Polite Task-oriented Dialog Agents: To Generate or to Rewrite?

Diogo Silva, David Semedo, João Magalhães
NOVA LINCS - Universidade NOVA de Lisboa
Lisbon, Portugal
dmgc.silva@campus.fct.unl.pt, {df.semedo, jmag}@fct.unl.pt

Abstract

For task-oriented dialog agents, the tone of voice mediates user-agent interactions, playing a central role in the flow of a conversation. Distinct from domain-agnostic politeness constructs, in specific domains such as online stores, booking platforms, and others, agents need to be capable of adopting highly specific vocabulary, with significant impact on lexical and grammatical aspects of utterances. Then, the challenge is on improving utterances' politeness while preserving the actual content, an utterly central requirement to achieve the task goal. In this paper, we conduct a novel assessment of politeness strategies for task-oriented dialog agents under a transfer learning scenario. We extend existing generative and rewriting politeness approaches, towards overcoming domain-shifting issues, and enabling the transfer of politeness patterns to a novel domain. Both automatic and human evaluation is conducted on customer-store interactions, over the fashion domain, from which contribute with insightful and experimentally supported lessons regarding the improvement of politeness in task-specific dialog agents.

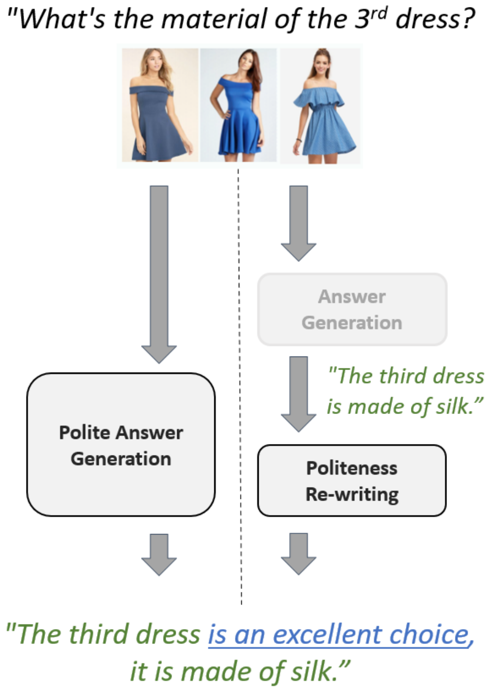

"What's the material of the 3rd dress?
Answer Generation
"The third dress is made of silk."
Polite Answer Generation
Politeness Re-writing
"The third dress is an excellent choice, it is made of silk."

Figure 1: Politeness can be introduced either by incorporating it in the generation step or as a rewritting step. In this example the politeness strategy adopted is the use of a positive lexicon.

1 Introduction

In a conversational scenario, the tone of voice used by interlocutors is a key aspect towards achieving fruitful, engaging, and natural user-agent interactions (Brown et al., 1987; Niu and Bansal, 2018). This is deeply rooted in the fact that discoursing in a polite manner, is a social trait of human conversations, that when left unattended by dialog agents, can lead to an immediate perception of artificial discourse and lack of intelligent behavior, which in turn leads to poor engagement.

Task-oriented dialog agents require simultaneously keeping the user engaged while achieving the task goal, whether it is selling a product, booking a restaurant or simply providing assistance. This requires *informative and correct answers, embedding domain-specific language, while keeping a polite tone of voice*. Being able to accomplish this, has an impact that extrapolates isolated conversations. For example, in the fashion world, the tone and the way the customer is addressed are strongly linked to the brand culture (Sousa et al., 2021) (e.g. more eloquent vs. more casual and youthful discourse).

While politeness is a deeply seeded cultural concept and difficult to fully generalize (Meier, 1995), it has been recently approached from a computational perspective (Danescu-Niculescu-Mizil et al., 2013; Niu and Bansal, 2018; Madaan et al., 2020) under the framework of (Brown et al., 1987), which divides politeness strategies in *a) negative politeness* - where polite discourse is achieved by expressing restraint, thus avoiding being direct - and

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 304 - 314
May 26, 2022 ©2022 Association for Computational Linguistics

b) positive politeness - where an explicit attempt of expressing solidarity, optimism and gratitude is made. Danescu-Niculescu-Mizil et al. (2013) took a pioneering approach and proposed to approximate these strategies by creating a human annotated politeness corpora, and training a classifier to capture general linguistic patterns of both negative and positive politeness. Recent works leverage on such classifier to develop either generative (Niu and Bansal, 2018; Firdaus et al., 2020) or rewriting-based (Madaan et al., 2020; Fu et al., 2020) approaches. Figure 1 contrasts these approaches, with respect to politeness strategies. While these have been applied to generic and domain-agnostic scenarios, it remains unclear how well such principles transfer to task-specific domains.

In this work, we assess under a transfer-learning scenario, the applicability of both generative and rewriting politeness approaches to a novel domain. Specifically, we use the challenging fashion domain as a use case, given its vocabulary complexity and highly specific-nature [1]. Namely, we propose to overcome the lack of labeled data by extending state-of-the-art generative (Niu and Bansal, 2018) and rewriting (Madaan et al., 2020) approaches, respectively, towards allowing each of them to overcome the domain-shift, and transferring linguistic politeness constructs to a novel (fashion) domain.

This is one of the first works to study politeness approaches for task-oriented dialog agents, contributing with:

- An adaptation of generative and rewriting politeness approaches (section 3), enabling transfer learning for specific domains.
- Comprehensive experiments (section 5), leading to valuable insights regarding how politeness approaches deal with the content-preservation vs. politeness improvement trade-off, in task-oriented dialog agents.
- A user-centered study that supports and confirms the conclusions of the automatic evaluation (section 5).
- Explored politeness on a novel domain, conversational assistants on the fashion domain (Saha et al., 2018), exposing the opportunities for improving politeness.

