Polite Dialogue Generation Without Parallel Data

Tong Niu and **Mohit Bansal**
UNC Chapel Hill
{tongn, mbansal}@cs.unc.edu

Abstract

Stylistic dialogue response generation, with valuable applications in personality-based conversational agents, is a challenging task because the response needs to be fluent, contextually-relevant, as well as paralinguistically accurate. Moreover, parallel datasets for regular-to-stylistic pairs are usually unavailable. We present three weakly-supervised models that can generate diverse, polite (or rude) dialogue responses without parallel data. Our late fusion model (Fusion) merges the decoder of an encoder-attention-decoder dialogue model with a language model trained on stand-alone polite utterances. Our label-fine-tuning (LFT) model prepends to each source sequence a politeness-score scaled label (predicted by our state-of-the-art politeness classifier) during training, and at test time is able to generate polite, neutral, and rude responses by simply scaling the label embedding by the corresponding score. Our reinforcement learning model (Polite-RL) encourages politeness generation by assigning rewards proportional to the politeness classifier score of the sampled response. We also present two retrieval-based, polite dialogue model baselines. Human evaluation validates that while the Fusion and the retrieval-based models achieve politeness with poorer context-relevance, the LFT and Polite-RL models can produce significantly more polite responses without sacrificing dialogue quality.

1 Introduction

Generating stylistic, personality-based language is crucial to developing engaging, convincing, and trustworthy conversational agents, for their effective application in intelligent tutoring, home assistance, online reservations/purchasing, health care, etc. Most current chatbots and conversational models lack any such style, which can be a social issue because human users might learn biased styles from such interactions, e.g., kids learning to be rude because the dialogue system encourages short, curt responses, and also does not itself use politeness to set an example.[1] In this work, we focus on the important and diverse paralinguistic style axis of politeness vs. rudeness (Brown and Levinson, 1987).

Generating stylistic dialogue responses is a substantially challenging task because the generated response needs to be syntactically and semantically fluent, contextually-relevant to the conversation, as well as convey accurate paralinguistic features. This is further complicated by the fact that content and style are only available in separate unpaired datasets, as opposed to translation-type parallel datasets containing regular-to-stylistic text pairs. Hence, we need indirectly-supervised models that can incorporate style into the generated response in absence of parallel data (i.e., where the training data for the conversation, versus style components, comes from two different datasets or domains), while still maintaining conversation relevance.

In this work, we present three such weakly-supervised models[2] that can generate diverse, natural, and contextually-relevant polite (and rude) di-

[1] https://qz.com/701521/parents-are-worried-the-amazon-echo-is-conditioning-their-kids-to-be-rude/

[2] The first version of this paper with the three Fusion, Discrete-LFT, and Polite-RL models was submitted on Oct 1, 2017. The two retrieval baselines and the continuous version

Transactions of the Association for Computational Linguistics, vol. 6, pp. 373–389, 2018. Action Editor: Colin Cherry.
Submission batch: 10/2017; Revision batch: 2/2018; Published 6/2018.

alogue responses, using data from separate style and dialogue domains: the *Stanford Politeness Corpus* (Danescu-Niculescu-Mizil et al., 2013) with Wikipedia and Stack Exchange requests, and the *MovieTriples Dialogue Corpus* (Serban et al., 2016) with IMSDB movie scripts, respectively. Each of our three models is based on a state-of-the-art politeness classifier and a sequence-to-sequence dialogue model. The first model (Fusion) employs a late fusion technique to merge the response generation decoder of the dialogue model with a language model trained on polite utterances chosen by the politeness classifier. The second label-fine-tuning (LFT) model prepends to the input utterance a single politeness label whose embedding is continuously scaled by the politeness score of the target sequence during training. This score is determined by feeding the corresponding ground-truth target sequence to our politeness classifier. During test time, we show that the LFT model is able to control the politeness level of generated responses by simply scaling the label's embedding by the continuous target politeness score of our choice. Our third reinforcement-based model (Polite-RL) encourages politeness generation by using the continuous-scale politeness score of the decoder-sampled sentence as a reward (via mixed-objective policy gradient methods), i.e., polite utterances are encouraged with positive reward, and rude ones discouraged with negative reward.

Hence, our models only need a style classifier (without parallel data) to automatically influence and encourage continuous-scale stylistic language generation in a complex dialogue setup, which also requires maintaining relevance to conversational context. Each of these models requires minimal changes to the architecture of either the underlying sequence-to-sequence (Seq2seq) dialogue base model or the style classifier, and hence can modularly update the architecture with the latest state-of-the-art dialogue models or style classifiers (and for diverse styles). In addition, we also employ two retrieval-based models, where we output the response which has the highest match with the input context from a set of classifier-picked polite responses or manually-picked generic polite utterances. These two retrieval models serve as parallel investigations on the performance of our three proposed generative models above.

We conducted multiple human evaluations (for style and dialogue quality) on Amazon Mechanical Turk (*MTurk*) (Buhrmester et al., 2011) for all three models plus the base sequence-to-sequence dialogue model and the retrieval-based models, and show that while the Fusion and the two retrieval models increase the politeness level of responses at the cost of poorer dialogue quality, both our LFT and Polite-RL models can successfully produce polite responses (capturing several politeness strategies discussed by Brown and Levinson (1987)), without sacrificing dialogue coherence and relevance compared to the base Seq2seq model (hence better balance between politeness and dialogue quality). We also compare the output dialogue politeness levels of the continuous LFT model for three different politeness levels. Finally, we present several detailed qualitative and quantitative analyses, including positive and negative output examples, automatic metric results on output responses, classifier error analysis, and visualization of the RL rewards.

2 Related Works

2.1 Models for Style Transfer

Style Transfer with Parallel Data There have been multiple works on style transfer with parallel data. These tasks can often be solved by directly applying some variation of translation-based Seq2seq model discussed in the previous section. For example, Xu et al. (2012) use a phrase-based statistical model, and Jhamtani et al. (2017) use a standard Seq2seq model to convert modern language to Shakespeare-style language by treating style transfer as a translation task. Some labeled sequence transduction methods have also been proposed (Kobus et al., 2017; Yamagishi et al., 2016; Johnson et al., 2017). For example, Kikuchi et al. (2016) are able to control the length of the summarization text by feeding to the Seq2seq base model a label that indicates the intended output length in addition to the source input. Our LFT model also adopts this labeling idea, and is able to handle a similar situation but without parallel data, because by labeling each target sequence in the training set with its politeness

of the LFT model were added to the Feb 1, 2018 resubmission based on reviewer discussions.

classifier score, we are essentially converting non-parallel data to (noisy) parallel data (by using a classifier with high accuracy).

