Multi-Task Neural Models for Translating Between Styles Within and Across Languages

Xing Niu
University of Maryland
xingniu@cs.umd.edu

Sudha Rao
University of Maryland
raosudha@cs.umd.edu

Marine Carpuat
University of Maryland
marine@cs.umd.edu

Abstract

Generating natural language requires conveying content in an appropriate style. We explore two related tasks on generating text of varying formality: monolingual formality transfer and formality-sensitive machine translation. We propose to solve these tasks jointly using multi-task learning, and show that our models achieve state-of-the-art performance for formality transfer and are able to perform formality-sensitive translation without being explicitly trained on style-annotated translation examples.

1 Introduction

Generating language in the appropriate style is a requirement for applications that generate natural language, as the style of a text conveys important information beyond its literal meaning (Hovy, 1987). Heylighen and Dewaele (1999) and Biber (2014) have argued that the formal-informal dimension is a core dimension of stylistic variation. In this work, we focus on the problem of generating text for a desired formality level. It has been recently studied in two distinct settings: (1) Rao and Tetreault (2018) addressed the task of *Formality Transfer* (FT) where given an informal sentence in English, systems are asked to output a formal equivalent, or vice-versa; (2) Niu et al. (2017) introduced the task of *Formality-Sensitive Machine Translation* (FSMT), where given a sentence in French and a desired formality level (approximating the intended audience of the translation), systems are asked to produce an English translation of the desired formality level. While FT and FSMT can both be framed as Machine Translation (MT), appropriate training examples are much harder to obtain than for traditional machine translation tasks. FT requires sentence pairs that express the same meaning in two different styles, which rarely occur naturally and are therefore only available in small quantities. FSMT can draw from existing parallel corpora in diverse styles, but would ideally require not only sentence pairs, but e.g., sentence triplets that contain a French input, its formal English translation, and its informal English translation.

We hypothesize that FT and FSMT can benefit from being addressed jointly, by sharing information from two distinct types of supervision: sentence pairs in the same language that capture style difference, and translation pairs drawn from corpora of various styles. Inspired by the benefits of multi-task learning (Caruana, 1997) for natural language processing tasks in general (Collobert and Weston, 2008; Liu et al., 2015; Luong et al., 2016), and for multilingual MT in particular (Johnson et al., 2017), we introduce a model based on Neural Machine Translation (NMT) that jointly learns to perform both monolingual FT and bilingual FSMT. As can be seen in Figure 1, given an English sentence and a tag (formal or informal), our model paraphrases the input sentence into the desired formality. The same model can also take in a French sentence, and produce a formal or an informal English translation as desired.

Designing this model requires addressing several questions: Can we build a single model that performs formality transfer in both directions? How to best combine monolingual examples of formality transfer and bilingual examples of translation? What kind of bilingual examples are most useful for the joint task? Can our joint model learn to perform FSMT without being explicitly trained on style-annotated translation examples? We explore these questions by conducting an empirical study on English FT and

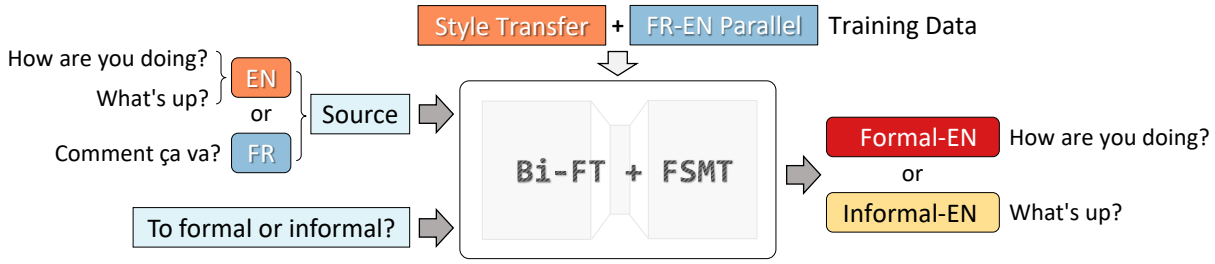


Figure 1: System overview: our multi-task learning model can perform both bi-directional English formality transfer and translate French to English with desired formality. It is trained jointly on monolingual formality transfer data and bilingual translation data.

French-English FSMT, using both automatic and human evaluation. Our results show the benefits of the multi-task learning approach, improving the state-of-the-art on the FT task, and yielding competitive performance on FSMT without style-annotated translation examples. Along the way, we also improve over prior results on FT using a single NMT model that can transfer between styles in both directions.

2 Background

Style Transfer can naturally be framed as a sequence to sequence translation problem given sentence pairs that are paraphrases in two distinct styles. These parallel style corpora are constructed by creatively collecting existing texts of varying styles, and are therefore rare and much smaller than machine translation parallel corpora. For instance, Xu et al. (2012) scrape modern translations of Shakespeare's plays and use a phrase-based MT (PBMT) system to paraphrase Shakespearean English into/from modern English. Jhamtani et al. (2017) improve performance on this dataset using neural translation model with pointers to enable copy actions. The availability of parallel standard and simple Wikipedia (and sometimes additional human rewrites) makes text simplification a popular style transfer task, typically addressed using machine translation models ranging from syntax-based MT (Zhu et al., 2010; Xu et al., 2016), phrase-based MT (Coster and Kauchak, 2011; Wubben et al., 2012) to neural MT (Wang et al., 2016) trained via reinforcement learning (Zhang and Lapata, 2017).

