Moses (Koehn et al., 2007) for our experiments. A 5-gram language model is trained by KenLM (Heafield et al., 2013). We tokenize the data by Moses. We turn all the data into lowercase for training and use our own script to correct the wrong cases in the result of PBMT, because we find in our experiments that this is better than training with the original case. Rao and Tetreault (2018) train PBMT on the output of a rule-based system. Since their code is not publicly available, we design our own rules and successfully reproduce such a rule-based approach for training our PBMT model.

C.3 Data Augmentation

Usually, a manually annotated parallel corpus is small. It would be beneficial to make use of a large unlabeled corpus for data augmentation. We follow Rao and Tetreault (2018) and augment our data with the unlabeled Yahoo Answer corpus. However, the augmented corpus and engineering details of Rao and Tetreault (2018) are not available. Therefore, we develop our own data augmentation method as follows.

We denote the large non-parallel dataset by $\mathcal{D}_u = \{\mathbf{u}_i\}_{i=1}^M$, where M is the number of unlabeled sentences. In this work, the large unlabeled corpus is constructed from the Yahoo Answers; most of these sentences are informal, although some may be more formal than others. Generally, the corpus is noisy, requiring *ad hoc* preprocessing before data augmentation.

We first train a GRU-based classifier with the parallel corpus $\mathcal{D}_p$ to classify if a sentenece is formal or informal. It achieves 92% accuracy on the validation data and is able to classify the formality of a sentence reasonably well. Then, we assign a formality score for each sentence in $\mathcal{D}_u$; the score is the predicted probability of being formal. We select those sentences with a high confidence (≥ 0.8) of being formal, and obtain the pseudo-formal corpus $\widetilde{\mathcal{D}}_f = \{\mathbf{u}_i^{(\mathrm{f})}\}_{i=1}^L$, where L is the number of samples in the augmented dataset.

Then a phrase-based machine translation (PBMT) model is trained for formal-to-informal transfer using the parallel corpus $\mathcal{D}_p$, whose language model is trained on the sentences in the corresponding domain of the Yahoo Answers L6 Corpus. For each pseudo-formal sentence $\mathbf{u}_i^{(\mathrm{f})} \in \widetilde{\mathcal{D}}_f$, we generate an informal sentence $\widetilde{\mathbf{x}}_i$ with PBMT. In this way, a pseudo-parallel dataset $\widetilde{\mathcal{D}}_p = \{\widetilde{\mathbf{x}}_i, \mathbf{u}_i^{(\mathrm{f})}\}_{i=1}^{L'}$ is constructed with some post-processing (L' is the number of samples).[6] Finally, we merge $\mathcal{D}_p$ and $\widetilde{\mathcal{D}}_p$, in which the sentences in D_p is up-sampled to balance the synthetic corpus and the human-constructed one. For data

[6] Pseudo-parallel data are filtered by length, repetition words, the language model score and the number of <UNK>.

Original Informal	S2S-SLS(RNN)-Combined	NMT-Combined*
After i while i t passes and i still love him	After I while I am not passes and I still love him .	After I while I still love him .
though i'm 19, i like anything up to 30?	I am nineteen years old .	I am 19 years old , I like anything up to 30 ?
Has he just dumped a girl and left before?	Has he broke up with a girl and left before ?	Has he just dumped a girl and left before ?
his answer is good enough for me...	His answer is good enough for me .	His answer is good enough for me .
(i mess up alot at times) but she don't forget.	I mess up a lot at times , but she does not forget .	I mess up a lot at times , but she do not forget .
i think hse will crash him!	I think hse will crash him .	I think hse will crash him .
Because believe it or not, women hate cocky men.	Because believe it or not , women dislike arrogance .	Because it or not , women dislike arrogant men .
Its really fun to read, i think!	It is really fun to read .	It is really fun to read .
RAP..4everrr.. .. also i like metal, & rap/rock.. like Linkin park.. but plain rock SUCKS..	I like metal and rap music .	I like metal , and rap rock music .
my favorite character is jess, he is so cute	My favorite character is Jess , he is so cute .	My favorite character is Jess , he is so cute .

Table 7: Random sampled examples of informal sentences and formal outputs produced by S2S-SLS(RNN)-Combined and NMT-Combined*.

augmentation, we synthesize 1.7M pairs for the F&R domain and 3.4M pairs for the E&M domain.

Self-training (Ueffing, 2006) and back-translation (Sennrich et al., 2016) are common methods for data augmentation of generation tasks. We tried both, but finally adopted the above *ad hoc* augmentation method because it demonstrates the best performance in our experiment.

Although the PBMT models are trained on the result of the rule-based approach, we find that we can still improve the quality of the augmented data by rule-based scripts. Concretely, we post-process the augmented data with our rule-based scripts. Then we remove sentences that contain URLs, have more than three unknown words, and are shorter than 6 or longer than 25 words. We also remove pairs if informal and formal sentences are the same. To further reduce the noise of the augmented data, we use a GRU-based language model to filter out sentences with too low or too high scores for each domain. The training data of the language model is the in-domain data of the Yahoo Answers.

C.4 RNN-Based S2S-SLS

The encoder and decoder are one-layer GRU with 300 hidden units. The dimension of the pre-trained embedding is 300. The embeddings are shared for both source and target. We employ the Adam algorithm (Kingma and Ba, 2014) to train our model with a batch size of 128. We set the initial learning rate as 0.001, and reduce it by half if the BLEU score on validation decreases. We stop training if validation BLEU decreases in two successive epochs.

C.5 GPT-Based S2S-SLS

We use the pre-trained GPT-2 model (117M) released by OpenAI[7] to initialize the shared encoder and the two decoders. We use a GTX-1080Ti GPU to run the model. Our batch size is only 16 due to the limitation of the GPU memory. We use Adam (Kingma and Ba, 2014) to optimize our model with an initial learning rate 0.0001. We also use learning rate decay and early stopping strategies as in RNN-based S2S-SLS.

[7] https://github.com/openai/gpt-2