Style Transfer without Parallel Data Several previous works have looked at style transfer without parallel data, in both vision (Gatys et al., 2016; Zhu et al., 2017; Liu and Tuzel, 2016; Liu et al., 2017; Taigman et al., 2016; Kim et al., 2017; Yi et al., 2017), and text (Sennrich et al., 2016a; Hu et al., 2017; Ghosh et al., 2017; Zhao et al., 2017; Mueller et al., 2017; Wang et al., 2017; Luan et al., 2017). Among these models, some are bag-of-words based, i.e., they use style-related keywords to annotate the target sequences in the training set. For example, to control how formal the output sequences are in a EN-DE translation task, Sennrich et al. (2016a) labeled each target sequence based on whether it contains formal or informal verbs and pronouns (honorifics). To build a language model that generates utterances with the desired style, Ficler and Goldberg (2017) annotated their text with meta-data and keywords/POS tags based heuristics, while Ghosh et al. (2017) also adopted keyword spotting based on a dictionary of emotional words. The basic ideas of their models are similar to that of our LFT model. However, these keyword-spotting approaches do not fully extend to our politeness generation task, because politeness strategies follow complex patterns of grammar, word order, and phrasing (Danescu-Niculescu-Mizil et al., 2013). For example, the politeness of *please* depends on where it occurs in a sentence, and what other politeness markers it co-occurs with (e.g., 'could/would you' style counterfactual modals vs. 'can/will you' style indicative modals). Therefore, our novel polite dialogue models are based on an accurate neural classifier, which is better at capturing several compositional paralinguistic features (as visualized in Aubakirova and Bansal (2016), whose politeness classifier we extend). Moreover, our LFT and Polite-RL models can generate a continuum of style levels based on the continuously-scaled (by the politeness score) label embedding or reinforcement rewards.

Lastly, there have also been style transfer models that rely on the latent representation of text and use variational auto-encoders or cross-alignment to disentangle the representation of content and style in text (Hu et al., 2017; Shen et al., 2017; Zhao et al., 2017; Fu et al., 2018). During inference time, the latent style representation is combined with new content to generate stylized, content-preserving text. Although both fall into the category of style transfer, our task differs in two important aspects from their tasks. First, as opposed to the task of strict content preservation when rephrasing a sentence to a different style, our task is about maintaining good relevance to the context when adding style, especially useful for dialogue-based tasks. Another distinctive trait of our task is that politeness resides in a spectrum rather than a fixed category or topic (e.g., Shakespearean), and our models can treat politeness as a continuum, i.e., controlling the politeness level by adjusting the fusion rate in the Fusion model, the magnitude of the continuous label in the LFT model, or the RL weight in the Polite-RL model.

2.2 Multi-Task Learning and Style Transfer

In order to obtain a persona-based conversational agent, Luan et al. (2017) proposed a multi-task learning (MTL) based approach: they train a Seq2seq model with conversation data and an autoencoder with non-conversational persona-related data from target speakers, and share the decoder parameters of these two models so that the generated responses can be adapted to the style of the target-speaker. This way of incorporating MTL into Seq2seq learning was first investigated by Dong et al. (2015) and Luong et al. (2016) to achieve multilingual NMT. In addition, Sennrich et al. (2016b) also employed MTL to improve NMT models with monolingual (non-parallel) data. These approaches are related to our Fusion model, because we use our classifier to obtain noisy polite target sequences (non-parallel data) that a polite language model trains on; next, during inference, we combine the parameters of the language model with a generative dialogue model trained on parallel data. In general, our models are also related to previous works like Johnson et al. (2017), who adopted labeled sequence transduction methods for MTL tasks, because our task also involves adapting generated responses to different politeness styles and optimizing two sub-tasks' (namely response and politeness generation) loss functions (related to a multi-task setup).

2.3 Politeness Studies

Danescu-Niculescu-Mizil et al. (2013) created the Stanford Politeness Corpus and trained an SVM classifier using a list of useful linguistic features based on strategies from Brown and Levinson's *theory of politeness* (Brown and Levinson, 1987). Aubakirova and Bansal (2016) recently took an end-to-end neural approach to this politeness classification task by training a CNN model that directly learns to identify polite requests without using any hand-engineered features, while still improving on prediction accuracy. They also visualized what features the CNN model was learning and discovered some new features along the way. Our classifier mainly extends their work by adding a bi-directional LSTM layer (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997) before the CNN layer to capture long-distance relationships in the sentence, which leads to higher cross-domain performance.

A related early work in personality-based dialogue is Mairesse and Walker (2007), who studied introvert/extrovert personality language based on templated content and sentence planning (via personality dimensions such as hedges, tag questions, negations, subject implicitness, etc.). Relatedly, Sennrich et al. (2016a) use an English to German translation task to present a model that can generate target sequences that are either formal or informal, specifically based on honorifics-related verbs and pronouns. Our task is more general, taking into account several politeness-related paralinguistic features of Brown and Levinson (1987) and allowing end-to-end trainable stylistic dialogue generation with a polite-to-rude spectrum (based on a politeness classifier, without relying on parallel data). Moreover, our approaches allow simply replacing the politeness classifier with any other emotion or personality based language classifier to generate stylistic dialogue for that new style dimension.

3 Politeness Classification Model

In order to develop an accurate politeness classifier for effective use in stylistic dialogue response generation, we extend and improve upon the state-of-the-art CNN model of Aubakirova and Bansal (2016), and propose a bi-directional LSTM followed by a convolutional layer (see Figure 1), in order to both capture long-distance relationships in the sentence as well as windowed filter based features. For a sentence $v_{1:n}$ (where each token v_i is a d-dim word embedding vector), the LSTM layer first produces hidden states $h_{1:n}$ (where h_t is the concatenation of forward and backward hidden states at time step t). A filter m is then applied on a window of u hidden states. This produces a convolution feature $c_i = f(m * v_{i:i+u-1} + b)$, where f is a non-linear function and b is a bias term. Every feature map $c \in \mathbb{R}^{n-u+1}$ is applied to each window, so that $c = [c_1, ..., c_{n-u+1}]$. The output of the convolutional layer is then fed to a max-pooling layer (Collobert et al., 2011) which gives $C = \max\{c\}$ for the filter. Filters of various sizes are used to obtain multiple features. The result is then passed to a fully-connected softmax layer that outputs probabilities over two labels, namely *Polite* and *Rude*.

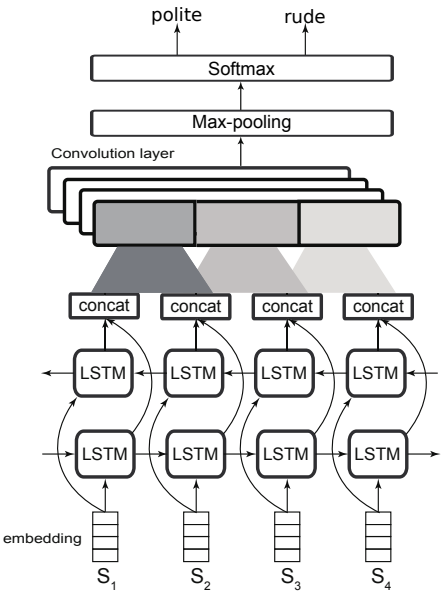

Figure 1: Our LSTM-CNN politeness classifier.

