Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer

Sudha Rao
University of Maryland, College Park*
raosudha@cs.umd.edu

Joel Tetreault
Grammarly
joel.tetreault@grammarly.com

Abstract

Style transfer is the task of automatically transforming a piece of text in one particular style into another. A major barrier to progress in this field has been a lack of training and evaluation datasets, as well as benchmarks and automatic metrics. In this work, we create the largest corpus for a particular stylistic transfer (formality) and show that techniques from the machine translation community can serve as strong baselines for future work. We also discuss challenges of using automatic metrics.

1 Introduction

One key aspect of *effective communication* is the accurate expression of the style or tone of some content. For example, writing a more *persuasive email* in a marketing position could lead to increased sales; writing a more *formal email* when applying for a job could lead to an offer; and writing a more *polite note* to your future spouse's parents, may put you in a good light. Hovy (1987) argues that by varying the style of a text, people convey more information than is present in the literal meaning of the words. One particularly important dimension of style is formality (Heylighen and Dewaele, 1999). Automatically changing the style of a given content to make it more formal can be a useful addition to any writing assistance tool.

In the field of style transfer, to date, the only available dataset has been for the transformation of modern English to Shakespeare, and it led to the application of phrase-based machine translation (PBMT) (Xu et al., 2012) and neural machine translation (NMT) (Jhamtani et al., 2017) models to the task. The lack of an equivalent or larger dataset for any other form of style transfer has blocked progress in this field. Moreover, prior work has mainly borrowed metrics from machine translation (MT) and paraphrase communities for evaluating style transfer. However, it is not clear if those metrics are the best ones to use for this task. In this work, we address these issues through the following three contributions:

- **Corpus:** We present Grammarly's Yahoo Answers Formality Corpus (GYAFC), the largest dataset for any style containing a total of 110K informal / formal sentence pairs. Table 1 shows sample sentence pairs.
- **Benchmarks:** We introduce a set of learning models for the task of formality style transfer. Inspired by work in low resource MT, we adapt existing PBMT and NMT approaches for our task and show that they can serve as strong benchmarks for future work.
- **Metrics:** In addition to MT and paraphrase metrics, we evaluate our models along three axes: *formality*, *fluency* and *meaning preservation* using existing automatic metrics. We compare these metrics with their human judgments and show there is much room for further improvement.

Informal:	*I'd say it is punk though.*
Formal:	*However, I do believe it to be punk.*
Informal:	*Gotta see both sides of the story.*
Formal:	*You have to consider both sides of the story.*

Table 1: Informal sentences with formal rewrites.

In this paper, we primarily focus on the *informal* to *formal* direction since we collect our dataset for this direction. However, we evaluate our models on the *formal* to *informal* direction as well.[1] All data, model outputs, and evaluation results have been made public[2] in the hope that they will encourage more research into style transfer.

*This research was performed when the first author was at Grammarly.

[1] Results are in the supplementary material.

[2] https://github.com/raosudha89/GYAFC-corpus

Proceedings of NAACL-HLT 2018, pages 129–140
New Orleans, Louisiana, June 1 - 6, 2018.

In the following two sections we discuss related work and the GYAFC dataset. In §4, we detail our rule-based and MT-based approaches. In §5, we describe our human and automatic metric based evaluation. In §6, we describe the results of our models using both human and automatic evaluation and discuss how well the automatic metrics correlate with human judgments.

2 Related Work

Style Transfer with Parallel Data: Sheikha and Inkpen (2011) collect pairs of formal and informal words and phrases from different sources and use a natural language generation system to generate informal and formal texts by replacing lexical items based on user preferences. Xu et al. (2012) (henceforth XU12) was one of the first works to treat style transfer as a sequence to sequence task. They generate a parallel corpus of 30K sentence pairs by scraping the modern translations of Shakespeare plays and train a PBMT system to translate from modern English to Shakespearean English.[3] More recently, Jhamtani et al. (2017) show that a copy-mechanism enriched sequence-to-sequence neural model outperforms XU12 on the same set. In text simplification, the availability of parallel data extracted from English Wikipedia and Simple Wikipedia (Zhu et al., 2010) led to the application of PBMT (Wubben et al., 2012a) and more recently NMT (Wang et al., 2016) models. We take inspiration from both the PBMT and NMT models and apply several modifications to these approaches for our task of transforming the formality style of the text.