2 Related work

The importance of politeness in social interactions and its impact in the projected self-image during social interactions has been studied for decades (Brown et al., 1978, 1987). These concepts were later reviewed and refined (Watts, 2019) with new work (Bargiela-Chiappini, 2003) proposing the label of 'polite behavior' to separate it from the theoretical and cultural baggage of the term face-work. More recently, Danescu-Niculescu-Mizil et al. (2013) introduced a labeled dataset (Stanford Politeness Corpus), along with a politeness classifier to enable further research as an NLP task. Additionally, they look into how politeness relates to the speaker's status and power within their community. Later work (Aubakirova and Bansal, 2016) introduced a new politeness classifier and several visualization techniques to gain further insight into linguistic markers of politeness. These visualization techniques reveal novel politeness strategies not considered originally, namely how punctuation affected politeness scores. The introduced politeness classifier uses a CNN and does not use politeness strategies as features while having higher accuracy.

Politeness as an NLP task has seen recent interest. Niu and Bansal (2018) uses the Stanford Politeness Corpus to investigate politeness generation models. Politeness generation here is treated as part of the answer generation task with models producing answers already in their polite form, using Reinforcement Learning and a novel politeness classifier. A Multilingual approach is taken in (Firdaus et al., 2020) where courteous responses are generated in a customer care scenario. Madaan et al. (2020) sees politeness as a style transfer task where politeness is introduced onto an utterance by rewriting it. This work uses a politeness classifier to label the Enron corpus (Klimt and Yang, 2004), and applies a transformer-based (Vaswani et al., 2017) style transfer pipeline to the utterance, using a tagger and generator approach. In a similar vein, in (Golchha et al., 2019) the authors transform neutral customer service replies into courteous ones.

Hence, we follow a similar line of work and propose to enrich fashion dialog agents with politeness. Saha et al. (2018) introduced a large-scale multimodal fashion dialog dataset (MMD) built semi-automatically, using field experts, accompanied by two RNN (Cho et al., 2014) models capable of emulating the system responses in a multimodal

[1] The established politeness classifier of Danescu-Niculescu-Mizil et al. (2013) lacks $\approx$ 15k terms from the considered fashion dialog corpus.

scenario. Due to its domain, it carries mainly neutral and polite dialog. To the best of our knowledge, there is no task-oriented conversational dataset to study politeness and we propose to fill this research gap.

3 Task-Specific Polite Dialog Agents

We consider two distinct methods of producing politeness and evaluate how each deals with domain changes: **polite answer generation** and **politeness rewriting**. We adapt each model to allow it to use transfer learning, in particular, transfer politeness patterns to a different domain, the fashion domain.

3.1 Politeness through Utterance Generation

Politeness can be improved in a generative manner, where an answer generation model learns to do so, by *merging answer generation and politeness generation in the same task*. This type of approach makes the work of the decoder two-fold: it needs to be able to accurately understand the context and produce an accurate answer, but it also needs to improve the politeness of the produced answer.

We adopted the Polite-RL generative approach (Niu and Bansal, 2018) based on a Seq2Seq model that receives the conversation history to produce a polite answer. The model is trained with Reinforcement Learning that leverages a Politeness Classifier (we will refer to as Classifier) to estimate the politeness of a sampled answer. Polite-RL uses the politeness score of a sampled utterance as a measure of politeness that acts as the Reinforcement Learning component of the loss function (see appendix A.4), to guide the generation towards a more polite output. We focused on improving the used embeddings to include a novel lexicon, given that the fashion domain (Saha et al., 2018) differs significantly from the training data (Danescu-Niculescu-Mizil et al., 2013), making out-of-vocabulary situations a major issue. Originally, this model uses embeddings initialized using a Word2Vec model trained on the Google News dataset (Mikolov et al., 2013). Despite its vocabulary size, the dataset's vocabulary can still leave out a significant portion of the terms used in the fashion-specific datasets (mainly clothes' names and attributes), due to its highly specific domain (Saha et al., 2018).

Looking at Table 7, we observe that politeness can be applied in several different ways, making it important to take into account the utterance as a whole to better understand how phrase structures affect its tone. In the Polite-RL (Niu and Bansal, 2018) model, these strategies are introduced implicitly by the politeness Classifier as the Seq2Seq model is not explicitly trained on politeness data. With this in mind, to improve the adaptability of this implementation and reduce the impact of this separation in the vocabulary, we introduce a new set of embeddings that accounts for the additional tokens from the novel domain dataset. These embeddings were obtained by training a Word2Vec (Mikolov et al., 2013) model on a concatenation of the MMD (Saha et al., 2018) - a conversational dataset on the fashion domain (see section 4.1) - and the original Politeness corpus. We will refer to these embeddings as Domain-Extended embeddings (DE).

3.2 Politeness through Utterance Rewriting

Politeness rewriting *separates the task of politeness generation from answer generation*. This enables tackling politeness individually, and avoid its dependence on the answer generation task.

For this approach, we adopt Tag-and-Generate (Madaan et al., 2020), which is composed of two main components: Tagger and Generator. The **Tagger** is responsible for extracting style makers from the utterance and adding a [TAG] token where new markers should be introduced. The style markers are defined using a TF-IDF-based approach that compares the relevance of an n-gram on the polite and rude subset of data. The **Generator** takes the tagged utterance and replaces the [TAG] token with polite style markers. This approach follows the assumption that the extracted style markers are good markers for politeness, meaning that if the model is dealing with a poor set of style markers then the results can be destructive and nonsensical.