Our classification model achieves comparable in-domain accuracy and improved cross-domain accuracy over the state-of-the-art results reported in Danescu-Niculescu-Mizil et al. (2013) and Aubakirova and Bansal (2016). We will discuss these results in detail in Section 6.

4 Polite-Style Dialogue Models

In this section, we first describe our base dialogue model, i.e., the core (backbone) dialogue architecture upon which the three proposed politeness mod-

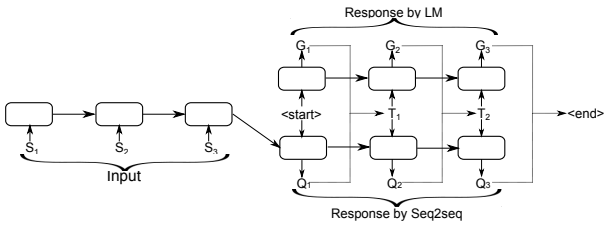

Figure 2: Fusion model: the output probability distributions of the decoder and the polite-LM are linearly mixed to generate the final decoded outputs.

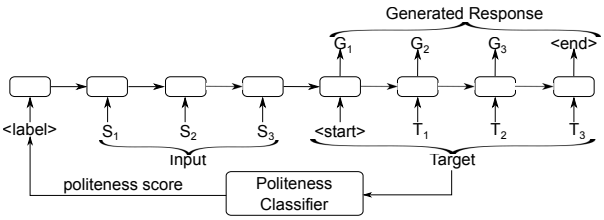

Figure 3: Label-Fine-Tuning model: during training, the embedding of the prepended label is scaled by the style classifier's continuous score on the ground-truth (target) sequence. During testing, we scale the embedding of the label by the desired (continuous) politeness score of the generated response.

els are built, and then present these three models that can generate polite dialogue responses. As a parallel investigation on the performance of our proposed models, we also employ two retrieval-based polite dialogue models toward the end.

4.1 Base Seq2seq Dialogue Model

Our base dialogue model is a simple sequence-to-sequence (Seq2seq) model that consists of a two-layer bi-directional LSTM-RNN encoder to encode the conversation history turns, and a four-layer LSTM-RNN decoder to generate the response. Additive attention from the output of the encoder is applied to the last layer of the decoder. This architecture is almost identical to that proposed by Bahdanau et al. (2015), except with more layers (similar to Shao et al. (2017)). Our base dialogue model achieves perplexity and word error rate results on par with those reported for the popular hierarchical HRED models in Serban et al. (2016), thus serving as a good base model to incorporate style into. Details will be discussed in Section 6.

4.2 Fusion Model

Inspired by the 'late fusion' approach in Venugopalan et al. (2016), our Fusion model (Fig. 2) combines the response generation decoder of the base Seq2seq dialogue model with a language model (polite-LM) trained exclusively on polite utterances. These utterances are chosen by feeding the classifier all response utterances in the MovieTriples training set, and only keeping those with politeness scores great than a certain threshold (set to 0.8 in our experiments, as will be discussed in Section 4.5). The polite-LM model is a two-layer LSTM-RNN based on Jozefowicz et al. (2016).

During inference time, we used the language model to re-score the final output of the Seq2seq decoder (for each time step) by computing a linear combination of the output vocabulary distributions proposed by the Seq2seq model and polite-LM. Specifically, let p_t^{S2S} and p_t^{LM} denote the output probability distributions proposed by the Seq2seq model and the LM model at time t, respectively. The final 'fused' distribution p_t for that time step is:

$$p_t = \alpha\, p_t^{\text{S2S}} + (1 - \alpha)\, p_t^{\text{LM}} \tag{1}$$

where the *fusion ratio* α is a hyperparameter that indicates how much Seq2seq output will influence the final output.

4.3 Label-Fine-Tuning Model

There are at least two drawbacks of the Fusion model. First, half of its output is determined by a polite language model that has not attended to the conversation context, making the response more likely to be irrelevant. Second, the model does not learn politeness during training, but is forced to be polite only during inference time. To address these two issues, we present our label-fine-tuning (LFT) model, which prepends a predicted continuous style label at the beginning of each input sentence to specify the intended politeness level.

Specifically, we add to the vocabulary a single politeness label and attach with it a trainable word embedding, just like what we would do to a normal token. Then, the way we make it continuous is by scaling its embedding vector with the (intended) politeness score of the target sequence. During training, this score is obtained by feeding the ground-truth target sequence (response) to the politeness classi-

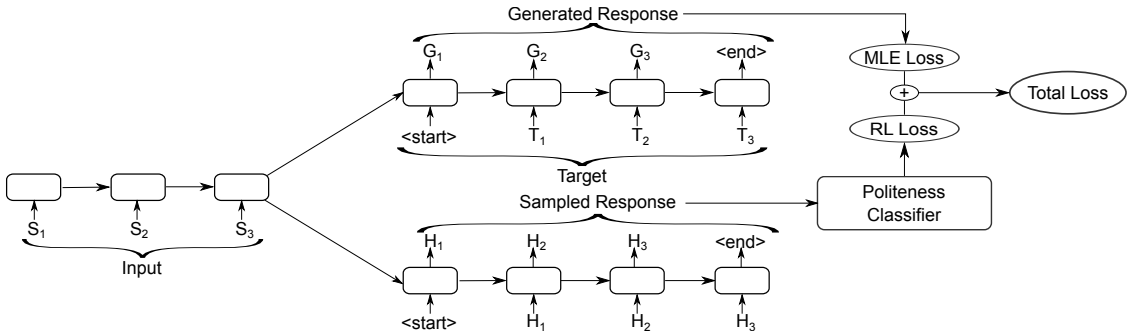

Figure 4: Polite-RL model: upper-right shows max-likelihood (ML) training with generated and ground-truth target sequences; lower-right shows RL training with a randomly sampled response generated by the model and the reward it generates after getting fed into the style classifier. Note that the attention mechanism is not shown here for clarity.

fier (see Figure 3), while during test time, we are free to scale the prepended politeness label with different scores of our choice (i.e., when we want the model to generate a polite response, we scale the label's embedding by a score between 0.5 and 1.0, whereas, to generate a rude response, we scale the embedding by a score between 0.0 and 0.5). This approach is related to the 'numerically-grounded' language model (Spithourakis et al., 2016), except that we scale the politeness label embedding by its corresponding politeness score, rather than concatenating the two as input to the LSTM.[3]