Style Transfer without Parallel Data: Another direction of research directly controls certain attributes of the generated text *without* using parallel data. Hu et al. (2017) control the sentiment and the tense of the generated text by learning a disentangled latent representation in a neural generative model. Ficler and Goldberg (2017) control several linguistic style aspects simultaneously by conditioning a recurrent neural network language model on specific style (professional, personal, length) and content (theme, sentiment) parameters. Under NMT models, Sennrich et al. (2016a) control the politeness of the translated text via side constraints, Niu et al. (2017) control the level of formality of MT output by selecting phrases of a requisite formality level from the k-best list during decoding. In the field of text simplification, more recently, Xu et al. (2016) learn large-scale paraphrase rules using bilingual texts whereas Kajiwara and Komachi (2016) build a monolingual parallel corpus using sentence similarity based on alignment between word embeddings. Our work differs from these methods in that we mainly address the question of how much leverage we can derive by collecting a large amount of informal-formal sentence pairs and build models that learn to transfer style directly using this parallel corpus.

Identifying Formality: There has been previous work on detecting formality of a given text at the lexical level (Brooke et al., 2010; Lahiri et al., 2011; Brooke and Hirst, 2014; Pavlick and Nenkova, 2015), at the sentence level (Pavlick and Tetreault, 2016) and at the document level (Sheikha and Inkpen, 2010; Peterson et al., 2011; Mosquera and Moreda, 2012). In our work, we reproduce the sentence-level formality classifier introduced in Pavlick and Tetreault (2016) (PT16) to extract informal sentences for GYAFC creation and to automatically evaluate system outputs.

Evaluating Style Transfer: The problem of style transfer falls under the category of natural language generation tasks such as machine translation, paraphrasing, etc. Previous work on style transfer (Xu et al., 2012; Jhamtani et al., 2017; Niu et al., 2017; Sennrich et al., 2016a) has re-purposed the MT metric BLEU (Papineni et al., 2002) and the paraphrase metric PINC (Chen and Dolan, 2011) for evaluation. Additionally, XU12 introduce three new automatic style metrics based on cosine similarity, language model and logistic regression that measure the degree to which the output matches the target style. Under human based evaluation, on the other hand, there has been work on a more fine grained evaluation where human judgments were separately collected for adequacy, fluency and style (Xu et al., 2012; Niu et al., 2017). In our work, we conduct a more thorough evaluation where we evaluate model outputs on the three criteria of *formality*, *fluency* and *meaning* using both automatic metrics and human judgments.

[3] https://github.com/cocoxu/Shakespeare

Domain	Total	Informal	Formal
All Yahoo Answers	40M	24M	16M
Entertainment & Music	3.8M	2.7M	700K
Family & Relationships	7.8M	5.6M	1.8M

Table 2: Yahoo Answers corpus statistics

		Informal to Formal		*Formal to Informal*	
	Train	Tune	Test	Tune	Test
E&M	52,595	2,877	1,416	2,356	1,082
F&R	51,967	2,788	1,332	2,247	1,019

Table 3: GYAFC dataset statistics

3 GYAFC **Dataset**

3.1 Creation Process

Yahoo Answers,[4] a question answering forum, contains a large number of informal sentences and allows redistribution of data. Hence, we use the Yahoo Answers L6 corpus[5] to create our GYAFC dataset of informal and formal sentence pairs. In order to ensure a uniform distribution of data, we remove sentences that are questions, contain URLs, and are shorter than 5 words or longer than 25. After these preprocessing steps, 40 million sentences remain. The Yahoo Answers corpus consists of several different domains like *Business, Entertainment & Music, Travel, Food,* etc. PT16 show that the formality level varies significantly across different genres. In order to control for this variation, we work with two specific domains that contain the most informal sentences and show results on training and testing within those categories. We use the formality classifier from PT16 to identify informal sentences. We train this classifier on the *Answers* genre of the PT16 corpus which consists of nearly 5,000 randomly selected sentences from Yahoo Answers manually annotated on a scale of -3 (very informal) to 3 (very formal).[6] We find that the domains of *Entertainment & Music* and *Family & Relationships* contain the most informal sentences and create our GYAFC dataset using these domains. Table 2 shows the number of formal and informal sentences in all of Yahoo Answers corpus and within the two selected domains. Sentences with a score less than 0 are considered as informal and sentences with a score greater than 0 are considered as formal.

Next, we randomly sample a subset of 53,000 informal sentences each from the *Entertainment & Music* (E&M) and *Family & Relationships* (F&R) categories and collect one formal rewrite per sentence using Amazon Mechanical Turk. The workers are presented with detailed instructions, as well as examples. To ensure quality control, four experts, two of which are the authors of this paper, reviewed the rewrites of the workers and rejected those that they felt did not meet the required standards. They also provided the workers with reasons for rejection so that they would not repeat the same mistakes. Any worker who repeatedly performed poorly was eventually blocked from doing the task. We use this train set to train our models for the style transfer tasks in both directions.