Models such as this, apply politeness strategies in an explicit manner, Table 7. The Generator learns the best way to add each politeness strategy onto a given utterance, by observing how each style marker is used throughout the training data. For honorifics, ideally, the model learns to place them immediately before surnames.

With the Tagger architecture in mind, we focused on using Transfer Learning to better adapt the model to the fashion domain. For the rewriting part, we hypothesize that using the style markers previously learned on the original dataset (Enron) will lead to improved politeness scores. To deal with the out-of-the-domain-vocabulary problem, we propose to curate the extracted style markers, by excluding domain-specific words and terms from being classified as style markers, thus leading to more representative style markers. To assess how this affects generation quality, we define four training setups:

- **RW-Enron:** Original model trained on the Enron dataset (Klimt and Yang, 2004).
- **RW-Fashion:** Model trained on the fashion-domain dataset, using polite and rude utterances, i.e. utterances with a politeness score above 0.9 and between 0.5-0.6 respectively.
- **RW-Fashion-Clean:** Similar to the previous model, but we force the model to ignore style markers associated to product nouns. For example, "scarf" and "trousers", shouldn't be counted as a style marker of politeness.
- **RW-Mixed:** This model learns the style markers on the original domain (Enron) and is trained on the fashion dataset. This way the model circumvents the noisy style markers extracted from the fashion data. Effectively transferring knowledge learned on politeness annotated data to the fashion domain.

4 Experimental Setup

4.1 Datasets and Protocols

In our experiments, 3 datasets were considered:

Stanford Politeness Corpus (SPC) - This is the dataset used for politeness conditioning, by training the Politeness proxy classifier (Niu and Bansal, 2018). This corpus is composed of requests (Wikipedia and Stack Exchange) annotated by 5 humans. We follow (Niu and Bansal, 2018) and use the original data splits.

Enron - Collection of emails exchanged in the Enron company (Klimt and Yang, 2004) - originally used to train the Tag-and-Generate model (Madaan et al., 2020) - that we adopt as the original domain, in a domain-transfer scenario. We consider an automatically annotated subset of Enron, with 212k polite and 51k rude utterances for training, 27k polite and 5.8k rude for validation, and 26k polite and 5.8k rude utterances for testing.

MMD - This dataset comprises multi-turn dialogs for the fashion domain (Saha et al., 2018), which we use as the target domain. We first create the **MMD-R** subset, comprised by system utterances that correspond to product(s) recommendations(s) to expose the model to domain-specific product lexicon, resulting in 380k/81k/81k utterances for training/validation/testing. A second subset is created, **MMD-A**, comprising all *neutral*[2] and *polite* system utterances with more than 5 tokens, resulting in 453k/116k/116k utterances. The **MMD-A** subset generalizes **MMD-R** to include utterances from multiple dialog intents.

Please kindly refer to Appendix A.2 for more details regarding each dataset (annotation protocol, splits, and others).

4.2 Metrics

For evaluation, we will focus mainly on two aspects of the generated utterances: **a) Politeness Improvement** and **b) Content Preservation**. With **a)**, we focus on understanding if each resulting utterance is in fact more polite than the original one. For this, to automatically quantify politeness, we follow (Niu and Bansal, 2018) and compute the average Politeness Score (*Pol.*) using its politeness classifier, where 1.0 is polite, 0.5 neutral and 0.0 is rude. In **b)**, we focus on understanding whether or not the model can preserve the original content. Thus, we follow previous work (Niu and Bansal, 2018; Madaan et al., 2020) and evaluate the results using BLEU **(B)** (Papineni et al., 2002), ROUGE **(R)** (Lin, 2004), and METEOR **(M)** (Denkowski and Lavie, 2011). Given the subjective nature of the task, we complement our evaluation with human evaluation.

4.3 Model Variants and Implementation

For evaluation, we refer to politeness answer generation variants as **Gen** and rewriting variants as **RW**. For **Gen**, we use the original embeddings. The generative approach with domain extented embeddings (section 3.1) is referred as **Gen+DE**. For rewriting, the 4 proposed variants (section 3.2) are referred as **RW-Enron**, **RW-Fashion**, **RW-Fashion-C**(lean) and **RW-Mixed**.

Regarding models implementation, for RW variants we use the original hyper-parameters, and both components are trained using a 4-layer 4-

[2] Due to its nature, the number of *rude* utterances in MMD is minimal, leading to a high imbalance.

Models	B	R	M	Pol.
Gen	68.54†	85.30	48.80†	69.95
Gen+DE	66.32	85.55†	45.02	64.47
Gen+DE (β=5)	33.56	63.63	27.64	75.21
Gen+DE (β=10)	29.41	58.66	24.90	78.77†
RW-Enron	70.38	86.68	51.51	**82.24**‡
RW-Fashion	85.03	83.72	58.99	79.76
RW-Fashion-C	86.71	86.44	60.23	78.42
RW-Mixed	**87.78**‡	**87.22**‡	**60.80**‡	80.70

Table 1: Politeness generation vs. rewriting results. † represents the highest result among Polite-RL (Gen) variations and ‡ represents the highest results among Tag-and-Generate (RW) models.

head transformer block and 512-dimensional embeddings, for 5 epochs. For Polite-RL we also use the original hyper-parameters, but we tuned the batch size b and the β parameter, the weight of the politeness component of the computed loss. Refer to Appendix A.4 for model tuning details.