Naturally occurring examples of parallel formal-informal sentences are harder to find. Prior work relied on synthetic examples generated based on lists of words of known formality (Sheikha and Inkpen, 2011). This state of affairs recently changed, with the introduction of the first large scale parallel corpus for formality transfer, GYAFC (Grammarly's Yahoo Answers Formality Corpus). 110K informal sentences were collected from Yahoo Answers and they were rewritten in a formal style via crowd-sourcing, which made it possible to benchmark style transfer systems based on both PBMT and NMT models (Rao and Tetreault, 2018). In this work, we leverage this corpus to enable multi-task FT and FSMT.

Recent work also explores how to perform style transfer without parallel data. However, this line of work considers transformations that alter the original meaning (e.g., changes in sentiment or topic), while we view style transfer as meaning-preserving. An auto-encoder is used to encode a sequence to a latent representation which is then decoded to get the style transferred output sequence (Mueller et al., 2017; Hu et al., 2017; Shen et al., 2017; Fu et al., 2018; Prabhumoye et al., 2018).

Style in Machine Translation has received little attention in recent MT architectures. Mima et al. (1997) improve rule-based MT by using extra-linguistic information such as speaker's role and gender. Lewis et al. (2015) and Niu and Carpuat (2016) equate style with domain, and train conversational MT systems by selecting in-domain (i.e. conversation-like) training data. Similarly, Wintner et al. (2017) and Michel and Neubig (2018) take an adaptation approach to personalize MT with gender-specific or speaker-specific data. Other work has focused on specific realizations of stylistic variations, such as T-V pronoun selection for translation into German (Sennrich et al., 2016a) or controlling voice (Yamagishi et al., 2016). In contrast, we adopt the broader range of style variations considered in our prior work, which introduced the FSMT task (Niu et al., 2017): in FSMT, the MT system takes a desired formality level as

an additional input, to represent the target audience of a translation, which human translators implicitly take into account. This task was addressed via n-best re-ranking in phrase-based MT — translation hypotheses whose formality are closer to desired formality are promoted.

By contrast, in this work we use neural MT which is based on the **Attentional Recurrent Encoder-Decoder** model (Bahdanau et al., 2015; Luong et al., 2016). The input is encoded into a sequence of vector representations while the decoder adaptively computes a weighted sum of these vectors as the context vector for each decoding step.

In the joint model, we employ **Side Constraints** as the formality input to restrict the generation of the output sentence (Figure 1). Prior work has successfully implemented side constraints as a special token added to each source sentence. These tokens are embedded into the source sentence representation and control target sequence generation via the attention mechanism. Sennrich et al. (2016a) append <T> or <V> (i.e. T-V pronoun distinction) to the source text to indicate which pronoun is preferred in the German output. Johnson et al. (2017) and Niu et al. (2018) concatenate parallel data of various language directions and mark the source with the desired output language to perform multilingual or bi-directional NMT. Kobus et al. (2017) and Chu et al. (2017) add domain tags for domain adaptation in NMT.

3 Approach

We describe our unified model for performing FT in both directions (Section 3.1), our FSMT model with side constraints (Section 3.2) and finally our multi-task learning model that jointly learns to perform FT and FSMT (Section 3.3). All models rely on the same NMT architecture: attentional recurrent sequence-to-sequence models.

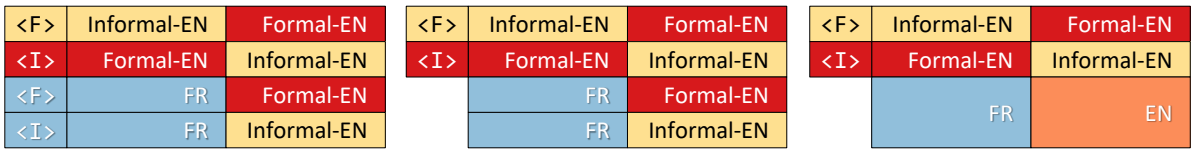
3.1 Bi-Directional Formality Transfer

Rao and Tetreault (2018) used independent neural machine translation models for each formality transfer direction (`informal`→`formal` and `formal`→`informal`). Inspired by the bi-directional NMT for low-resource languages (Niu et al., 2018), we propose a unified model that can handle either direction — we concatenate the parallel data from the two directions of formality transfer and attach a tag to the beginning of each source sentence denoting the desired target formality level i.e. <F> for transferring to formal and <I> for transferring to informal. This enables our FT model to learn to transfer to the correct style via attending to the tag in the source embedding. We train an NMT model on this combined dataset. Since both the source and target sentences come from the same language, we encourage their representations to lie in the same distributional vector space by (1) building a shared Byte-Pair Encoding (BPE) model on source and target data (Sennrich et al., 2016b) and (2) tying source and target word embeddings (Press and Wolf, 2017).