Thus, the LFT model is able to simultaneously produce polite, neutral and rude responses depending on the prepended label, similar to recent multi-label, multi-space, and zero-shot machine translation work using language identity or style labels (Sennrich et al., 2016a; Johnson et al., 2017; Ghosh et al., 2017). Intuitively, this prepended label serves as the prior for the intended style of the generated response sequence, while the source utterance serves as the prior for the content of the generated sequence. In other words, the label and the source sentence cooperatively determine what the overall response looks like.[4]

4.4 Polite Reinforcement Learning Model

The LFT model incorporates style more directly into its training procedure than the fusion model, but it still does not fully exploit the value of the style classifier since it only supervises the dialogue model once by initially classifying the style of all the target sequences in the training set. Ideally we would want the classifier to constantly monitor and influence what style the model produces. Moreover, many contexts do not naturally elicit a polite response,[5] in which case we do not want to force the model to generate an utterance that matches the target politeness score, but rather to ask the model to generate as polite and natural a response as it could. These limitations motivate us to propose the third model: Polite Reinforcement Learning model (Polite-RL), where the style classifier regularly updates the model parameters (via sampling-based policy gradient) with continuous-spectrum rewards that encourage decoder-generated response samples to be polite and discourage them from being rude.

Following work from Paulus et al. (2018), our loss function consists of two terms. The first term is the traditional maximum likelihood loss (L_{ML}), which we refer to as the *teacher forcing part*. The other one is the reinforcement learning loss (L_{RL}) based on politeness scores, which we refer to as the *reinforce part*. The total loss L then takes the form:

$$L = L_{\text{ML}} + \beta\, L_{\text{RL}} \tag{2}$$

[3] Although we trained the politeness classifier to be binary, its outputs are probabilities ranging from 0.0 to 1.0. This allows us to interpret the outputs as continuous politeness scores.

[4] Note that the position of the label did not affect the results much (e.g., Sennrich et al. (2016a) appended the label at the end of the input sequence). Moreover, our models use a bidirectional encoder, which does not distinguish between the beginning and end of the source sequence.

[5] For example, it is hard to be polite in answering questions like "What's your name?" (The most "legitimate" answer would be "My name is XX.", rather than "Thanks for asking! My humble name is XX if you would allow me to say so.")

where β is a hyperparameter indicating how much weight we want to give to the style reward component of the loss. The teacher forcing part minimizes the average of the maximum-likelihood loss at each decoding step. Specifically, let $y^* = \{y_1^*, y_2^*, ..., y_n^*\}$ be the ground-truth response for a given source (conversation history) utterance sequence x. The maximum-likelihood training objective is the minimization of the loss:

$$L_{\text{ML}} = -\sum_{t=1}^{n} \log p(y_t^* | y_1^*, ..., y_{t-1}^*, x). \quad (3)$$

We use a policy gradient method (Williams, 1992; Sutton et al., 2000) to calculate the second term in the objective function. Specifically, we sample a generated response for each input sequence (conversation history) x, and assign to it a reward R, which in our case is the politeness classifier's probability of the response classified as polite. Let $y^s = \{y_1^s, y_2^s, ..., y_n^s\}$ be the sampled response, then the reinforce part of the loss is:

$$L_{\text{RL}} = -(R - R_b) \sum_{t=1}^{n} \log p(y_t^s | y_1^s, ..., y_{t-1}^s, x) \quad (4)$$

where R_b is a baseline that helps reduce variance during training (Ranzato et al., 2016).

Note that we can invert the classifier scores or reward (by flipping the first minus sign in Eq. 4), if we want to encourage rudeness as the style, instead of politeness. This also shows that an advantage of our implementations of the LFT model over the Polite-RL model (at the cost of shallower training) is that the LFT model can multitask to simultaneously produce responses of different style labels at test time, whereas reward-based reinforcement learning can only work in one direction at a time (based on the reward sign).[6]

4.5 Retrieval-based Models

We employ two retrieval-based baseline models as a sanity check to the proposed approaches' performance: the first with oracle-level fluency, the second with additional oracle-level politeness.

Classifier-based Retrieval Following Lowe et al. (2015), for a $[X_1, Y, X_2]$ triple, our retrieval model treats the context (X_1, Y) and each response (X_2) as two documents and converts them to their TF-IDF based vectors (Ramos, 2003) to check for similarity. Specifically, we first obtain all candidate responses in the training set that are polite,[7] and calculate their TF-IDF vectors. Then for each context TF-IDF vector in the test set, we calculate its cosine similarity with that of each such polite-classified candidate response, and output the one with the highest value. Intuitively, for each context we are choosing a response that is both polite and most relevant to (having the most word overlaps with) the context.

Generic-10 This approach is similar to the one above but uses the 10 manually-chosen most-polite generic utterances as candidate responses for each context. Specifically, we collect all ground-truth polite requests from the Stanford Politeness Corpus, split each one into sentences, and then manually pick the most frequent 10 polite sentences.[8] We then determine which one to retrieve as a response for each input context, based again on the TF-IDF vector similarity method described above.

5 Experimental Setup

5.1 Datasets

As discussed above, we propose models that can deal with style data coming from an unpaired, non-parallel domain, different from the domain of the dialogue dataset. For our style (politeness) domain, we use the *Stanford Politeness Corpus* (Danescu-Niculescu-Mizil et al., 2013), which contains a collection of requests from *Wikipedia* (WIKI) editor's

[6] However, to make the reward-based model capable of multitasking, one could also prepend various politeness labels to each of the context in the training set (thus generating several examples out of one context), and encourage the generated response to be consistent with the given label. We will explore this extension in future work.

[7] We treat only responses in the higher, more-accurate percentile of $[0.8, 1.0]$ range as *polite* (and $[0.0, 0.2]$ range as *rude*).

[8] The 10 final polite sentences for Generic-10 are "thanks.", "can you help?", "can you clarify?", "no problem.", "you're welcome.", "interesting question.", "thanks for the answer.", "could you help please?", "can you elaborate?" and "nice.". The 2 rejected ones are "what have you tried?" and "what do you think?". This shortlist needed some human filtering because in the Stanford Politeness Corpus, each polite example consists of two sentences, and sometimes not both of them are polite, i.e., one of them could be neutral (more generic and task-based).

talk pages and the *Stack Exchange* (SE) question-answering communities. Based on scores from human annotators, these requests are labeled with either *Polite* or *Rude*, with each class equally consisting of 1,089 requests for the Wikipedia domain and 1,651 requests for the Stack Exchange domain. For the content (dialogue) domain, we use the popular *MovieTriples* dialogue corpus (Serban et al., 2016), which contains 245K conversations extracted from IMSDB movie scripts in *X-Y-X* triplet-utterance format, where X and Y correspond to two movie characters (and the model's task is to generate the last response).