5 Results and Discussion

5.1 Politeness Generation or Rewriting?

Automatic-Evaluation. We start by comparing how the adapted generative (section 3.1) and rewriting (section 3.2) politeness approaches perform on the fashion domain, in terms of politeness and content preservation. The Gen and RW models were evaluated on the **MMD-R** and **MMD-A** datasets, respectively. Table 1 shows the evaluation results for both models and their variants. From these results, it is evident that rewriting variations (RW) outperform the generation-based ones (Gen) across all metrics, due to their need to attend to two tasks. For content preservation, the results from Gen are consistently behind its RW counterparts, with all variations of the RW model outperforming the generation-based models. Regarding politeness, the scores paint a similar picture with Gen models trailing behind and only reaching near when content preservation is significantly neglected (higher β value, the weight given to politeness in Polite-RL). Despite this, all models are able to post the politeness score on the polite spectrum (Pol. > 0.5), according to the politeness classifier.

Human-Evaluation. Automatic metrics offer a quick and reproducible way of evaluating work, however, they lack the depth needed to accurately evaluate subjective topics like politeness (Danescu-Niculescu-Mizil et al., 2013). To supplement our previous automatic analysis of the proposed changes to the RW setup, we ran a crowdsourcing experiment to assess the tone and grammar[3] of generated utterances. For this, we randomly sample 100 utterances from the **MMD-A** test set and then collect the generated utterance for each of the RW variations and Gen+DE. For each utterance, 3 annotators were asked to rate its tone on a scale of 1 to 3 (1=Rude, 2=Neutral, 3=Polite). For grammar, annotators were asked to give a binary rating, whether the utterance was grammatically correct or not (0=No, 1=Yes). Annotators were provided an example utterance for each possible value. We obtained $\approx 82\%$ agreement on grammar and $\approx 77\%$ for politeness. The results are show in Table 2. For a given utterance, the agreement was the measure of how many annotators labeled it the same.

Models	Politeness	Grammar
Reference	2.453	0.793
Gen+DE	2.170	0.583
RW	**2.497**	0.770
RW-Fashion	2.437	0.733
RW-Fashion-C	**2.497**	**0.790**
RW-Mixed	2.453	0.767

Table 2: Human evaluation results for Politeness and Grammar, on 100 utterances.

Regarding politeness, the Gen+DE model scored lower than the Reference, whereas all RW setups matched or improved on it. In particular, only RW-Fashion failed to improve and both RW-Fashion-C and RW were able to outscore the reference. When looking into rating distribution, we noted that of the 1800 annotations, only in 28 occasions did an annotator consider the utterance rude, and never 2 annotators agreed that an utterance was rude, showing that all models are able to keep the text neutral. For grammar, none of the models were able to score higher than the reference, and Gen+DE was rated significantly lower.

From the ratings, we note that there is a significant gap between Generative and Rewriting approaches, similar to the automatic evaluation. Additionally, Gen+DE underperforming with respect to the reference shows that generally, the model was not able to improve on the utterances' tone often leading to incoherent generations. On another note, the performance of the RW-Fashion-C shows

[3] We included grammar to understand if the models were reducing the quality of the re-written utterances.

that style marker curation can be the way forward for rewriting approaches.

5.2 Style Marker Domain Transfer

In the previous section, we observed that politeness rewriting is the top-performing approach to improve the politeness of task-oriented dialog agents. In this section, we perform a finer-grain evaluation of the rewriting model variants, i.e. the style marker-based model (**RW**), under a domain transfer setting. As this corresponds to an utterance rewriting task, we look for high content preservation paired with high politeness score.

To perform this experiment, we use the **MMD-A** subset, comprising diverse system utterances (recommendation, answering product questions, etc.). Then, we follow Madaan et al. (2020) and use the politeness classifier to split the subset into 10 buckets, corresponding to a 10-bin histogram over politeness scores.

Table 1 also depicts the results, where we compare the four RW model variations on the **MMD-A** test set. We observe that for content preservation, top performance is achieved by the RW-Mixed model, across all three content metrics. Additionally, we note that the RW-Fashion-C model is a step up on RW-Fashion, showing that excluding domain-specific words from the style attribute list helps preserve content. However, for the RW-Enron model, which is restricted to the original domain, the results were significantly lower.

Regarding Politeness scores, the RW-Enron model outperforms all the others specifically trained on **MMD-A** data. The Mixed model also performed better than its RW-Fashion and RW-Fashion-C counterparts. The success of RW-Enron in politeness score and RW-Mixed in content preservation shows that leveraging out-of-domain style markers yields positive results for neutral domains, where it is difficult to extract style markers. This also shows how models can vary in the way they add style markers. While RW-Enron does several and significant changes, thus having lower content preservation scores, RW-Mixed does less but more informed and in-domain additions.

5.2.1 Utterance Tone and Length Impact

To identify the impact of utterance tone (rude or neutral) and utterance size on the models' performance, we prepared a set of distinct scenarios covering the different aspects of utterances' tone and length. These two utterance traits were chosen due to the following reasons:

	RW Model:	Enron	Fashion	Fashion-C	Mixed
BLEU	SN	83.24	88.30	**91.11**	89.45
	MN	82.31	96.22	**97.10**	96.28
	L	65.71	95.71	97.55	**97.68**
ROUGE	SN	91.04	91.32	**93.25**	92.22
	MN	89.18	94.76	**95.67**	95.40
	L	84.77	97.69	98.63	**99.05**
METEOR	SN	56.53	60.23	**64.24**	61.43
	MN	56.22	70.75	**72.07**	70.44
	L	47.98	**73.54**	70.83	72.06

Table 3: Utterance length impact in content preservation. We fix utterances tone to *neutral*.

Utterance tone - It is important to gauge models' ability to adapt to different levels of politeness. Namely, the difficulty of improving from rude to polite differs from neutral to polite. Additionally, models are trained on the neutral politeness bucket of data(to perform style transfer to polite tone), which may bias their performance towards a particular tone.