3.2 Formality-Sensitive Machine Translation with Side Constraints

Inspired by Sennrich et al. (2016a), we use side constraints on parallel translation examples to control output formality. At training time, this requires a tag that captures the formality of the target sentence for every sentence pair. Given the vast range of text variations that influence style, we cannot obtain tags using rules as for T-V pronoun distinctions (Sennrich et al., 2016a). Instead, we categorize French-English parallel data into formal vs. informal categories by comparing them to the informal and formal English from the GYAFC corpus.

We adopt a data selection technique, Cross-Entropy Difference (CED) (Moore and Lewis, 2010), to rank English sentences in the bilingual corpus by their relative distance to each style. First, we consider formal English as the target style and define $CED(s) = H_{formal}(s) - H_{informal}(s)$, where $H_{formal}(s)$ is the cross-entropy between a sentence s and the formal language model. Smaller CED indicates an English sentence that is more similar to the formal English corpus and less similar to the informal English corpus. We rank English sentences by their CED scores and select the top N sentences (choice of N discussed in Section 6). Pairing these N English sentences with their parallel French source, we get the formal sample of our bilingual data. Similarly, we construct the informal sample using informal English as the target style. Finally, we combine the formal and the informal samples, attach the <F> and <I>



(a) Formality tags on bilingual data + 2-Style selection

(b) No formality tags on bilingual data + 2-Style selection

(c) No formality tags on bilingual data + Random selection

Figure 2: The training data used for multi-task learning models. The bi-directional formality transfer data and the bilingual data (e.g. FR-EN) of equivalent size are always concatenated.

tags to corresponding source French sentences (i.e. the bottom two rows of data in Figure 2a) and train an NMT model for our FSMT task.

3.3 Multi-Task Learning

We propose a multi-task learning model to jointly perform FT and FSMT using a many-to-one (i.e. multi-language to English) sequence to sequence model (Luong et al., 2016). Following Johnson et al. (2017), we implement this approach using shared encoders and decoders. This approach can use existing NMT architectures without modifications. We design three models to investigate how to best incorporate side constraints at training time, and the benefits of sharing representations for style and language.

MultiTask-tag-style is a straightforward combination of the transfer and translation models above. We hypothesize that using the bilingual parallel data where English is the target could enhance English FT in terms of target language modeling, especially when the bilingual data has similar topics and styles. We therefore combine equal sizes of formality tagged training data (selected as described in Section 3.2) from our FT and FSMT tasks in this configuration (Figure 2a).

MultiTask-style is designed to test whether formality tags for bilingual examples are necessary. We hypothesize that the knowledge of controlling target formality for the FSMT task can be learned from the FT data since the source embeddings of formality tags are shared between the FT and the FSMT tasks. We therefore combine the formality tagged FT data with the MT data without their tags (Figure 2b).

MultiTask-random investigates the impact of the similarity between formality transfer and bilingual examples. Selecting bilingual data which is similar to the GYAFC corpus is not necessarily beneficial for the FSMT task especially when French-English bilingual examples are drawn from a domain distant from the GYAFC corpus. In this configuration, we test how well our model performs FSMT if bilingual examples are randomly selected instead (Figure 2c).

4 Experimental Setup

FT data: We use the GYAFC corpus introduced by Rao and Tetreault (2018) as our FT data. This corpus consists of 110K informal sentences from two domains of Yahoo Answers (*Entertainment and Music (E&M)* and *Family and Relationships (F&R)*) paired with their formal rewrites by humans. The train split consists of 100K informal-formal sentence pairs whereas the dev/test sets consist of roughly 5K source-style sentences paired with four reference target-style human rewrites for both transfer directions.

FSMT data: We evaluate the FSMT models on a large-scale French to English (FR-EN) translation task. Examples are drawn from OpenSubtitles2016 (Lison and Tiedemann, 2016) which consists of movie and television subtitles and is thus more similar to the GYAFC corpus compared to news or parliament proceedings. This is a noisy dataset where aligned French and English sentences often do not have the same meaning, so we use a bilingual semantic similarity detector to select 20,005,000 least divergent examples from ∼27.5M deduplicated sentence pairs in the original set (Vyas et al., 2018). Selected examples are then randomly split into a 20M training pool, a 2.5K dev set and a 2.5K test set.

Preprocessing: We apply four pre-processing steps to both FT and MT data: normalization, tokenization, true-casing, and joint source-target BPE with 32,000 operations for NMT (Sennrich et al., 2016b).

NMT Configuration: We use the standard attentional encoder-decoder architecture implemented in the Sockeye toolkit (Hieber et al., 2017). Our translation model uses a bi-directional encoder with a single

LSTM layer (Bahdanau et al., 2015) of size 512, multilayer perceptron attention with a layer size of 512, and word representations of size 512. We apply layer normalization and tie the source and target embeddings as well as the output layer's weight matrix. We add dropout to embeddings and RNNs of the encoder and decoder with probability 0.2. We train using the Adam optimizer with a batch size of 64 sentences and checkpoint the model every 1000 updates (Kingma and Ba, 2015). Training stops after 8 checkpoints without improvement of validation perplexity. We decode with a beam size of 5. We train four randomly seeded models for each experiment and combine them in a linear ensemble for decoding.[1]

5 Evaluation Protocol

5.1 Automatic Evaluation

We evaluate both FT and FSMT tasks using BLEU (Papineni et al., 2002), which compares the model output with four reference target-style rewrites for FT and a single reference translation for FSMT. We report case-sensitive BLEU with standard WMT tokenization.[2] For FT, Rao and Tetreault (2018) show that BLEU correlates well with the overall system ranking assigned by humans. For FSMT, BLEU is an imperfect metric as it conflates mismatches due to translation errors and due to correct style variations. We therefore turn to human evaluation to isolate formality differences from translation quality.