Since we want our tune and test sets to be of higher quality compared to the train set, we recruit a set of 85 expert workers for this annotation who had a 100% acceptance rate for our task and who had previously done more than 100 rewrites. Further, we collect multiple references for the tune/test set to adapt PBMT tuning and evaluation techniques to our task. We collect four different rewrites per sentence using our expert workers by randomly assigning sentences to the experts until four rewrites for each sentence are obtained.[7] To create our tune and test sets for the *informal* to *formal* direction, we sample an additional 3,000 informal sentences for our tune set and 1,500 sentences for our test set from each of the two domains.

To create our tune and test sets for the *formal* to *informal* direction, we start with the same tune and test split as the first direction. For each formal rewrite[8] from the first direction, we collect three different informal rewrites using our expert workers as before. These three informal rewrites along with the original informal sentence become our set of four references for this direction of the task. Table 3 shows the exact number of sentences in our train, tune and test sets.

3.2 Analysis

The following quantitative and qualitative analyses are aimed at characterizing the changes between the original informal sentence and its formal

[4] https://answers.yahoo.com/answer

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

[6] http://www.seas.upenn.edu/~nlp/resources/formality-corpus.tgz

[7] Thus, note that the four rewrites are not from the same four workers for each sentence

[8] Out of four, we pick the one with the most edit distance with the original informal. Rationale explained in Section 3.2

rewrite in the GYAFC train split.[9] We present our analysis here on only the E&M domain data since we observe similar patterns in F&R.

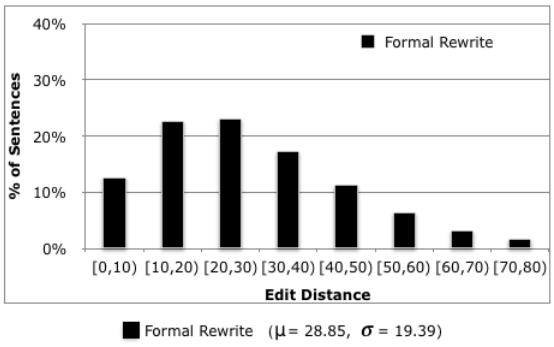

Figure 1: Percentage of sentences binned according to formality score in train set of E&M.

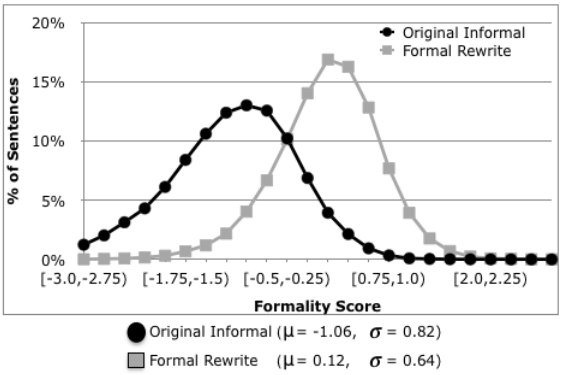

Figure 2: Percentage of sentences binned according to formality score in train set of E&M

Quantitative Analysis: While rewriting sentences more formally, humans tend to make a wide range of lexical/character-level edits. In Figure 1, we plot the distribution of the character-level Levenshtein edit distance between the original informal and the formal rewrites in the train set and observe a standard deviation of $\sigma = 19.39$ with a mean $\mu = 28.85$. Next, we look at the difference in the formality level of the original informal and the formal rewrites in GYAFC. We find that the classifier trained on the *Answers* genre of PT16 dataset correlates poorly (Spearman $\rho = 0.38$) with human judgments when tested on our domain specific datasets. Hence, we collect formality judgments on a scale of -3 to +1, similar to PT16, for an additional 5000 sentences each from both domains and obtain a formality classifier with higher correlation (Spearman $\rho = 0.56$). We use this retrained classifier for our evaluation in §5 as well.

In Figure 2, we plot the distribution of the formality scores on the original informal sentence and their formal rewrites in the train set and observe an increase in the mean formality score as we go from informal (-1.06) to formal rewrites (0.12). As compared to edit distance and formality, we observe a much lower variation in sentence lengths with the mean slightly increasing from informal (11.93) to their formal rewrites (12.56) in the train set.

Qualitative Analysis: To understand what stylistic choices differentiate formal from informal text, we perform an analysis similar to PT16 and look at 50 rewrites from both domains and record the frequency of the types of edits that workers made when creating a more formal sentence.[10] In contrast to PT16, we observe a higher percentage of phrasal paraphrases (47%), edits to punctuations (40%) and expansion of contractions (12%). This is reflective of our sentences coming from very informal domains of Yahoo Answers. Similar to PT16, we also observe capitalization (46%) and normalization (10%).

