Intention and Face in Dialog

💻🏛Adil Soubki, 🗣🏛Owen Rambow
💻Department of Computer Science, 🗣Department of Linguistics
🏛Institute for Advanced Computational Science, Stony Brook University
asoubki@cs.stonybrook.edu, owen.rambow@stonybrook.edu

Abstract

The notion of face described by Brown and Levinson (1987) has been studied in great detail, but a critical aspect of the framework, that which focuses on how intentions mediate the planning of turns which impose upon face, has received far less attention. We present an analysis of three computational systems trained for classifying both intention and politeness, focusing on how the former influences the latter. In politeness theory, agents attend to the desire to have their wants appreciated (positive face), and a complementary desire to act unimpeded and maintain freedom (negative face). Similar to speech acts, utterances can perform so-called face acts which can either raise or threaten the positive or negative face of the speaker or hearer. We begin by using an existing corpus to train a model which classifies face acts, achieving a new SoTA in the process. We then observe that every face act has an underlying intention that motivates it and perform additional experiments integrating dialog act annotations to provide these intentions by proxy. Our analysis finds that dialog acts improve performance on face act detection for minority classes and points to a close relationship between aspects of face and intent.

Keywords: Discourse Annotation, Representation and Processing, Cognitive Methods, Dialogue

1. Introduction

Brown and Levinson (1987) introduce an influential theory of politeness based on the concept of “face”, which they claim to be culturally universal. In this theory, face – i.e., the public image one seeks to claim – is a two-sided coin. Agents attend to their desire to have their wants appreciated, which the theory calls positive face, as well as a complementary desire to act unimpeded and maintain freedom, which the theory calls negative face. The face of every agent is ensnared with that of every other agent – agents cannot have their desires appreciated if they cannot appreciate the desires of others. As a result, utterances can raise (+) or threaten (-) the positive (Pos) or negative (Neg) face of the speaker (S) or hearer (H).

For example, *pass the salt* would be labeled HNeg- as it imposes on the hearer’s freedom by requesting their action, while *that was my fault*, a self-critique, would be considered SPos-. A summary of possible face acts is shown in Table 1.

A face threat or face raising is not a property of particular linguistic choices, but of communicative intent. If I want to request information from you, then I necessarily need to threaten your negative face: if you recognize the speech act (and thus it is successful), I will oblige you to answer, and therefore I will restrict your choice of actions. In Brown and Levinson (1987)’s theory, discourse participants choose among various strategies for minimizing threats to face. These strategies are linguistic strategies (for example, using hedges), and the choice of strategy depends on many factors such as the discourse situation (who is talking to whom under what circumstances) and cultural conventions.

Work related to natural language processing has concentrated on studying linguistic manifestations of politeness (Walker et al., 1997; Danescu-Niculescu-Mizil et al., 2013), while largely disregarding the notion of face act (FA). The major exception is the seminal work of Dutt et al. (2020), who annotate a corpus of written dialogs for face acts, and then develop a system for predicting face acts for dialog turns. In this paper, we build on this work. Our goal is not primarily to improve on the tagging results (which we do), but to understand better how face acts interact with discourse intention, and to pave the way for a new phase of work on face acts. We believe face acts, as conceived of by Brown and Levinson (1987), are not a peripheral aspect of NLP, but can serve a crucial role in improving both interactive NLP and our understanding of how we humans use language. There are two reasons to make such a strong claim. First, in NLP it has been observed that, despite their great success in many engineering problems, large language models (LLMs) are typically pre-trained entirely on sequences of words and do not model intention (Bender et al., 2021; Bender and Koller, 2020). However, communicative intention and intention recognition is the fundamental mechanism of communication. Second, while much progress has been made (again partially thanks to LLMs) in processing multiple languages, there has not been much work in NLP that addresses the culture-specific ways in which language is used in context. We believe the study of face acts can address both issues. This paper is a first step in the direction of

LREC-COLING 2024, pages 9143–9153
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Face Act	Interpretation	Example Discourse Goals
HNEG-	Imposition	Requests, commands, questions, offers, promises, ...
HPOS-	Disagreement	Criticism, insults, disapproval, ...
HNEG+	Permissiveness	Granting permission, making exceptions, ...
HPOS+	Agreement	Seeking common ground, group cohesion, ...
SNEG-	Indebtedness	Thanking, accepting offers or thanks, commitments, ...
SPOS-	Apologies	Confessions, embarrassment, ...
SNEG+	Autonomy	Refusing requests, asserting freedoms, ...
SPOS+	Confidence	Self-promotion, signaling virtue, ...