5.2 Human Evaluation

Following Rao and Tetreault (2018), we assess model outputs on three criteria: *formality*, *fluency* and *meaning preservation*. Since the goal of our evaluation is to compare models, our evaluation scheme asks workers to compare sentence pairs on these three criteria instead of rating each sentence in isolation. We collect human judgments using CrowdFlower on 300 samples of each model outputs. For FT, we compare the top performing NMT benchmark model in Rao and Tetreault (2018) with our best FT model. For FSMT, we compare outputs from three representative models: NMT-constraint, MultiTask-random and PBMT-random.[3]

Formality. For FT, we want to measure the amount of style variation introduced by a model. Hence, we ask workers to compare the source-style sentence with its target-style model output. For FSMT, we want to measure the amount of style variation between two different translations by the same model. Hence, we ask workers to compare the "informal" English translation and the "formal" English translation of the same source sentence in French.[4] We design a five point scale for comparing the formality of two sentences ranging from one being much more formal than the other to the other being much more formal than the first, giving us a value between 0 and 2 for each sentence pair.[5]

Fluency. For both FT and FSMT tasks, we want to understand how fluent are the different model outputs. Hence, we ask workers to compare the fluency of two model outputs of the same target style. Similar to formality evaluation, we design a five point scale for comparing the fluency of two sentences, giving us a value between 0 and 2 for each sentence pair.

Meaning Preservation. For FT, we want to measure the amount of meaning preserved during formality transfer. Hence, we ask workers to compare the source-style sentence and the target-style model output. For FSMT, we want to measure the amount of meaning preserved between two different translations by the same model. Hence, we ask workers to compare the "informal" English translation and the "formal" English translation of the same source sentence in French. We design a four point scale to compare the meaning of two sentences ranging from the two being completely equivalent to the two being not equivalent, giving us a value between 0 and 3 for each sentence pair.

---

[1] Data and scripts available at `https://github.com/xingniu/multitask-ft-fsmt`.

[2] `https://github.com/EdinburghNLP/nematus/blob/master/data/multi-bleu-detok.perl`

[3] Note that we do not compare with the English reference translation. A more detailed description of the human annotation protocol can be found in the appendix.

[4] Evaluating which systems produces the most (in)formal output is an orthogonal question that we leave to future work.

[5] Details on the conversion from a five point scale to a value between 0 and 2 is in the appendix.

Model	Informal→Formal		Formal→Informal	
	E&M	F&R	E&M	F&R
PBMT (Rao and Tetreault, 2018)	68.22	72.94	33.54	32.64
NMT Baseline (Rao and Tetreault, 2018)	58.80	68.28	30.57	36.71
NMT Combined (Rao and Tetreault, 2018)	68.41	74.22	33.56	35.03
NMT Baseline	65.34	71.28	32.36	36.23
Bi-directional FT	66.30	71.97	34.00	36.33
+ training on E&M + F&R	69.20	73.52	35.44	37.72
+ ensemble decoding ($\times$4)	71.36	74.49	36.18	38.34
+ multi-task learning (MultiTask-tag-style)	**72.13**	**75.37**	**38.04**	**39.09**

Table 1: Automatic evaluation of Formality Transfer with BLEU scores. The bi-directional model with three stacked improvements achieves the best overall performance. The improvement over the second best system is statistically significant at $p < 0.05$ using bootstrap resampling (Koehn, 2004).

6 Formality Transfer Experiments

6.1 Baseline Models from Rao and Tetreault (2018)

PBMT is a phrase-based machine translation model trained on the GYAFC corpus using a training regime consisting of self-training, data sub-selection and a large language model.
NMT Baseline uses OpenNMT-py (Klein et al., 2017). Rao and Tetreault (2018) use a pre-processing step to make source informal sentences more formal and source formal sentences more informal by rules such as re-casing. Word embeddings pre-trained on Yahoo Answers are also used.
NMT Combined is Rao and Tetreault's best performing NMT model trained on the rule-processed GYAFC corpus, with additional forward and backward translations produced by the PBMT model.

6.2 Our Models

NMT Baseline: Our NMT baseline uses Sockeye instead of OpenNMT-py and is trained on raw datasets of two domains and two transfer directions.
Bi-directional FT: Our initial bi-directional model is trained on bi-directional data from both domains with formality tags. It is incrementally augmented with three modifications to get the final multi-task model (i.e. MultiTask-tag-style as described in Section 3.3): (1) We combine training sets of two domains (E&M+F&R) together and train a single model on it. (2) We use ensemble decoding by training four randomly seeded models on the combined data. (3) We add formality-tagged bilingual data and train the model using multi-task learning to jointly learn FT and FSMT. Suppose the amount of original bi-directional FT data is n, we always select kn bilingual data where k is an integer. We also duplicate FT data to make it match the size of selected bilingual data.