4 Models

We experiment with three main classes of approaches: a rule-based approach, PBMT and NMT. Inspired by work in low resource machine translation, we apply several modifications to the standard PBMT and NMT models and create a set of strong benchmarks for the style transfer community. We apply these models to both directions of style transfer: *informal* to *formal* and *formal* to *informal*. In our description, we refer to the two styles as *source* and *target*. We summarize the models below and direct the reader to supplementary material for further detail.

4.1 Rule-based Approach

Corresponding to the category of edits described in §3.2, we develop a set of rules to automatically make an informal sentence more formal where we capitalize first word and proper nouns, remove repeated punctuations, handcraft a list of expansion for contractions etc. For the *formal* to *informal* direction, we design a similar set of rules in the opposite direction.

[9] We observe similar patterns on the tune and test set.

[10] Examples of edits in supplementary material.

4.2 Phrase-based Machine Translation

Phrased-based machine translation models have had success in the fields of machine translation, style transfer (XU12) and text simplification (Wubben et al., 2012b; Xu et al., 2016). Inspired by work in low resource machine translation, we use a combination of training regimes to develop our model. We train on the output of the rule-based approach when applied to GYAFC. This is meant to force the PBMT model to learn generalizations *outside* the rules. To increase the data size, we use self-training (Ueffing, 2006), where we use the PBMT model to translate the large number of in-domain sentences from GYAFC belonging to the the source style and use the resultant output to retrain the PBMT model. Using subselection, we only select rewrites that have an Levenshtein edit distance of over 10 characters when compared to the source to encourage the model to be less conservative. Finally, we upweight the rule-based GYAFC data via duplication (Sennrich et al., 2016b). For our experiments, we use Moses (Koehn et al., 2007). We train a 5-gram language model using KenLM (Heafield et al., 2013), and use target style sentences from GYAFC and the sub-sampled target style sentences from out-of-domain Yahoo Answers, as in Moore and Lewis (2010), to create a large language model.

4.3 Neural Machine Translation

While encoder-decoder based neural network models have become quite successful for MT(Sutskever et al., 2014; Bahdanau et al., 2014; Cho et al., 2014), the field of style transfer, has not yet been able to fully take advantage of these advances owing to the lack of availability of large parallel data. With GYAFC we can now show how well NMT techniques fare for style transfer. We experiment with three NMT models:

NMT baseline: Our baseline model is a bi-directional LSTM (Hochreiter and Schmidhuber, 1997) encoder-decoder model with attention (Bahdanau et al., 2014).[11] We pretrain the input word embeddings on Yahoo Answers using GloVE (Pennington et al., 2014). As in our PBMT based approach, we train our NMT baseline model on the output of the rule-based approach when applied to GYAFC.

NMT Copy: Jhamtani et al., (2017) introduce a copy-enriched NMT model for style transfer to better handle stretches of text which should not be changed. We incorporate this mechanism into our NMT Baseline.

NMT Combined: The size of our parallel data is smaller than the size typically used to train NMT models. Motivated by this fact, we propose several variants to the baseline models that we find helps minimize this issue. We augment the data used to train NMT Copy via two techniques: 1) we run the PBMT model on additional source data, and 2) we use back-translation (Sennrich et al., 2016c) of the PBMT model to translate the large number of in-domain target style sentences from GYAFC. To balance the over one million artificially generated pairs from the respective techniques, we upweight the rule-based GYAFC data via duplication.[12]

5 Evaluation

As discussed earlier, there has been very little research into best practices for style transfer evaluation. Only a few works have included a human evaluation (Xu et al., 2012; Jhamtani et al., 2017), and automatic evaluations have employed BLEU or PINC (Xu et al., 2012; Chen and Dolan, 2011), which have been borrowed from other fields and not vetted for this task. In our work, we conduct a more thorough and detailed evaluation using both humans and automatic metrics to assess transformations. Inspired by work in the paraphrase community (Callison-Burch, 2008), we solicit ratings on how formal, how fluent and how meaning-preserving a rewrite is. Additionally, we look at the correlation between the human judgments and the automatic metrics.

5.1 Human-based Evaluation

We perform human-based evaluation to assess model outputs on the four criteria: *formality*, *fluency*, *meaning* and *overall*. For a subset of 500 sentences from the test sets of both *Entertainment & Music* and *Family & Relationship* domains, we collect five human judgments per sentence per criteria using Amazon Mechanical Turk as follows:

[11] Details are in the supplementary material.