5.2 Evaluation Methods

Human To evaluate our models' ability to generate polite responses without sacrificing dialogue quality, we conducted several comprehensive human evaluation studies on Amazon Mechanical Turk (MTurk). Specifically, we compare the three stylistic models w.r.t. the base model on both dialogue quality (i.e., context relevance and coherence) and politeness level.[9] For this, we randomly sampled 300 contexts covering all types of conversations and their generated responses from the Seq2seq base model, the three stylistic models, and the retrieval-based models. For each source input, the six responses are randomly shuffled to anonymize model identities. Each response was then annotated by two human evaluators that were located in the US, had an approval rate greater than 98%, and had at least 10,000 approved HITs (Human Intelligence Tasks) on record (to prevent those who had just started using MTurk and hence unconditionally enjoyed a high acceptance rate.). All our human evaluations are performed by two annotators (for both dialogue quality and politeness level) in order to calculate inter-rater agreement, for which we employ Cohens Kappa κ (Cohen, 1968), a score that measures the level of inter-rater agreement between two annotators on a classification problem (Artstein and Poesio, 2008). For both dialogue quality and politeness evaluations, the human raters were shown the conversation context (input) and the six shuffled responses (from the six models). Clear instructions (closely following those from Wang et al. (2017)) corresponding to each score were shown in the interface. More specifically, we asked the annotators to first read the context and each of the generated/retrieved responses, and assign a score to each response. They then scored each response on a five-point Likert scale (Likert, 1932) (for both politeness and dialogue quality), hence providing absolute measurements but in an overall comparative (relative) setting.[10] We explicitly stated that it is possible for them to find some conversation disconnected or lacking context, and encouraged them to make the best guess when in doubt. Using similar instructions (and a 300-sized sample), we also performed a separate 3-way LFT model comparison by setting its target politeness scores to 1.0, 0.5, and 0.0, respectively.

Automatic Since there do not exist ground-truth stylized versions of the response to the MovieTriples conversations, we only use automatic evaluation metrics as complementary and trend-verification information to the primary human perception studies in this work: we compute BLEU (a phrase-matching based metric; (Papineni et al., 2002)) as an approximation of dialogue quality as used by some previous work (Ritter et al., 2011; Galley et al., 2015; Li et al., 2016c). Note that we choose to report BLEU scores not to draw any immediate conclusion (Liu et al. (2016) found that BLEU does not correlate well with human studies on dialogue quality), but rather to check for match with the trends from

[9] We opted for dialogue quality rather than several separated, fine-grained metrics such as relevance, specificity, informativeness because Lowe et al. (2017) found that little additional information was provided by adding in more metrics on top of overall dialogue quality, and it also confused MTurkers in many scenarios. We had similar observations in our initial human study on MTurk.

[10] The Likert scale is a bipolar scaling method that maps each score to a text item that describes the score, e.g., our politeness level interface uses 'Polite', 'Slightly Polite', 'Neutral', 'Slightly Rude', 'Rude'; and our dialogue quality study uses 'Very good', 'Good', 'Acceptable', 'Poor', and 'Very poor', instead of the abstract scores 1-5. Note that we did not adopt pairwise comparisons because first, it will create several independent sets of pairwise results (15 sets in our case), which also raises the cost substantially, and secondly, pairwise comparison does not tell us "by how much" a response is better/equal/worse than the other. In contrast, our absolute scores can help future research compare more directly to our results. We will release our detailed instructions and MTurk interfaces, plus our annotation scores on our webpage.

human evaluation. We also compute the politeness classifier's scores as an approximation of politeness level. Sec. 6.3 discusses these results.

5.3 Training Details

We now present some important training details.[11]

Embedding Initialization For all our models, we initialized the embedding matrix with word2vec trained on Google News dataset (about 100 billion words)[12] (Mikolov et al., 2013); we use *Xavier* initializer (Glorot and Bengio, 2010) for out-of-vocabulary words.

Pretraining Following Serban et al. (2016), we pretrained the Seq2seq base model for 4 epochs with Q-A SubTle corpus (Ameixa et al., 2014), which contains around 5.5M movie subtitle Q&A pairs.

Implementation Details We used 300-dim embeddings, the *AdamOptimizer* (Kingma and Ba, 2015) with a learning rate of 0.001, and a dropout rate of 0.2. All models were trained with a mini-batch of size 96. The classifier was trained for 3 epochs, and the three proposed stylistic models were each trained for 35 epochs. The polite language model used in the Fusion model was trained until there was no improvement for perplexity on a held-out dev-set (all tuning decisions were made on the respective dev-sets). We use a balanced value of 0.5 for the fusion ratio (α in Eq. 1), and 2.0 for the RL weight (β in Eq. 4) after some light empirical tuning. Due also to the nearly perfect balance between the number of polite and rude examples in the Stanford Politeness Corpus, we set the baseline reward of Polite-RL (R_b in Eq. 4) to a constant 0.5 at all times.[13] Note that for effective and non-confusing MTurk studies, for all our models (the base model and the three stylistic models), we avoid UNK tokens to appear in the generated response, by not back-propagating the MLE loss for these tokens. We also do the same for a short list (around 10) of very offensive swear words (from Wiktionary).

[11] We will add all reproducibility details and more analysis examples in a post-publication supplement on our webpage.

[12] `https://code.google.com/archive/p/word2vec/`

[13] We also tried using a self-critical baseline as in Rennie et al. (2017), but found that our way of setting the constant-based baseline led to better responses. We speculate that this is because a self-critical approach tries to make an utterance as polite as possible, which usually leads to a few very generic and very polite responses at convergence (because the model gets a positive reward only when the sampled utterance is more polite than the greedy-decoded one).

	WIKI	SE
SVM	82.6%	65.2%
CNN	**85.8%**	66.4%
LSTM-CNN	85.0%	**70.2%**

Table 1: Politeness classification accuracies. Top results are boldfaced.

6 Results

In this results section, we first briefly present our politeness classifier (Sec. 3) and base dialogue model (Sec. 4.1) results, and then focus on the stylistic-dialogue results (retrieval and generative).

6.1 Politeness Classification Results

Following Danescu-Niculescu-Mizil et al. (2013), we use accuracy (i.e., percentage of correctly labeled messages for binary polite/rude labels) to evaluate our politeness classifier's generalization ability. Specifically, we used data from the training set of WIKI, and test on both the test set of WIKI and the entire SE (Stack Exchange) corpus. We used the same train-validation-test split setup (7:1:2) as in Aubakirova and Bansal (2016).[14] As shown in Table 1, our LSTM-CNN model improved cross-domain accuracy (while maintaining comparable in-domain accuracy) compared to that of the SVM and CNN models reported in Aubakirova and Bansal (2016). This is similar to how Zhou et al. (2015) also found that a combination of LSTM-RNNs and CNNs is superior to an LSTM-RNN or CNN alone for sentiment classification, likely because the joint model captures both long-distance relationships as well as local windowed filter-based features, and this could make it easier to separate in-domain and out-of-domain properties. We also observe more improvement on cross-domain accuracy because it has much more space for improvement, as opposed to

[14] Note that this train/dev/test split is only for verifying the strength of the classification model. The classifier used for the three proposed polite-dialogue models was trained on the entire Stanford Politeness Corpus (due to the small amount of politeness-labeled data available).