Utterance Length - During the initial experiments, we observed that the models tended to leave longer utterances untouched, and we wanted to measure the extent of that behavior for different utterance sizes.

To assess these two aspects, we evaluated our proposed four RW model variations, under a set of scenarios obtained by systematically varying the length and tone utterance properties, resulting in the following 5 scenarios:

Long (L) - Comprises of neutral[4]long utterances from the **MMD-A** test set. These correspond to recommendation of products thus being very rich in fashion-specific terms. We obtained 88 utterances.

Short & Rude (SR) - Short utterances obtained from the **MMD-A** test set. This corresponds to utterances belonging to the P_0 or P_1 buckets, i.e. utterances deemed *rude*, with less than 17 tokens. In total, we obtain 134 utterances.

Short & Neutral (SN) - Same strategy as **SR** but utterances are picked from the P_5 bucket instead - halfway of the politeness scale, meaning that utterances are deemed as *neutral*. In total we obtain 2.5k utterances.

[4] Due to the low utterance count, a long and rude test scenario was not viable.

	RW Model:	Enron	Fashion	Fashion-C	Mixed
BLEU	SR	70.22	84.28	**87.27**	86.36
	SN	83.24	88.30	**91.11**	89.45
	MR	71.35	87.62	**91.71**	90.75
	MN	82.31	96.22	**97.10**	96.28
ROUGE	SR	80.40	87.31	**88.95**	87.01
	SN	91.04	91.32	**93.25**	92.22
	MR	82.78	88.95	**91.68**	91.11
	MN	89.18	94.76	**95.67**	95.40
METEOR	SR	47.82	57.86	**60.23**	57.11
	SN	56.53	60.23	**64.24**	61.43
	MR	49.36	60.66	**63.10**	62.25
	MN	56.22	70.75	**72.07**	70.44

Table 4: Impact of tone of voice - Neutral (N) and Rude (R) - in both Short (S) and Medium (M) length utterances.

RW Model:	Enron	Fashion	Fashion-C	Mixed
Enron	+6.50	+13.15	+17.26	**+22.36**
MMD-A	**+5.76**	+3.28	+1.94	+4.22
SR	**+10.99**	+6.55	+5.34	+7.18
SN	**+8.83**	+6.49	+3.56	+6.89
MR	**+9.34**	+5.07	+3.72	+5.88
MN	**+6.11**	+1.21	+1.00	+1.96
L	**+2.45**	+1.02	+2.20	+2.43

Table 5: Relative improvement of the generated utterances over the target sentences (i.e. score=Scenario Score - Target Sentences Politeness Score), across all scenarios.

Medium & Rude (MR) - Similar to **SR** but with utterances length between 16 and 32 tokens, totaling 138 utterances.

Medium & Neutral (MN) - Similar to **SN** but with utterances length between 16 and 32 tokens, resulting in a total of 2.2k utterances.

Table 3 shows the results of each model over *neutral* utterances, but of varying lengths. The RW-Fashion-C model achieved the highest content preservation scores on the short and medium utterance scenarios. It is interesting to point that in all scenarios, models trained on **MMD-A** (all except RW-Enron), showed the same pattern: they make more changes in shorter utterances and less in longer ones, producing very few changes or even leaving utterances unaltered in the latter case. The RW-Enron model showed the opposite trend, making significant changes in longer utterances.

With respect to style changes (Table 4 and Table 5), for every model, there was a clear difference between neutral scenarios (SN and MN) and their rude counterparts (SR and MR). On average, the rude scenarios scored 7% lower on content preservation metrics than the neutral tests. This result should not lead to the conclusion that models perform better on neutral data. Actually, after inspecting the results, we observed that models obtained higher scores in neutral utterances because they are less capable of identifying what needs to be replaced or added to improve politeness. This is supported by the politeness variation, shown in Table 5. Here we observe that all models produce a higher improvement in rude utterances, but the difference in the relative improvement on neutral utterances is small, meaning that the utterance would still fall on the rude split.

Regarding Politeness scores, the models trained on the **MMD-A** (**Fashion**, **Fashion-C** and **Mixed**) show significant improvement on the Enron test. However, after a second inspection of the generated utterances, it was evident that the Politeness score increase did not translate to tone improvements given the generation being of low quality. Namely, models simply add **MMD-A** excerpts with no apparent criteria.

Overall, these models perform better in short rude utterances. When dealing with neutral text, they tend to produce a lesser amount of changes meaning that for such models to be applied as part of a pipeline of a task-oriented dialog agent, it is important to perform fine-tuning, towards overcoming domain shift issues. We also observed that, based on common politeness strategies (see section A.1), most of the politeness strategies employed were Gratitude and Positive Lexicon, as is common in a costumer-store interaction, on the fashion domain.

5.2.2 Qualitative Analysis of Style Markers

In this section, we conduct a qualitative evaluation over a set of three utterances, in order to further pinpoint each rewriting variant's characteristics.

Table 6 illustrates 2 sample output utterances, and the resulting output of each RW variant. For the first example, we have a polite-sounding utterance where the ideal behavior would be to leave the utterance untouched, given that it is already in a very polite form. For this case, the RW-Mixed model produced a slightly improved form, making the utterance less generic and more fashion-related. The utterance generated by RW-Enron and RW-Fashion-C could have been a successful case had the correct semantics been applied to the added text. In the second example, we see mixed results.