Table 1: Face acts with a short label, which serves as an general interpretation of the face act, and some examples of their related discourse goals.

a larger research project.

The principal contribution of this paper lies in a set of experiments in which we train FA taggers using information from dialog act taggers. We show in an extensive analysis how this helps for specific FAs, which are hard to tag without such information. Furthermore, we provide a relatively straightforward application of generative neural techniques to the FA tagging problem, and we obtain a new state-of-the-art (increasing the state-of-the-art from 69% to 73% F-measure).

This paper is structured as follows. We start with a review of relevant literature (§2) and an outline of our approach to modeling and evaluation (§3). We then discuss our experiments using only face act information (§4) as well as follow-up experiments which integrate intention through the use of dialog acts (§5). We then conduct an extensive error analysis of all system variants (§6) and report our conclusions along with a discussion of future work for this research program (§7).

2. Related Work

Brown and Levinson (1987) provide a theory of politeness which has been fundamental to work in various fields concerned with how language is used. We have provided a brief summary in Section 1. Curiously, in NLP there has not been much work building explicitly on (Brown and Levinson, 1987). Danescu-Niculescu-Mizil et al. (2013) concentrate on one type of face-threatening act (FTA), namely the negative face-threatening act of a request, and investigate the strategies used for doing this FTA. To do this, they use crowd sourcing to rate the requests on a politeness scale. They then develop a model which predicts the politeness of these requests and use it to study how this affects the interactions between users on Wikipedia and StackExchange.

The face acts (FAs) themselves are the object of Dutt et al. (2020). In addition to developing a data set annotated with FAs, they present a FA classifier based on a neural architecture they devise on top of BERT, which achieves 69% F-measure (60% macro). As the data involves participants convincing others to donate to a charity, they also use this corpus to investigate the relationship between face acts and persuasion by predicting if a participant chose to donate. We use this data set in our work on FA tagging.

A salient aspect of the work of Brown and Levinson (1987) is that they situate the notion of politeness within a larger theory of rational interaction, as outlined by Grice (1975). One consequence of this is that successful communication requires the recognition of intent: a speaker's request cannot threaten the hearer's negative face if the hearer does not recognize the speech act as a request. There is a large body of work on speech acts and intent, starting with Austin (1962). We do not provide a summary of all relevant work here, but the notion of speech act was extended in NLP as a dialog act, and given several fine-grained inventories that go well beyond the initial high-level distinctions of speech acts (Anderson et al., 1991; Core and Allen, 1997; Stolcke et al., 2000a). The corresponding classification task, dialog act tagging, is a crucial component in creating dialog systems, as it allows for a simple way of modeling the user's communicative intent through text classification. In this paper, we do not make contributions to dialog act tagging, but we use existing work.

3. Approach

In this section we outline the data, modeling techniques, and evaluation measures used throughout the paper.

3.1. Face Act Data

As discussed in Section 2, we use the face data set developed by Dutt et al. (2020) for our experiments. Wang et al. (2019) introduce a corpus of dyadic, persuasion-oriented conversations sourced from an online task where Amazon Mechanical Turk workers must convince their addressee to donate part of their task earnings to a charity, Save the Children. The conversations are carried out through a chat interface with one worker acting as the persuader (ER) and the other as the persuadee (EE). The participants were informed that the dialog must last at least 10 turns and that their reward is not penalized should they fail to convince their partner to donate. Dutt et al. (2020) augment conversations from this corpus with eight face act annotations (see Table 2) based on Brown and Levinson (1987). They take some small departures from politeness theory in their annotation. Most notably, thanking is annotated as HPos+ rather than SNeg- and Other is used to indicate that no face act is present. The authors selected the corpus as their starting point for two main reasons. First, the goal-oriented nature of the conversations incentivizes face-threatening acts, which are typically avoided unless necessary. Second, each participant is on equal ground which helps mitigates potentially confounding issues, like power distance.