[12] Training data sizes for different methods are summarized in the supplementary material.

Formality: Following PT16, workers rate the formality of the source style sentence, the target style reference rewrite and the target style model outputs on a discrete scale of -3 to +3 described as: *-3: Very Informal, -2: Informal, -1: Somewhat Informal, 0: Neutral, 1: Somewhat Formal, 2: Formal and 3: Very Formal.*

Fluency: Following Heilman et al. (2014), workers rate the fluency of the source style sentence, the target style reference rewrite and the target style model outputs on a discrete scale of 1 to 5 described as: *5: Perfect, 4: Comprehensible, 3: Somewhat Comprehensible, 2: Incomprehensible*. We additionally provide an option of *1: Other* for sentences that are incomplete or just a fragment.

Meaning Preservation: Following the annotation scheme developed for the Semantic Textual Similarity (STS) dataset (Agirre et al., 2016), given two sentences i.e. the source style sentence and the target style reference rewrite or the target style model output, workers rate the meaning similarity of the two sentences on a scale of 1 to 6 described as: *6: Completely equivalent, 5: Mostly equivalent, 4: Roughly equivalent, 3: Not equivalent but share some details, 2: Not equivalent but on same topic, 1: Completely dissimilar*.

Overall Ranking: In addition to the fine-grained human judgments, we collect judgments to assess the overall ranking of the systems. Given the original source style sentence, the target style reference rewrite and the target style model outputs, we ask workers to rank the rewrites in the order of their overall formality, taking into account both fluency and meaning preservation. We then rank the model using the equation below:

$$rank(model) = \frac{1}{|S|}\sum_{s \in S}\frac{1}{|J|}\sum_{j \in J} rank(s_{model}, j) \quad (1)$$

where, $model$ is the one of our models, S is a subset of 500 test set sentences, J is the set of five judgments, s_{model} is the model rewrite for sentence s, and $rank(s_{model}, j)$ is the rank of s_{model} in judgment j.

The two authors of the paper reviewed these human judgments and found that in majority of the cases the annotations looked correct. But as is common in any such crowdsourced data collection process, there were some errors, especially in the overall ranking of the systems.

5.2 Automatic Metrics

We cover each of the human evaluations with a corresponding automatic metric:

Formality: We use the formality classifier described in PT16. We find that the classifier trained on the *answers* genre of PT16 dataset does not perform well when tested on our datasets. Hence, we collect formality judgments for an additional 5000 sentences and use the formality classifier re-trained on this in-domain data.

Fluency: We use the reimplementation[13] of Heilman et al. (2014) (H14 in Table 4) which is a statistical model for predicting the grammaticality of a sentence on a scale of 0 to 4 previously shown to be effective for other generation tasks like grammatical error correction (Napoles et al., 2016).

Meaning Preservation: Modeling semantic similarity at a sentence level is a fundamental language processing task, and one that is a wide open field of research. Recently, He et al., (2015) (HE15 in Table 4) developed a convolutional neural network based sentence similarity measure. We use their off-the-shelf implementation[14] to train a model on the STS and use it to measure the meaning similarity between the original source style sentence and its target style rewrite (both reference and model outputs).

Overall Ranking: We experiment with BLEU (Papineni et al., 2002) and PINC (Chen and Dolan, 2011) as both were used in prior style evaluations, as well as TERp (Snover et al., 2009).

6 Results

In this section, we discuss how well the five models perform in the *informal* to *formal* style transfer task using human judgments (§6.1) and automatic metrics (§6.2), the correlation of the automatic metrics and human judgments to determine the ef-

[13] https://github.com/cnap/grammaticality-metrics/tree/master/heilman-et-al

[14] https://github.com/castorini/MP-CNN-Torch

Model	Formality		Fluency		Meaning		Combined		Overall		
	Human	PT16	Human	H14	Human	HE15	Human	Auto	BLEU	TERp	PINC
Original Informal	*-1.23*	*-1.00*	*3.90*	*2.89*	–	–	–	–	*50.69*	*0.35*	*0.00*
Formal Reference	0.38	0.17	4.45	3.32	4.57	3.64	5.68	4.67	100.0	0.37	69.79
Rule-based	-0.59	-0.34	4.00	3.09	**4.85**	**4.41**	5.24	4.69	61.38	0.27	26.05
PBMT	-0.19*	0.00*	3.96	3.28*	4.64*	4.19*	5.27	4.82*	67.26*	**0.26**	44.94*
NMT Baseline	**0.05***	0.07*	4.05	**3.52***	3.55*	3.89*	4.96*	**4.84***	56.61	0.38*	**56.92***
NMT Copy	0.02*	**0.10***	4.07	3.45*	3.48*	3.87*	4.93*	4.81*	58.01	0.38*	56.39*
NMT Combined	-0.16*	0.00*	**4.09***	3.27*	4.46*	4.20*	**5.32***	4.82*	**67.67***	**0.26**	43.54*

Table 4: Results of models on 500 test sentences from E&M for *informal* to *formal* task evaluated using human judgments and automatic metrics for three criteria of evaluation: formality, fluency and meaning preservation. Scores marked with * are significantly different from the rule-based scores with $p < 0.001$.