	Hi, please tell me what I can help you with?
RW-Enron	Hi, please tell me what I can help you get together with?
RW-Fashion	Hi, please tell me what I can help you with? please note
RW-Fashion-C	Hi, please tell me what I can help you show with?
RW-Mixed	Hi, please tell me what I can help fit you with?
	Hello, what I can help you with today?
RW-Enron	Hello, what I can help you get together with today?
RW-Fashion	Hello, what I can help you move with today?
RW-Fashion-C	Hello, what I can help you get ready for today?
RW-Mixed	He can go well withrit, what I can help fit fit you with today?
	Great. I think that's a great choice.
RW-Enron	Great. I think that's could be a great choice.
RW-Fashion	Great. I do not think i have a great choice but would you like something in other types.
RW-Fashion-C	Great. I think that's a great choice.
RW-Mixed	Great. I think that's a great choice thank you for shopping with us.

Table 6: Politeness rewriting output utterances analysis. Changes made by the model are highlighted with red, meaning a negative change (Grammar error or Rude tone) occurred, yellow for neutral changes, and green for positive changes. lighter green indicates a positive change but less impactful than a darker green one.

Both the RW-Enron and RW-Fashion-C models were able to improve on the utterance's sentiment by adding in-domain knowledge. The RW-Mixed model produced a bad generation, adding duplicated words and low-quality excerpts.

Overall, under the correct circumstances, we see that most of the models can successfully improve politeness. The RW-Fashion makes mostly low-quality additions, showing that there is a need for style marker curation. We also observed that the models are often more successful when improving on an already polite utterance rather than when dealing with neutral utterances. We believe this behavior is a product of the model architecture that looks for style markers to replace and said style markers are not present in neutral-sounding text.

6 Conclusion

In this work, we address the research gap regarding the development of polite task-oriented dialog agents. We demonstrate that while politeness language constructs tend not to be domain-specific, their application is, requiring politeness approaches to cope with domain-specific vocabulary. Particularly, we show that when improving politeness in task-specific utterances, rewriting approaches consistently deliver better results, given that generative alternatives need to attend to two tasks.

In summary, the key takeaways are:

- Politeness through rewriting results in the most robust approach, providing a good balance between delivering polite utterances and preserving content.
- Politeness answer generation is less stable. By definition, generation and politeness improvement need to be addressed jointly, which is too ambitious in a domain-transfer setting.
- Bringing politeness to task-oriented dialog agents, characterized by operating over highly specific domains, is achievable with the proposed model domain adaptations.

As future work, we plan to extend our work and research methods that select the best politeness strategies while accounting for the specificity of distinct conversation phases (e.g. greeting vs. product description utterances).

Acknowledgments

This work has been partially funded by the iFetch project, Ref. 45920, co-financed by ERDF, COMPETE 2020, NORTE 2020 and FCT under CMU Portugal, by the FCT Ph.D. scholarship grant Ref. PRT/BD/152810/2021, and by the FCT project NOVA LINCS Ref. (UIDB/04516/2020).

References

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, Austin, Texas. Association for Computational Linguistics.

Francesca Bargiela-Chiappini. 2003. Face and politeness: new (insights) for old (concepts). *Journal of Pragmatics*, 35(10):1453–1469.

Penelope Brown, Stephen Levinson, and E Goody. 1978. *Universal in Language Usage: Politeness Phenomena*, volume 8, pages 56–311.

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Michael J. Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *WMT@EMNLP*.

Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Incorporating politeness across languages in customer care responses: Towards building a multi-lingual empathetic dialogue agent. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4172–4182, Marseille, France. European Language Resources Association.

Liye Fu, Susan Fussell, and Cristian Danescu-Niculescu-Mizil. 2020. Facilitating the communication of politeness through fine-grained paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5127–5140, Online. Association for Computational Linguistics.

Hitesh Golchha, M. Firdaus, Asif Ekbal, and P. Bhattacharyya. 2019. Courteously yours: Inducing courteous behavior in customer care responses using reinforced pointer generator network. In *NAACL-HLT*.

Bryan Klimt and Yiming Yang. 2004. Introducing the enron corpus. In *CEAS*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Ardith J. Meier. 1995. Defining politeness: Universality in appropriateness. *Language Sciences*, 17:345–356.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

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

Amrita Saha, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 696–704. AAAI Press.

Ricardo Gamelas Sousa, Pedro Miguel Ferreira, Pedro Moreira Costa, Pedro Azevedo, Joao Paulo Costeira, Carlos Santiago, Joao Magalhaes, David Semedo, Rafael Ferreira, Alexander I. Rudnicky, and Alexander Georg Hauptmann. 2021. *IFetch: Multimodal Conversational Agents for the Online Fashion Marketplace*, page 25–26. Association for Computing Machinery, New York, NY, USA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Richard J. Watts. 2019. *2. Linguistic politeness and politic verbal behaviour: Reconsidering claims for universality*, pages 43–70. De Gruyter Mouton.

Politeness Strategies
Using Honorifics: *How can I help you Ms Smith?* **Description:** Through the usage of Honorifics, the speaker conveys respect towards another person.
Gratitude: ***Thank you very much*** *for your help.* **Description:** One of the easiest ways of sounding polite is through the expression of gratitude.
Tag Questions: *Take a seat,* ***won't you?*** **Description:** Tag Questions at the end of a utterance can define the tone used, making it sound more friendly or the opposite, depending on the usage.
Positiveness: ***Great choice!*** *That's a* ***fantastic*** *watch.* **Description:** Using a positive lexicon keeps the utterance tone positive and conveys that same feeling to the listener.
Greeting: ***Hello, welcome!*** *Can I interest you in some ...* **Description:** A greeting can help convey a polite and respectful tone to the other interlocutor.

Table 7: Example strategies to improve utterances politeness, with style markers highlighted in blue.