6.3 Results

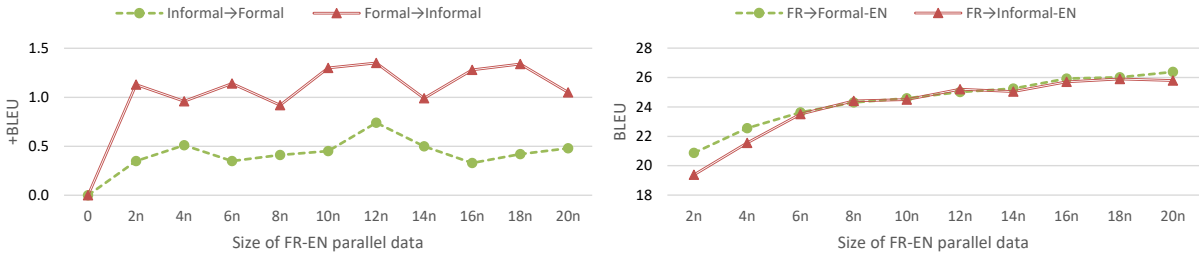
Automatic Evaluation. As shown in Table 1, our NMT baselines yield surprisingly better BLEU scores than those of Rao and Tetreault (2018), even without using rule-processed source training data and pre-trained word embeddings. We attribute the difference to the more optimized NMT toolkit we use.

Initial bi-directional models outperforms uni-directional models. This matches the behavior of bi-directional NMT in low-resource settings studied by Niu et al. (2018) — we work with a relatively small amount of training data ($\sim$50K), and FT models benefit from doubling the size of training data without being confused by mixing two transfer directions. For the same reason, increasing the training data by combining two domains together improves performance further. Ensemble decoding is a consistently effective technique used by NMT and it enhances our NMT-based FT models as expected.

Incorporating the bilingual parallel data by multi-task learning yields further improvement. The target side of bilingual data is selected based on the closeness to the GYAFC corpus, so we hypothesize that the higher quality comes from better target language modeling by training on more English text.

	Model A	Model B	Formality Diff Range = [0,2]		Meaning Preservation Range = [0,3]
			I→F	F→I	
FT	Source	NMT Combined	0.54	0.45	2.94
	Source	MultiTask-tag-style	0.59	**0.64**	2.92
FSMT	NMT-constraint I	NMT-constraint F	**0.35**		2.95
	NMT MultiTask-random I	NMT MultiTask-random F	**0.32**		2.90
	PBMT-random I	PBMT-random F	0.05		2.97

Table 2: Human evaluation of formality difference and meaning preservation. MultiTask-tag-style generates significantly more informal (F→I) English than NMT Combined (p<0.05 using the t-test, see Section 6.3). PBMT-random does not control formality effectively when comparing its informal (I) and formal (F) output (Section 7.2). Formality scores are relatively low because workers rarely choose "much more (in)formal". All models preserve meaning equally well.



(a) BLEU improvements for formality transfer.

(b) BLEU scores for machine translation.

Figure 3: BLEU improvements or scores for four transfer/translation directions vs. the size of FR-EN parallel data. n in x-axis equals to the original size of bi-directional style transfer training data. Formality transfer improves with bilingual data and the performance converges quickly. The translation quality increases monotonously with the size of training data.

Human Evaluation. The superior performance of the best FT model (i.e. MultiTask-tag-style) is also reflected in our human evaluation (see Table 2). It generates slightly more formal English (0.59 vs 0.54) and significantly more informal English (0.64 vs 0.45) than NMT Combined. This is consistent with BLEU differences in Table 1 which show that MultiTask-tag-style yields bigger improvements when transferring formal language to informal. Both models have good quality with respect to meaning preservation (2.94 vs 2.92) and workers can hardly find any fluency difference between outputs of these two models by assigning 0.03 in average in the fluency test (0 means no difference).

Impact of Bilingual Data Size. We evaluate the impact of selected bilingual data size on the combination of development sets from two domains in GYAFC and show the results in Figure 3. The quality of formality transfer improves instantly when using bilingual data and it soon converges when more data is used. Meanwhile, the translation quality increases monotonously with the size of training data. The optimal point is a hyper-parameter that can be determined on the development set. We empirically choose $n = 12$ since it works best for formality transfer and yields reasonable translation quality.

6.4 Qualitative Analysis

We manually inspect 100 randomly selected samples from our evaluation set and compare the target-style output of our best model (MultiTask-tag-style) with that of the best baseline model (NMT-Combined) from Rao and Tetreault (2018). Table 3 shows some samples representative of the trends we find for `informal`→`formal` (3a) and `formal`→`informal` (3b) tasks.