Model	PPL	PPL@L	WER	WER@L
RNN	27.09	26.67	64.10	64.07
HRED	27.14	26.60	64.10	64.03
HRED-Bidir.	26.81	26.31	**63.93**	**63.91**
Seq2seq	**25.96**	**25.85**	64.27	64.25

Table 2: PPL, WER results computed on $\{U_1, U_2, U_3\}$ and PPL@L, WER@L computed on $\{U_3\}$ conditioned on $\{U_1, U_2\}$. Lower is better for all metrics. Top results are boldfaced.

in-domain accuracy which is already very close to human performance. The higher accuracy is also important because we need a cross-domain-accurate style classifier so that it can effectively stylize responses in diverse dialogue corpora domains such as MovieTriples.

6.2 Base Dialogue Model Results

Next, in Table 2, we show that our starting point, base dialogue model is comparable in quality to a popular, representative previous model of Serban et al. (2016), trained on the same corpora with similar model architectures. We use their *Perplexity* (*PPL*) and *Word Error Rate* (*WER*) metrics. In order to have a meaningful perplexity (i.e., the probability of regenerating a reference response) comparison between two language generation models, they should have the same vocabulary set. Since the vocabulary of our politeness dialogue models is a combination of vocabulary sets drawn from the MovieTriples and Stanford Politeness corpora, for fair comparison in this section, we separately train a base Seq2seq model following exactly the vocabulary ($10,000$ most frequent tokens, plus an UNK for the rest) and preprocessing protocols from Serban et al. (2016). We bootstrapped the model with 4 epochs on the SubTle corpus (see Sec. 5.3), and then trained on MovieTriples until there was no improvement on perplexity for the validation set. The comparison for this base model with their hierarchical-encoder HRED models is presented in Table 2. As shown, we get comparable results overall on all metrics, and hence we have a good starting-point dialogue model, to which we add politeness, via the following three approaches.

6.3 Stylistic Dialogue Model Results

Primary Human Evaluation Results In this section, we present our primary human evaluation (MTurk) results on both politeness level and dialogue quality (context-relevance) of the generated response, based on two annotators and a 300-sized test sample. Table 3 shows the annotator-average scores for each of these two metrics and their absolute difference, based on our Likert scales of 1 to 5 (see Sec. 5.2). We can first see that all three of our stylistic generative models improve on politeness compared to the Seq2seq base model. However, the Fusion model's politeness gain is not statistically significant,[15] and moreover it achieves this minor politeness level improvement at the cost of significantly compromising dialogue quality (because its output is half-determined by a standalone politeness-trained LM that ignores context).

	Politeness	Quality	Difference
Retrieval	3.57	3.15	0.42
Generic-10	**3.66**	2.99	0.67
Seq2seq	3.11	3.42	0.31
Fusion	3.23	3.05	0.18
LFT	3.63	3.39	0.24
Polite-RL	3.50	**3.54**	**0.04**

Table 3: MTurk human evaluation results on politeness level and dialogue quality (as well as the absolute value difference between the two, to show balance) of the Retrieval Models, Seq2seq and the three proposed generative models (avg. of two annotators is shown here). Top results are boldfaced.

Next, we see that the LFT model is the most polite (stat. significance of $p < 0.01$ over the Seq2seq model), and also has dialogue quality close (statistically equal) to that of Seq2seq. Our final Polite-RL model wins over Seq2seq on politeness (stat. significance of $p < 0.01$) as well as achieves a small improvement in dialogue quality (though not at stat. significance level; but it is stat. significantly better in quality than Retrieval, Generic-10 and Fusion.). Moreover, the politeness levels of the LFT and Polite-RL models are statistically equal. Therefore, both models, with their training depth and multitasking trade-offs (see Sec. 4), can produce strong levels of stylistic content, without harming context-relevance.

Lastly, we can also see that our two retrieval-based models are both very polite (but not stat. sig-

[15] We test stat. significance via the bootstrap test (Noreen, 1989; Efron and Tibshirani, 1994) with 100K samples.

nificantly better over LFT); and as expected, they both have dialogue quality lower than Seq2seq, Polite-RL and LFT (stat. significance of $p < 0.01$). They also feature two of the worst balances between average politeness and dialogue quality score. This is the type of sacrifice we want to avoid from imposing on dialogue quality when building a stylistic dialogue model.

For inter-annotator agreement, the Kappa score was 0.35 (fair[16]) on Dialogue Quality and 0.46 (moderate) on Politeness. If we employ a collapsed-Likert version, where the more ambiguous and extreme scores of $\{1, 2\}$ and $\{4, 5\}$ are bucketed together,[17] we obtained a Kappa score of 0.42 (moderate) on Dialogue Quality and 0.55 (moderate) on Politeness.

Human Evaluation Results on 3-way LFT Models We also present results on a 3-way politeness level comparison MTurk study among the Polite-LFT, Neutral-LFT, and Rude-LFT models, i.e., the LFT model with three levels (scores) of scaling the prepended style label, corresponding to politeness scores 1.0, 0.5 and 0.0, respectively (Table 4, *Continuous-LFT* column). The table shows that Polite-LFT is significantly more polite than Neutral-LFT (stat. significance of $p < 0.01$), and Neutral-LFT is in turn more polite than Rude-LFT (stat. significance of $p < 0.01$). For inter-annotator agreement on this 3-way LFT study, we get a Kappa of 0.51 (moderate), and 0.61 (substantial) for the collapsed-Likert case.

We also experimented earlier with a discrete version of LFT, where we treated responses in the $[0.8, 1.0]$ range as *polite*, $[0.2, 0.8]$ as *neutral*, and $[0.0, 0.2]$ as *rude*. Instead of scaling a single label embedding with continuous politeness scores (as described in Section 4.3), we assigned to each response one of these three labels with no scaling, according to its corresponding politeness bin. The human evaluation scores for that model were 3.52, 3.09 and 2.93, respectively, which features less score difference between *neutral* and *rude* (Table 4 *Discrete-LFT* column).

	Continuous-LFT	Discrete-LFT
Polite	3.70	3.52
Neutral	3.15	3.09
Rude	1.19	2.93

Table 4: MTurk human evaluation results on politeness level of 3 LFT models, for both the continuous and the discrete versions.