A Appendix

A.1 Politeness Strategies

Table 7 shows some of the Politeness Strategies that can be identified when dealing with polite dialog.

A.2 Dataset details

Stanford Politeness Corpus (SPC) - Dataset used for politeness conditioning, by training the Politeness proxy classifier (Niu and Bansal, 2018). This corpus is composed of requests made by editors in Wikipedia and by requests made on Sack Exchange, all of which have been annotated by 5 humans. For Wikipedia and Stack Exchange requests, 4,353 out of 35,661 and 6,603 out of 373,519 were annotated, respectively. Request scores were z-score normalized and averaged. We used the data splits used originally with Polite-RL (Niu and Bansal, 2018), so we only considered the top and bottom 25% utterances for polite and rude respectively. 3808 utterances were used for training and 1056 for testing.

Enron - This dataset is a collection of emails exchanged in the Enron company (Klimt and Yang, 2004), that was originally used to train the Tag-and-Generate model (Madaan et al., 2020). We consider the subset of the Enron dataset that the authors automatically annotated using the Politeness Classifier (Niu and Bansal, 2018). This dataset is used in our work to establish an initial domain for a task-oriented dialog agent. For training, 212k polite and 51k rude sentences are considered, for validation 27k polite and 5.8k rude, and for testing 26k polite and 5.8k rude utterances.

Multimodal Dialogs Dataset (MMD) - MMD (Saha et al., 2018) comprises multi-turn multimodal dialogs for the fashion domain. MMD is used as use-case for a second domain task-oriented dialog agent, where we define two distinct (but overlapping) subsets: **MMD-R** and **MMD-A**. In **MMD-R**, based on the provided intent annotations, we only keep system utterances corresponding to product(s) recommendation(s), resulting in 380k/81k/81k utterances for training/validation/testing. The goal of this first subset is to expose the model to the domain-specific product lexicon of the fashion domain. Namely, these utterances comprise scenarios in which the system recommends and describes one or more products to the user. In **MMD-A** we keep all *neutral* and *polite* system utterances, with more than 5 tokens, totaling 39k/10k/10k and 414k/96k/96k, neutral and polite utterances, respectively, for training/validation/testing. Here we considered a style change from neutral to polite, rather than rude to polite, since the number of rude utterances is minimal (2.5k).

Model Name (b, β)	**BLEU**	**Politeness**
Reference(96, 2.0)	**66.30**	64.47
Model 1(32, 2.0)	49.64	72.40
Model 2 (128, 2.0)	63.60	60.24
Model 3 (96, 5.0)	33.56	75.21
Model 4 (96, 10.0)	29.41	**78.77**

Table 8: Experimental results of the 4 tested scenarios vs a reference model. From now on, we scale up the politeness scores into a 0 to 100 scale.

A.3 MMD Sample dialog

A sample dialog from the MMD (Saha et al., 2018) dataset can be found in Figure 2.

A.4 Polite-RL Model tuning

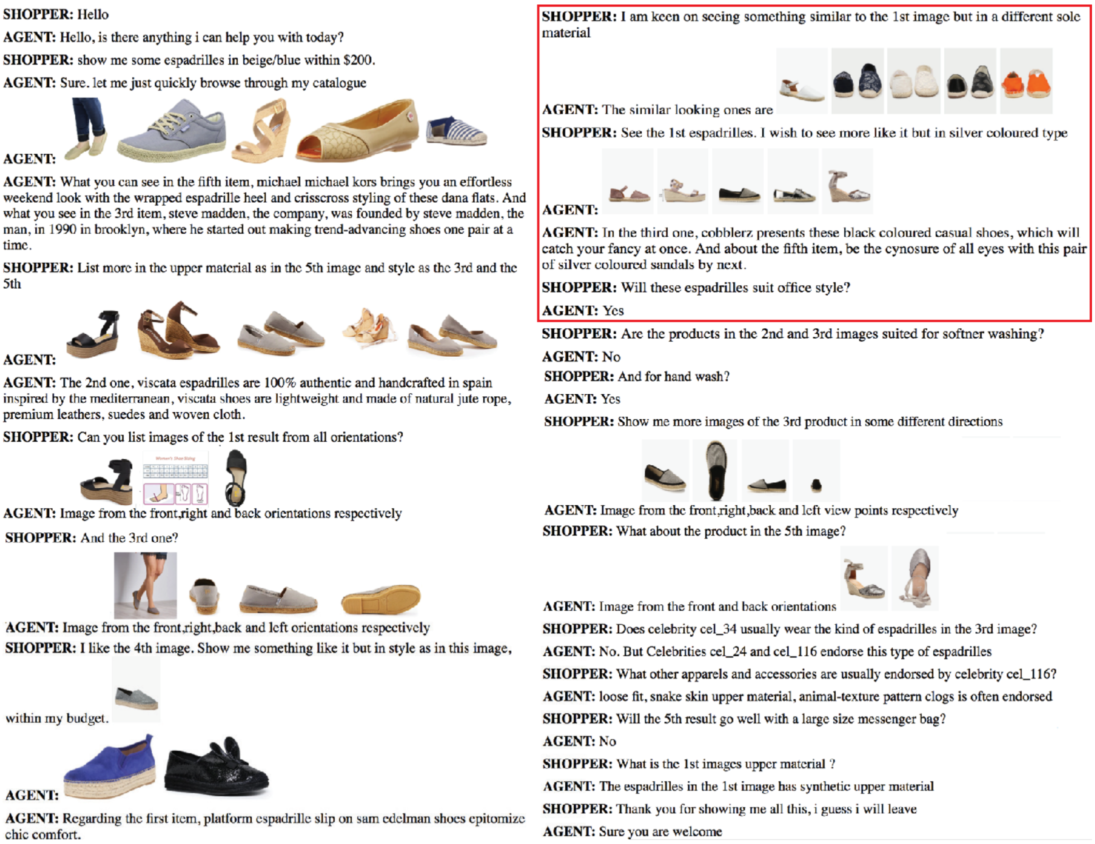
For the model parameter tuning, the two parameters, β and batch size (b), were tested separately, and, for each parameter, we tested 2 variations of their values. To measure the impact on the results, we use BLEU and the politeness score on the test set. As for baselines, we use a version of the model trained on the MMD data with the default values for each parameter.