In majority of the cases, the two models produce similar outputs as can be expected since they use similar NMT architectures. In cases where the two outputs differ, in the I→F task, we find that our model produces a more formal output by introducing phrasal level changes (first sample in 3a) or by moving

3a: informal→formal		
1	Original I	chill out sweetie everything will be just fine eventually
	NMT-Combined F	Can you chill out sweetie everything will be just fine eventually.
	MultiTask-tag-style F	Calm down, sweetie, everything will be fine eventually.
2	Original I	Dakota Fanning.....I know that she is only 12 but she is really famous.
	NMT-Combined F	Dakota Fanning.i know that she is only twelve, but she is famous.
	MultiTask-tag-style F	I know that Dakota Fanning is only twelve, but she is really famous.
3	Original I	depends....usully they are about ur personailty but not wat ur gonna do iwith ur life.
	NMT-Combined F	Depends.usully they are about your personailty, but not what your going to do iwith your life.
	MultiTask-tag-style F	It depends. They are about your personality, but not what you are going to do with your life.
4	Original I	THAT DEPENDS...ARE YOU A HOTTIE W/A BODY?
	NMT-Combined F	That depends, are you a hottie with a body?
	MultiTask-tag-style F	That depends. Are you a HOTTIE W / A BODY?
3b: formal→informal		
1	Original F	Therefore I would say that they do succeed but not frequently. I hope this is helpful.
	NMT-Combined I	So I would say that they do failing but not frequently, I hope this is helps.
	MultiTask-tag-style I	so i would say they do it but not all the time, hope this helps.
2	Original F	I am simply inquiring because people behave as though they are no longer interested in them.
	NMT-Combined I	I am just asking because people act as though they are no longer interested in them.
	MultiTask-tag-style I	I'm just asking because people act like they don't like them anymore.
3	Original F	Hello, I am interested in visiting your country.
	NMT-Combined I	Hi, I'm interested in visiting your country.
	MultiTask-tag-style I	hi, I'm going to go to your country.

Table 3: Sample model outputs for the Formality Transfer (FT) task.

phrases around (second sample in 3a), both of which happens frequently during machine translation thus showcasing the benefit of our multi-task approach. Our model very often makes the output sentence more complete (and thereby more formal) by inserting pronouns like 'it', 'they' at the start of the sentence or by removing conjunctions like 'usually', 'and', 'but', 'however' from the beginning of a sentence (sample three in 3a). Likewise, in the F→I task, our model produces more informal sentences compared to the baseline by introducing more phrasal level changes (first and second sample in 3b).

Error analysis: In the I→F task, our model performs worse than the baseline when the original informal sentence consists of all uppercased words (fourth sample in 3a). This is primarily because the baseline model pre-lowercases them using rules. Whereas, we rely on the model to learn this transformation and so it fails to do so for less frequent words. In the F→I task, in trying to produce more informal outputs, our model sometimes fails to preserve the original meaning of the sentence (third sample in 3b). In both tasks, very often our model fails to make transformations for some pairs like ('girls','women'), which the baseline model is very good at. We hypothesize that this could be because for these pairs, human rewriters do not always agree on one of the words in the pair being more informal/formal. This makes our model more conservative in making changes because our bi-directional model combines FT data from both directions and when the original data contains instances where these words are not changed, we double that and learn to copy the word more often than change it.

7 Formality-Sensitive Machine Translation Experiments

7.1 Models

NMT-constraint: We first evaluate the standard NMT model with side constraints introduced in Section 3.2 and then compare it with three variants of FSMT models using multi-task learning as described in Section 3.3 (i.e. **MultiTask-tag-style**, **MultiTask-style** and **MultiTask-random**). The best performing system for FT is MultiTask-tag-style with $12n$ ($\sim$2.5M) bilingual data. For fair comparison, we select this size of bilingual data for all FSMT models either by data selection or randomly.

PBMT-random: We also compare our models with the PBMT-based FSMT system proposed by Niu et al. (2017). Instead of tagging sentences in a binary fashion, this system scores each sentence using a lexical formality model. It requests a desired formality score for translation output and re-ranks n-best

Model	+Tag?	Random?	FR→Formal-EN	FR→Informal-EN
NMT-constraint	✓		27.15	26.70
NMT MultiTask-tag-style	✓		25.02	25.20
NMT MultiTask-style			23.25	23.41
NMT MultiTask-random		✓	25.24	25.14
PBMT-random (Niu et al.)		✓	29.12	29.02

Table 4: BLEU scores of various FSMT models. "+Tag" indicates using formality tags for bilingual data while "Random" indicates using randomly selected bilingual data.

translation hypotheses by their closeness to the desired formality level. We adapt this system to our evaluation scenario — we calculate median scores for informal and formal data (i.e. -0.41 and -0.27 respectively) in GYAFC respectively by a PCA-LSA-based formality model (Niu and Carpuat, 2017; Niu et al., 2017) and use them as desired formality levels.[6] The bilingual training data is randomly selected.

7.2 Results

Automatic Evaluation. We compute BLEU scores on the held out test set for all models as a sanity check on translation quality. Because there is only one reference translation of unknown style for each input sentence, these BLEU scores conflate translation errors and stylistic mismatch, and are therefore not sufficient to evaluate FSMT performance. We include them for completeness here, as indicators of general translation quality, and will rely on human evaluation as primary evaluation method. As can be seen in Table 4, changing the formality level for a given system yields only small differences in BLEU. Based on BLEU scores, we select MultiTask-random as the representative of multi-task FSMT and compare it with NMT-constraint and PBMT-random during our human evaluation.