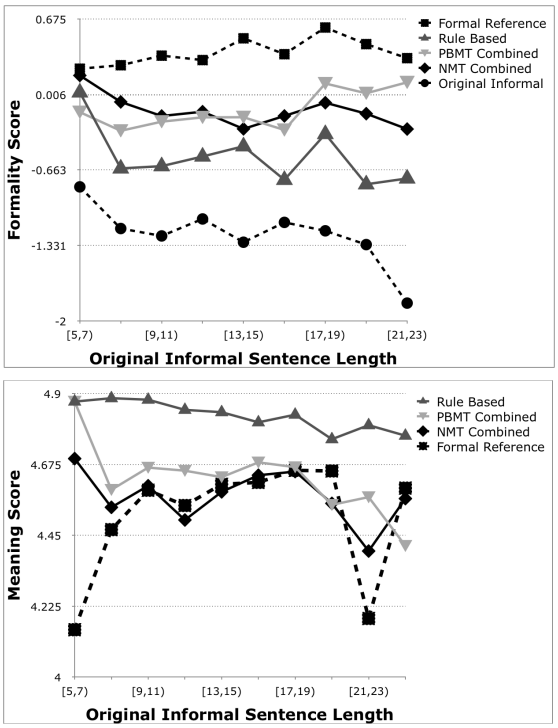

Figure 3: For varying sentence lengths of the original informal sentence the *formality* and the *meaning* scores from human judgments on different model outputs and on the original informal and the formal reference sentences.

ficacy of the metrics (§6.3) and present a manual analysis (§6.4). We randomly select 500 sentences from each test set and run all five models. We use the entire train and tune split for training and tuning. We discuss results only on the E&M domain and list results on the F&R domain in the supplementary material.

Table 4 shows the results for human §6.1 and automatic §6.2 evaluation of model rewrites. For all metrics except *TERp*, a higher score is better. For each of the automatic metrics, we evaluate against four human references. The row *'Original Informal'* contains the scores when the original informal sentence is compared with the four formal reference rewrites. Comparing the model scores to this score helps us understand how closer are the model outputs to the formal reference rewrites compared to initial distance between the informal and the formal reference rewrite.

6.1 Results using Human Judgments

The columns marked *'Human'* in Table 4 show the human judgments for the models on the three separate criteria of *formality*, *fluency* and *meaning* collected using the process described in Section 5.1.[15] The NMT Baseline and Copy models beat others on the formality axis by a significant margin. Only the NMT Combined model achieves a statistically higher fluency score when compared to the rule-based baseline model. As expected, the rule-based model is the most meaning preserving since it is the most conservative. Figure 3 shows the trend in the four leading models along *formality* and *meaning* for varying lengths of the source sentence. NMT Combined beats PBMT on *formality* for shorter lengths whereas the trend reverses as the length increases. PBMT generally preserves meaning more than the NMT Combined. We find that the fluency scores for all models decreases as the sentence length increases which is similar to the trend generally observed with machine translation based approaches.

Since a good style transfer model is the one that attains a balanced score across all the three axes, we evaluate the models on a combination of these metrics[16] shown under the column *'Combined'* in Table 4. NMT Combined is the only model having a combined score statistically greater than the rule-based approach.

[15] Out of the four reference rewrites, we pick one at random to show to Turkers.

[16] We recalibrate the scores to normalize for different ranges.

Finally, Table 5 shows the overall rankings of the models from best to worst in both domains. PBMT and NMT Combined models beat the rule-based model although not significantly in the E&M domain but significantly in the F&R domain. Interestingly, the rule-based approach attains third place with a score significantly higher than NMT Copy and NMT Baseline models. It is important to note here that while such a rule-based approach is relatively easy to craft for the formality style transfer task, the same may not be true for other styles like politeness or persuasiveness.