Automatic Metric Evaluation Results As discussed in Sec. 5.2, we also use some automatic evaluation metrics to complement and verify the MTurk human study results. In Table 5, we present the average politeness classifier and BLEU-4 scores of responses from each model. First, we can see that our politeness classifier agrees reasonably well with the human politeness judgments in Table 3, since both identify the Retrieval-based models and LFT as the most polite, followed by Polite-RL and Fusion in descending order. We quantified this 'agreement' concretely, and found high correlation between the six human Politeness scores (Table 3 *Politeness* column) and the six automatic classifier scores (Table 5 *Politeness Score* column): Pearson correlation is 0.827 (stat. significance $p = 0.0422$), and Spearman's rank-order correlation is 0.9276 ($p = 0.0077$). Next, for BLEU scores, although these scores (as percentages) are very low (consistent with the observation in Ritter et al. (2011) and Li et al. (2016b)), their relative system-ranking still roughly agrees with that of human judgments — we found reasonably high correlation between human Dialogue Quality and BLEU (based on the six scores in Table 3 *Quality* column and Table 5 *BLEU-4* column): Pearson correlation is 0.793 (stat. significance $p = 0.0597$), and Spearman's rank-order correlation is 0.771 ($p = 0.0724$).

Hence, overall, the automatic metric evaluation again shows that without politeness training, the base dialogue model produces neutral responses on average (0.49 score), while the retrieval-based models and all three proposed generative models improve on politeness score. Also, the BLEU scores show, similar to the human study results in Table 3, that among the three proposed models, the Fusion model sacrifices the most dialog quality to become more polite, whereas the LFT and RL models main-

[16] These levels were defined by Landis and Koch (1977); also see https://en.wikipedia.org/wiki/Cohens_kappa

[17] As discussed in Weijters et al. (2010), James et al. (1984), and https://en.wikipedia.org/wiki/Likert_scale, the 'central tendency bias' makes raters avoid using the two extreme response categories.

	Politeness Score	BLEU-4
Retrieval	0.88	0.59
Generic-10	**0.93**	0.03
Seq2seq	0.49	**1.05**
Fusion	0.61	0.78
LFT	0.72	1.02
Polite-RL	0.61	0.94

Table 5: Automatic metrics: avg. politeness and BLEU-4 scores for the two Retrieval models, Seq2seq and three proposed models. Also, the politeness score of Neutral-LFT and Rude-LFT are 0.48, 0.25, resp. Top results are boldfaced.

Target Sequence	**Score**
Polite Examples	
well , thanks . that 's . i appreciate that .	0.99
⟨num⟩ , ⟨num⟩ of them in los angeles . i checked . nice work , though .	0.98
nah . i have curfew . he starts to walk away , then stops . quincy oh , by the way . congratulations .	0.97
thank you , ma'am . um , may i ask what this is regarding ?	0.91
hi , ⟨person⟩ . how are you ?	0.89
i know . amazing .	0.89
Rude Examples	
you really should pay more attention to what you read , ⟨person⟩ .	0.11
they were in a car accident .	0.10
you calling more of your stupid friends again ? ya prick .	0.05
are you a bad boy ?	0.05
then she 's a liar .	0.04
excuse me , does that flask belong to this man ?	0.03
oh , well , excuse me all to hell .	0.02

Table 6: Analysis of politeness classifier: sampled responses from MovieTriples dialogue corpus (without cherry-picking) and their politeness score. The double line separates polite and rude classified responses.

tain comparable quality with improved politeness over the base model (Seq2seq). For the retrieval models, we again see that their politeness levels are better than LFT and RL models, but with a corresponding loss in dialogue quality.

7 Analysis

7.1 Analysis of Politeness Classifier

We start our analysis by providing qualitative examples of how well our politeness classifier performs on the target sequences from MovieTriples train dataset. This is important to check because the classifier is trained on Wikipedia (Wiki) admin request messages, and while our LSTM-CNN performs better on cross-domain Stack Exchange (SE) data, the MovieTriples dialogue corpus is still quite different and diverse in domain from both Wiki and SE. Hence, it is important to have a reasonably accurate politeness classifier such that it can provide useful labels and rewards for our polite-dialogue models. Table 6 presents some randomly-selected (i.e., non-cherry-picked) responses from MovieTriples and their politeness classifier scores. We can see that the classifier provides a reasonably correct score a majority of the time, capturing several psycholinguistic politeness strategies mentioned in Danescu-

Input	X: for what ? Y: for forcing me to face my life honestly for the first time .
Retrieval	X: what 's the c . w . for ?
Generic-10	X: thanks for the answer .
S2S	X: that 's not true .
Fusion	X: i do n't know . I mean , it 's just a feeling , but that 's not what i 'm saying .
LFT	X: you 'll be fine .
RL	X: i wish i could .
Input	X: no, but i wanted to. Y: are you a total lying , scumsucking pig , ⟨person⟩ ?
Retrieval	X: so ⟨person⟩ , what did it total out at ?
Generic-10	X: no problem .
S2S	X: oh , that 's ridiculous .
Fusion	X: well , you 're not a little funny .
LFT	X: i do n't think so , sir .
RL	X: you 're hurting me !
Input	X: you 're sweet to say so . Y: pretty song .
Retrieval	X: is there a song you do n't know ?
Generic-10	X: can you clarify ?
S2S	X: ⟨person⟩ , i know .
Fusion	X: you 're a good man .
LFT	X: i love this song .
RL	X: you sound like a goddess .
Input	X: i find it hard to believe myself . Y: his lips , unused as they are to such a contortion , twisted and stretched . into a smile .
Retrieval	X: thank you . and always smile .
Generic-10	X: nice .
S2S	X: how can you be so sure ?
Fusion	X: you 're a good man , mr .
LFT	X: your lips are well .
RL	X: your lips are so beautiful .

Table 7: Output dialogue response examples by Retrieval, Generic-10, Seq2seq (denoted as S2S) and the 3 generative polite models Fusion, LFT, and RL (shows conversation history turns of X and Y, and then the generated 3rd turn response by X).

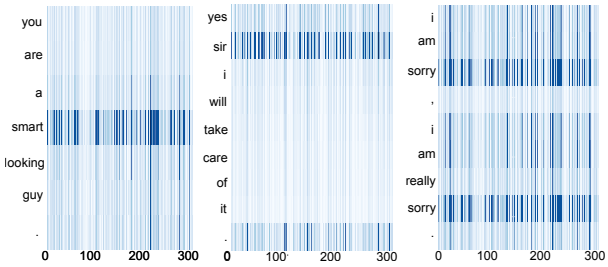

Figure 5: Saliency heatmaps of the classifier's attention (reward for sampled responses in Polite-RL model).