Human Evaluation. Table 2 shows that neural models control formality significantly better than PBMT-random (0.35/0.32 vs. 0.05). They also introduce more changes in translation: with NMT models, $\sim$80% of outputs change when only the input formality changes, while that is only the case for $\sim$30% of outputs with PBMT-random. Among neural models, MultiTask-random and NMT-constraint have similar quality in controlling output formality (0.32 vs. 0.35) and preserving meaning (2.90 vs. 2.95). They are also equally fluent as judged by humans. Interestingly, multi-task learning helps MultiTask-random perform as well as NMT-constraint with simpler examples that do not require the additional step of data selection to generate formality tags.

7.3 Qualitative Analysis

We randomly sample 100 examples from our test set and manually compare the formal and the informal translations of the French source by MultiTask-random, NMT-constraint and PBMT-random. Table 5 shows representative examples of the observed trends.

We find that in most cases, the difference between the formal and informal style translations is very minor in PBMT-random model, better in NMT-constraint model and the best in our MultiTask-random model (first sample in the table). In general, our MultiTask-random model does a good job of making very large changes while transferring the style, especially into informal (second sample in the table). We hypothesize that this is because our joint model is trained on the GYAFC corpus which consists of parallel sentences that differ heavily in style.

Error analysis: All FSMT models perform well in terms of meaning preservation, yet the human scores are not perfect (Table 2). They occasionally change not only the style but also the meaning of the input (e.g. the third sample of MultiTask-random in Table 5). This motivates future work that penalizes meaning changes more explicitly during training. In general, none of the models do a good job of changing the style when the source sentence is not skewed in one style. For example, consider the French sentence "Combien de fois vous l'ai-je dit?" and its English reference translation "How many

[6]The PCA-LSA-based formality model achieves lowest root-mean-square error on a scoring task of sentential formality as listed on https://github.com/xingniu/computational-stylistic-variations.

1	French Source	Impossible d'avoir accès à internet ici.
	English Reference	I don't know if you've tried yet, but it's impossible to get on the internet up here.
	MultiTask-random Formal	It is impossible to have access to the internet here.
	NMT-constraint Formal	It's impossible to have access to the Internet here.
	PBMT-random Formal	I can't access to the internet here.
	MultiTask-random Informal	Impossible to get to the internet here.
	NMT-constraint Informal	Couldn't have accessed the internet here.
	PBMT-random Informal	I can't access to the internet here.
2	French Source	Abstenez-vous de tout commentaire et répondez à la question, chef Toohey.
	English Reference	Refrain from the commentary and respond to the question, Chief Toohey.
	MultiTask-random Formal	You need to be quiet and answer the question, Chief Toohey.
	NMT-constraint Formal	Please refrain from any comment and answer the question, Chief Toohey.
	PBMT-random Formal	Please refrain from comment and just answer the question, the Tooheys's boss.
	MultiTask-random Informal	Shut up and answer the question, Chief Toohey.
	NMT-constraint Informal	Please refrain from comment and answer the question, chief Toohey.
	PBMT-random Informal	Please refrain from comment and answer my question, Tooheys's boss.
3	French Source	Essaie de présenter des requêtes supplémentaires dès que tu peux.
	English Reference	Try to file any additional motions as soon as you can.
	MultiTask-random Formal	You should try to introduce the sharks as soon as you can.
	NMT-constraint Formal	Try to present additional requests as soon as you can.
	PBMT-random Formal	Try to introduce any additional requests as soon as you can.
	MultiTask-random Informal	Try to introduce sharks as soon as you can.
	NMT-constraint Informal	Try to introduce extra requests as soon as you can.
	PBMT-random Informal	Try to introduce any additional requests as soon as you can.

Table 5: Sample model outputs for the Formality-Sensitive Machine Translation (FSMT) task.

times have I told you, right?". All models produce the same translation "How many times did I tell you?". In such cases, changing style requires heavier editing or paraphrasing of the source sentence that our current models are unable to produce.

8 Conclusion

We explored the use of multi-task learning to jointly perform monolingual FT and bilingual FSMT. Using French-English translation and English style transfer data, we showed that the joint model is able to learn from both style transfer parallel examples and translation parallel examples. On the FT task, the joint model significantly improves the quality of transfer between formal and informal styles in both directions, compared to prior work (Rao and Tetreault, 2018). The joint model interestingly also learns to perform FSMT without being explicitly trained on style-annotated translation examples. On the FSMT task, our model outperforms previously proposed phrase-based MT model (Niu et al., 2017), and performs on par with a neural model with side-constraints which requires more involved data selection.

These results show the promise of multi-task learning for controlling style in language generation applications. In future work, we plan to investigate other multi-task architectures and objective functions that better capture the desired output properties, in order to help address current weaknesses such as meaning errors revealed by manual analysis.

Acknowledgments

We thank the three anonymous reviewers for their helpful comments and suggestions. We thank Joel Tetreault for useful discussions and for making the GYAFC corpus available, as well as members of the Computational Linguistics and Information Processing (CLIP) lab at University of Maryland for helpful discussions. This work is supported by the Clare Boothe Luce Foundation and by the NSF grant IIS-1618193. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsors.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Douglas Biber. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in contrast*, 14(1):7–34.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *ACL*, pages 385–391.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *ICML*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167.