E&M	F&R
(2.03*) Reference	(2.13*) Reference
(2.47) PBMT	(2.38*) PBMT
(2.48) NMT Combined	(2.38*) NMT Combined
(2.54) Rule-based	(2.56) Rule-based
(3.03*) NMT Copy	(2.72*) NMT Copy
(3.03*) NMT Baseline	(2.79*) NMT Baseline

Table 5: Ranking of different models on the *informal* to *formal* style transfer task. Rankings marked with * are significantly different from the rule-based ranking with $p < 0.001$.

Automatic	**Human**	**E&M**	**F&R**
Formality	*Formality*	0.47	0.45
Fluency	*Fluency*	0.48	0.46
Meaning	*Meaning*	0.33	0.30
BLEU	*Overall*	-0.48	-0.43
TERp	*Overall*	0.31	0.30
PINC	*Overall*	0.11	0.08

Table 6: Spearman rank correlation between automatic metrics and human judgments. The first three metrics are correlated with their respective human judgments and the last three metrics are correlated with the *overall ranking* human judgments. All correlations are statistically significant with $p < 0.001$.

6.2 Results with Automatic Metrics

Under automatic metrics, the formality and meaning scores align with the human judgments with the NMT Baseline and NMT Copy winning on formality and rule-based winning on meaning. The fluency score of the NMT Baseline is the highest in contrast to human judgments where the NMT Combined wins. This discrepancy could be due to H14 being trained on *essays* which contains sentences of a more formal genre compared to Yahoo Answers. In fact, the fluency classifier scores the formal reference quite low as well. Under overall metrics, PBMT and NMT Combined models beat other models as per BLEU (significantly) and TERp (not significantly). NMT Baseline and NMT copy win over other models as per PINC which can be explained by the fact that PINC measures lexical dissimilarity with the source and NMT models tend towards making more changes. Although such an analysis is useful, for a more thorough understanding of these metrics, we next look at their correlation with human judgments.

6.3 Metric Correlation

We report the spearman rank correlation coefficient between automatic metrics and human judgments in Table 6. For *formality*, *fluency* and *meaning*, the correlation is with their respective human judgments whereas for BLEU, TERp and PINC, the correlation is with the overall ranking.

We see that the formality and the fluency metrics correlate moderately well while the meaning metric correlates comparatively poorly. To be fair, the HE15 classifier was trained on the STS dataset which contains more formal writing than informal. BLEU correlates moderately well (better than what XU12 observed for the Shakespeare task) whereas the correlation drops for TERp. PINC, on the other hand, correlates very poorly with a positive correlation with rank when it should have a negative correlation with rank, just like BLEU. This sheds light on the fact that PINC, on its own, is not a good metric for style transfer since it prefers lexical edits at the cost of meaning changes. In the Shakespeare task, XU12 did observe a higher correlation with PINC (0.41) although the correlation was not with overall system ranking but rather only on the style metric. Moreover, in the Shakespeare task, changing the text is more favorable than in formality.

6.4 Manual Analysis

The prior evaluations reveal the relative performance differences between approaches. Here, we identify trends per and between approaches. We sample 50 informal sentences total from both domains and then analyze the outputs from each model. We present sample sentences in Table 7.

The NMT Baseline and NMT Copy tend to have the most variance in their performance. This is likely due to the fact that they are trained on only 50K sentence pairs, whereas the other models are trained on much more data. For shorter sentences, these models make some nice formal transformations like from '*very dumb*' to '*very foolish*'. However, for longer sentences, these models make drastic meaning changes and drop some content altogether (see examples in Table 7). On the

Entertainment & Music	
Original Informal	Wow , I am very dumb in my observation skills .....
Reference Formal	I do not have good observation skills .
Rule-based	Wow , I am very dumb in my observation skills .
PBMT	Wow , I am very dumb in my observation skills .
NMT Baseline	I am very foolish in my observation skills .
NMT Copy	Wow , I am very foolish in my observation skills .
NMT Combined	I am very unintelligent in my observation skills .
Family & Relationship	
Original Informal	i hardly everrr see him in school either usually i see hima t my brothers basketball games .
Reference Formal	I hardly ever see him in school . I usually see him with my brothers playing basketball .
Rule-based	I hardly everrr see him in school either usually I see hima t my brothers basketball games .
PBMT	I hardly see him in school as well, but my brothers basketball games .
NMT	I rarely see him in school , either I see him at my brother 's basketball games .
NMT Copy	I hardly see him in school either , usually I see him at my brother 's basketball games .
NMT Combined	I rarely see him in school either usually I see him at my brothers basketball games .

Table 7: Sample model outputs with references from both E&M and F&R domains on the *informal* to *formal* task

other hand, the PBMT and NMT Combined models have lower variance in their performance. They make changes more conservatively but when they do, they are usually correct. Thus, most of the outputs from these two models are usually meaning preserving but at the expense of a lower formality score improvement.