$$L = L_{ML} + \beta L_{RL} \tag{1}$$

For the batch size (whose default value was 96), we tested the model with sizes 128 and 32, these values were picked to understand the model's behavior with an increase and decrease of the value. The β is an hyperparameter that dictates the weight given to the politeness reward component of the model's loss function, as shown in Equation 1 where L_{ML} is the maximum likelihood loss and L_{RL} is the politeness reward loss. For this parameter, we followed a different direction and tested with values 5 and 10, both significantly bigger than the default value of 2. This was done to understand the impact of the parameter in the politeness of the generated text and how it impacted generation quality. This is an important factor given that, for conversational agents, it is important to generate polite text but also retain high-quality question answering capabilities.

The results, shown in Table 8, are using the Classifier with custom embeddings. These results show that altering both parameters can lead to a noticeable change in the model's performance. Looking at the BLEU scores, none of the tested variations beat the base model, with only Model 2 coming close. For the two models where we changed the beta value, Model 3 and Model 4, the BLEU score took a nosedive, which was expected since by increasing the β we are changing the initial balance in the loss function making it highly favor polite generation over accurate question answering.

When looking at the inference results from both models, we see significant text degeneration on a large portion of the test sentences, with the same pattern repeating: the first dozen or so tokens are correctly predicted followed by a dozen or



SHOPPER: Hello

AGENT: Hello, is there anything i can help you with today?

SHOPPER: show me some espadrilles in beige/blue within $200.

AGENT: Sure. let me just quickly browse through my catalogue

AGENT:

AGENT: What you can see in the fifth item, michael michael kors brings you an effortless weekend look with the wrapped espadrille heel and crisscross styling of these dana flats. And what you see in the 3rd item, steve madden, the company, was founded by steve madden, the man, in 1990 in brooklyn, where he started out making trend-advancing shoes one pair at a time.

SHOPPER: List more in the upper material as in the 5th image and style as the 3rd and the 5th

AGENT:

AGENT: The 2nd one, viscata espadrilles are 100% authentic and handcrafted in spain inspired by the mediterranean, viscata shoes are lightweight and made of natural jute rope, premium leathers, suedes and woven cloth.

SHOPPER: Can you list images of the 1st result from all orientations?

AGENT: Image from the front,right and back orientations respectively

SHOPPER: And the 3rd one?

AGENT: Image from the front,right,back and left orientations respectively

SHOPPER: I like the 4th image. Show me something like it but in style as in this image, within my budget.

AGENT:

AGENT: Regarding the first item, platform espadrille slip on sam edelman shoes epitomize chic comfort.

SHOPPER: I am keen on seeing something similar to the 1st image but in a different sole material

AGENT: The similar looking ones are

SHOPPER: See the 1st espadrilles. I wish to see more like it but in silver coloured type

AGENT:

AGENT: In the third one, cobblerz presents these black coloured casual shoes, which will catch your fancy at once. And about the fifth item, be the cynosure of all eyes with this pair of silver coloured sandals by next.

SHOPPER: Will these espadrilles suit office style?

AGENT: Yes

SHOPPER: Are the products in the 2nd and 3rd images suited for softner washing?

AGENT: No

SHOPPER: And for hand wash?

AGENT: Yes

SHOPPER: Show me more images of the 3rd product in some different directions

AGENT: Image from the front,right,back and left view points respectively

SHOPPER: What about the product in the 5th image?

AGENT: Image from the front and back orientations

SHOPPER: Does celebrity cel_34 usually wear the kind of espadrilles in the 3rd image?

AGENT: No. But Celebrities cel_24 and cel_116 endorse this type of espadrilles

SHOPPER: What other apparels and accessories are usually endorsed by celebrity cel_116?

AGENT: loose fit, snake skin upper material, animal-texture pattern clogs is often endorsed

SHOPPER: Will the 5th result go well with a large size messenger bag?

AGENT: No

SHOPPER: What is the 1st images upper material ?

AGENT: The espadrilles in the 1st image has synthetic upper material

SHOPPER: Thank you for showing me all this, i guess i will leave

AGENT: Sure you are welcome

Figure 2: A sample dialog from the MMD dataset. Source: Saha et al. (2018)

more repetitions of the same token, a token favored by the politeness classifier. Still considering the BLEU scores, Model 1 presented some surprising results as we were not expecting text degeneration to also occur in this scenario, but it did, albeit not as accentuated as in the later 2 models.

When taking into account the politeness scores, we see that increasing the beta value clearly improves politeness generation, or so it would seem, as mentioned before all of the models that beat the reference in politeness generation did it by starting to generate the same token repeatedly mid-sentence.

These results showed us that when trying to encourage politeness generation, we cannot solely rely on token probability distributions, semantics need to be taken into account or, at least, vocabulary diversity at the classifier level, since any other way the model is not punished by simply outputting the classifier's favorite word, 'belt' in some cases and 'republic' in others. This means that, using the Polite-RL, the best balance that can be achieved results is a compromise in generation quality while not making the used tone polite. For conversational agents, this is important as the question answering quality needs to remain high throughout the conversation.