William Coster and David Kauchak. 2011. Learning to simplify sentences using wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9.

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

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

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

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

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *ACL (System Demonstrations)*, pages 67–72.

Catherine Kobus, Josep Maria Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *RANLP*, pages 372–378.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395.

Will Lewis, Christian Federmann, and Ying Xin. 2015. Applying cross-entropy difference for selecting parallel training data from publicly available sources for conversational machine translation. In *IWSLT*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *LREC*.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *HLT-NAACL*, pages 912–921.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *ICLR*.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *ACL*.

Hideki Mima, Osamu Furuse, and Hitoshi Iida. 1997. Improving performance of transfer-driven machine translation with extra-linguistic informatioon from context, situation and environment. In *IJCAI (2)*, pages 983–989.

Robert C. Moore and William D. Lewis. 2010. Intelligent selection of language model training data. In *ACL (Short Papers)*, pages 220–224.

Jonas Mueller, David K. Gifford, and Tommi S. Jaakkola. 2017. Sequence to better sequence: Continuous revision of combinatorial structures. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 2536–2544.

Xing Niu and Marine Carpuat. 2016. The UMD Machine Translation Systems at IWSLT 2016: English-to-French Translation of Speech Transcripts. In *IWSLT*.

Xing Niu and Marine Carpuat. 2017. Discovering stylistic variations in distributional vector space models via lexical paraphrases. In *Proceedings of the Workshop on Stylistic Variation*, pages 20–27.

Xing Niu, Marianna J. Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *EMNLP*, pages 2814–2819.

Xing Niu, Michael Denkowski, and Marine Carpuat. 2018. Bi-directional neural machine translation with synthetic parallel data. In *NMT@ACL*.

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.

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

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *EACL*, pages 157–163.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *HLT-NAACL*, pages 35–40.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *ACL*.

Fadi Abu Sheikha and Diana Inkpen. 2011. Generation of formal and informal sentences. In *ENLG*, pages 187–193.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *NIPS*, pages 6833–6844.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. In *NAACL-HLT*, pages 1503–1515.

Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016. Text simplification using neural machine translation. In *AAAI*, pages 4270–4271.

Shuly Wintner, Shachar Mirkin, Lucia Specia, Ella Rabinovich, and Raj Nath Patel. 2017. Personalized machine translation: Preserving original author traits. In *EACL*, pages 1074–1084.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *ACL*, pages 1015–1024.

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

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *TACL*, 4:401–415.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in japanese-to-english neural machine translation. In *WAT@COLING*, pages 203–210.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *EMNLP*, pages 584–594.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *COLING*, pages 1353–1361.

Appendix A. Details of Human-Based Evaluations

As described in the main paper, we assess the model outputs on three criteria of formality, fluency and meaning preservation. We collect these judgments using CrowdFlower.

Since we want native English speakers to perform this task, we restrict our set of annotators only to these three native English speaking countries: United States, United Kingdom and Australia. We create a sample of 51 gold questions for each of the three tasks (criteria). Annotators have to continually maintain an accuracy of above 70% to be able to contribute to the task.

We collect judgments on 300 samples of each model output and we collect three judgments per sample (i.e. sentence pair). Given the three judgments per sample, we calculate the aggregate score using the weighted average:

$$\frac{\sum_{i=1}^{3} score_i \times trust_i}{\sum_{i=1}^{3} trust_i}$$

where $score_i$ is the score given by an annotator and $trust_i$ is our trust on that annotator. This trust is the accuracy of the annotator on the gold questions.

Formality: Given two sentences, we ask workers to compare their formality using one of the following categories, regardless of fluency and meaning. We do not enumerate specific rules (e.g. typos or contractions) and encourage workers to use their own judgment.

Score	Category
2	*Sentence 1 is much more formal than Sentence 2*
1	*Sentence 1 is more formal than Sentence 2*
0	*No difference or hard to say*
-1	*Sentence 2 is more formal than Sentence 1*
-2	*Sentence 2 is much more formal than Sentence 1*

Fluency: Given two sentences, we ask workers to compare their fluency using one of the following categories, regardless of style and meaning. We define fluency as follows: *A sentence is fluent if it has a meaning and is coherent and grammatical well-formed.*

Score	Category
2	*Sentence 1 is much more fluent than Sentence 2*
1	*Sentence 1 is more fluent than Sentence 2*
0	*No difference or hard to say*
-1	*Sentence 2 is more fluent than Sentence 1*
-2	*Sentence 2 is much more fluent than Sentence 1*

Meaning preservation: Given two sentences, we ask workers to answer "how much of the first sentence's meaning is preserved in the second sentence", regardless of style.

Score	Category
3	*Equivalent since they convey the same key idea*
2	*Mostly equivalent since they convey the same key idea but differ in some unimportant details*
1	*Roughly equivalent since they share some ideas but differ in important details*
0	*Not equivalent since they convey different ideas*

While collecting formality and fluency annotations for sentence pairs, to avoid system-level bias, we randomly swap the two items in the pair and collect annotations on a symmetric range of [-2,2]. But while aggregating these scores, we recover the order and hence the final scores in Table 2 are only in the range of [0,2].