In most examples, all models are good at removing very informal words like '*stupid*', '*idiot*' and '*hell*', with PBMT and NMT Combined models doing slightly better. All models struggle when the original sentence is very informal or disfluent. They all also struggle with sentence completions that humans seem to be very good at. This might be because humans assume a context when absent, whereas the models do not. Unknown tokens, either real words or misspelled words, tend to wreak havoc on all approaches. In most cases, the models simply did not transform that section of the sentence, or remove the unknown tokens. Most models are effective at low-level changes such as writing out numbers, inserting commas, and removing common informal phrases.

7 Conclusions and Future Work

The goal of this paper was to move the field of style transfer forward by creating a large training and evaluation corpus to be made public, showing that adapting MT techniques to this task can serve as strong baselines for future work, and analyzing the usefulness of existing metrics for overall style transfer as well as three specific criteria of automatic style transfer evaluation. We view this work as rigorously expanding on the foundation set by XU12 five years earlier. It is our hope that with a common test set, the field can finally benchmark approaches which do not require parallel data.

We found that while the NMT systems perform well given automatic metrics, humans had a slight preference for the PBMT approach. That being said, two of the neural approaches (NMT Baseline and Copy) often made successful changes and larger rewrites that the other models could not. However, this often came at the expense of a meaning change.

We also introduced new metrics and vetted all metrics using comparison with human judgments. We found that previously-used metrics did not correlate well with human judgments, and thus should be avoided in system development or final evaluation. The formality and fluency metrics correlated best and we believe that some combination of these metrics with others would be the best next step in the development of style transfer metrics. Such a metric could then in turn be used to optimize MT models. Finally, in this work we focused on one particular style, formality. The long term goal is to generalize the methods and metrics to any style.

Acknowledgments

The authors would like to thank Yahoo Research for making their data available. The authors would also like to thank Junchao Zheng and Claudia Leacock for their help in the data creation process, Courtney Napoles for providing the fluency scores, Marcin Junczys-Dowmunt, Rico Sennrich, Ellie Pavlick, Maksym Bezva, Dimitrios Alikaniotis and Kyunghyun Cho for helpful discussion and the three anonymous reviewers for their useful comments and suggestions.

References

Eneko Agirre, Carmen Banea, Daniel M Cer, Mona T Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval@ NAACL-HLT*. pages 497–511.

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

Julian Brooke and Graeme Hirst. 2014. Supervised ranking of co-occurrence profiles for acquisition of continuous lexical attributes. In *COLING*. pages 2172–2183.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, pages 90–98.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, pages 196–205. http://www.aclweb.org/anthology/D08-1021.

David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 190–200.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation* page 103.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. *Proceedings of the Workshop on Stylistic Variation, EMNLP 2017* .

Hua He, Kevin Gimpel, and Jimmy J Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *EMNLP*. pages 1576–1586.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pages 690–696. https://kheafield.com/papers/edinburgh/estimate_paper.pdf.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 174–180. http://www.aclweb.org/anthology/P14-2029.

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

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

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

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

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

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pages 1147–1158.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, pages 177–180.

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu. 2011. Informality judgment at sentence level and experiments with formality score. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 446–457.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*. Association for Computational Linguistics, pages 220–224.

Alejandro Mosquera and Paloma Moreda. 2012. Smile: An informality classification tool for helping to assess quality and credibility in web 2.0 texts. In *Proceedings of the ICWSM workshop: Real-Time Analysis and Mining of Social Streams (RAMSS)*.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 2109–2115. https://aclweb.org/anthology/D16-1228.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 2804–2809.

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

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *HLT-NAACL*. pages 218–224.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics* 4:61–74.

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

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the enron corpus. In *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, pages 86–95.

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. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 371–376. http://www.aclweb.org/anthology/W16-2323.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 86–96. https://doi.org/10.18653/v1/P16-1009.

Fadi Abu Sheikha and Diana Inkpen. 2010. Automatic classification of documents by formality. In *Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on*. IEEE, pages 1–5.

Fadi Abu Sheikha and Diana Inkpen. 2011. Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation*. Association for Computational Linguistics, pages 187–193.

Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter?: exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 259–268.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Nicola Ueffing. 2006. Self-training for machine translation. In *NIPS workshop on Machine Learning for Multilingual Information Access*.

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

Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. 2012a. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pages 1015–1024.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012b. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Jeju Island, Korea, pages 1015–1024. http://www.aclweb.org/anthology/P12-1107.

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

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

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, pages 1353–1361.