Niculescu-Mizil et al. (2013), e.g., positive ones such as gratitude, deference, greeting, positive lexicon, indirection, indicative modal, and negative ones such as negative lexicon, direct question, direct start, 2nd person start. However, it does occasionally give strongly polite or rude scores to some mild or neutral responses, e.g., "they were in a car accident", showing scope for classifier improvements.

7.2 Output Examples of Stylistic Dialogue

Next, we show some output examples of our polite dialogue models w.r.t. the base Seq2seq model as well as the retrieval-based models. We use these examples to demonstrate the politeness strategies our proposed generative models have learned (in Table 7). In the first example, our stylistic models use politeness strategies such as indirection, positive lexicon and counterfactual modal (Danescu-Niculescu-Mizil et al., 2013). This example also illustrates the behavior of the Retrieval model, i.e., most of the time it just outputs an utterance that has word overlap with but totally irrelevant to the context. Thus although all its retrieved responses have oracle-level fluency and grammaticality, its average dialogue quality score in the human evaluation is still not as good as that of Seq2seq.

In the second example, Fusion uses indirection, while LFT is being polite even when disagreeing with the abusive language from Y. This example also shows that Generic-10, due to its limited space for retrieval, oftentimes fails to provide a relevant answer, although it is the most polite one since its candidate responses are manually picked. In the third example, Fusion and LFT both use positive lexicon, and RL makes a compliment. In the fourth example, each of the three proposed models uses positive lexicon. It is worth noting that in the last example, while LFT and Polite-RL seem to provide a relevant compliment, they are actually complimenting the wrong person. This kind of issue motivates us toward creating persona-based (Li et al., 2016c) politeness models for future work.

7.3 Visualization of Polite-RL Reward

Using derivative saliency (Simonyan et al., 2013; Li et al., 2016a; Aubakirova and Bansal, 2016), we also visualize how much each token in the sampled response contributes to the classifier's reward during Polite-RL model's training. Fig. 5 shows three such heatmaps that correspond to the magnitudes of the derivative in absolute value with respect to each dimension. The figures clearly show that the classifier has learned to identify multiple politeness strategies, e.g., "smart" (deference), "sir" (polite address), and the two "sorry"s (apologizing).

8 Conclusion and Future Work

We first presented three diverse generative models that can generate rich polite-to-rude spectrum dialogue responses (based on the politeness theories by Brown and Levinson (1987)), without using any parallel data (which is usually assumed for tasks such as machine translation) and only relying on a style classifier. Via multiple human evaluation studies and automatic metrics, we demonstrated that all three models generate more polite responses (displaying several politeness strategies discussed in previous psycholinguistic works), while LFT and Polite-RL are able to do so without losing dialogue quality, as opposed to the Fusion model as well as the two retrieval-based models.

In future work, there is still much room for improvement on the politeness as well as dialogue quality side, and one could employ more recent, advanced models such as variational, adversarial, and decoder-regulation techniques.

Though we focused on politeness for the scope of this paper, our models can be easily generalized to other emotion and personality styles (only relying on a style classifier), hopefully contributing towards the valuable paradigm of human-like and engaging intelligent tutors and personal assistants. In future work, our polite-RL model could also be extended to stylistic task-based dialogue generation, where both content preservation and style transfer are needed, potentially by disentangling politeness and content

of the generated response and then only feeding the politeness portion to the classifier for RL training.

Acknowledgments

We thank the action editor and the anonymous reviewers for their helpful comments and discussions. This work was supported by DARPA (YFA17-D17AP00022), Facebook ParlAI Research Award, Google Faculty Research Award, Bloomberg Data Science Research Grant, and Nvidia GPU awards. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the funding agency.

References

David Ameixa, Luisa Coheur, Pedro Fialho, and Paulo Quaresma. 2014. Luke, I am your father: Dealing with out-of-domain requests by using movies subtitles. In *International Conference on Intelligent Virtual Agents*, pages 13–21. Springer.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Malika Aubakirova and Mohit Bansal. 2016. Interpreting neural networks to improve politeness comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2035–2041.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of International Conference on Learning Representations*, pages 1–15.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*, volume 4. Cambridge University Press.

Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5.

Jacob Cohen. 1968. Weighted Kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 250–259.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732.

Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC press.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 663–670.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 445–450.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*, pages 249–256.

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

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, PMLR 70*, pages 1587–1596.

Lawrence R. James, Robert G. Demaree, and Gerrit Wolf. 1984. Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69(1):85.

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 Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics, v. 5*, pages 339–351.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR abs/1602.02410*.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1857–1865.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations*.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of Recent Advances in Natural Language Processing*, pages 372–378.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016a. Visualizing and understanding neural models in NLP. In *Proceedings of North American Chapter of the Association for Computational Linguistics-HLT*, pages 681–691.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A diversity-promoting objective function for neural conversation models. In *Proceedings of North American Chapter of the Association for Computational Linguistics-HLT*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016c. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 994–1003.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*.

Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 469–477.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Ryan Lowe, Nissan Pow, Iulian V. Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2015)*, pages 285–294.

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1116–1126.

Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 2017. Multi-task learning for speaker-role adaptation in neural conversation models. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 605–614.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *Proceedings of International Conference on Learning Representations*.

François Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representa-

tions in vector space. In *Proceedings of International Conference on Learning Representations Workshop*.

Jonas Mueller, David Gifford, and Tommi Jaakkola. 2017. Sequence to better sequence: Continuous revision of combinatorial structures. In *International Conference on Machine Learning*, pages 2536–2544.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses*. Wiley New York.

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*, pages 311–318.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *Proceedings of International Conference on Learning Representations*.

Juan Ramos. 2003. Using TF-IDF to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, volume 242, pages 133–142.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of International Conference on Learning Representations*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, page 1197.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 583–593.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of North American Chapter of the Association for Computational Linguistics*, pages 35–40.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *The Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 3776–3784.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6833–6844.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Georgios P. Spithourakis, Isabelle Augenstein, and Sebastian Riedel. 2016. Numerically grounded language models for semantic error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 987–992.

Richard S. Sutton, David Mcallester, Satinder Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, pages 1057–1063. MIT Press.

Yaniv Taigman, Adam Polyak, and Lior Wolf. 2016. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*.

Subhashini Venugopalan, Lisa Anne Hendricks, Raymond J. Mooney, and Kate Saenko. 2016. Improving LSTM-based video description with linguistic knowledge mined from text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1961–1966.

Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. Steering output style and topic in neural response generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2140–2150.

Bert Weijters, Elke Cabooter, and Niels Schillewaert. 2010. The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3):236–247.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*, pages 5–32. Springer.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2899–2914.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a

sentence in Japanese-to-English neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210.

Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised dual learning for image-to-image translation. In *Proceedings of International Conference on Computer Vision*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of International Conference on Computer Vision*.