It is possible for a single utterance to perform multiple face acts at once. For example, *Just stick to what you know* could be seen as both HNeg-, since it is issuing a command, and HPos-, as it entails the critique that they did not know what they were doing. However, Dutt et al. (2020) observed multi-labeled acts in only 2% of their data leading them to simplify the the problem to a single label per utterance. In the event of a multi-label annotation, they select one randomly. The resulting data set contains 10,716 turns averaging 10 words (or 51 characters) in length across 296 unique conversations. The label distribution (see Table 2) is highly imbalanced with vanishingly rare labels like SPos+ appearing only 12 times and labels like SNeg- actually vanishing.

3.2. Modeling

We model face act tagging as a text classification task. Given a sequence of n tokens $S = [t_1, t_2, \ldots, t_n]$, we wish to assign a label $y \in Y$ where Y represents the set containing the 8 possible face acts. Recently, similar classification tasks including sentiment analysis (Zhang et al., 2021) and event factuality prediction (Murzaku et al., 2022) have achieved state-of-the-art results training sequence-to-sequence models. We adopt this approach and utilize Flan-T5-base (Chung et al., 2022) for fine-tuning which allows us to re-frame the problem as a generation task with limited model-specific requirements.[1]

3.3. Data Representation

While generative approaches unify many aspects of the model design, they present challenges when it comes to determining effective input and output representations. We provide the models an input which contains an utterance prefixed with ER for persuaders or EE for persuadees along with up to two previous turns of dialog as context. Each turn is separated by a newline character which we treat as a special token when tokenizing.

```
[Input]
  ER: Are you interested in donating?
  EE: Possibly, I'm not sure.
  EE: I don't even know what the char-
  ity is.
[Output]
  sneg+
```

The target output is a string containing the correct label for the final utterance of the input text, in this case SNeg+ since the speaker is asserting their freedom of action. In preliminary experiments we found context to be a critical factor with a size of two, for a total of three utterances, performing best. As there are no previous turns for the first two turns in each dialog, those examples are provided in a similar format containing only one or no lines of context.

3.4. Evaluation

We evaluate model performance using F-measure for each of the eight represented classes as well as micro and macro F-measure aggregated over all labels. Since each utterance is assigned exactly one label, micro F-measure is the same as accuracy. We observe that the extreme rarity of SPos+ (12 occurrences) contributes to high variance in macro F1 and, as a result, advocate focusing on micro F1 and macro F1 with this label removed as the primary high-level metrics for this task. So as to maintain comparability with previous work, we report values for macro F-measure computed with all represented labels. Our experiments are performed using five-fold cross-validation on the same splits which Dutt et al. (2020) report their findings on.[2] We identified some conversations which were duplicated across folds and keep only the first appearance of these entries when performing evaluation. The evaluation metrics are averaged across all five folds.

Ideally, the output generated by the model would contain only sequences in our label set, but there were a small number of cases in which malformed output was produced. We repair these labels using a methodology inspired by (Zhang et al., 2021).

[1] Our choice of Flan-T5 was informed by a preliminary set of small experiments in which a variety of pre-trained models were were examined on single seed runs.

[2] https://github.com/xinru1414/Face_Act

	F1	F1	Prec.	Recall	Count
Macro	0.60	0.63	0.63	0.63	-
Micro	0.69	0.73	0.73	0.73	-
Other	-	0.75	0.76	0.73	4,300
HPos+	-	0.75	0.72	0.77	2,844
SPos+	-	0.74	0.74	0.75	1,589
HNeg-	-	0.74	0.71	0.76	1,073
HPos-	-	0.55	0.61	0.51	334
HNeg+	-	0.44	0.47	0.41	305
SNeg+	-	0.57	0.61	0.53	259
SPos-	-	0.47	0.39	0.58	12
Dutt et al. (2020)			Fos		

Table 2: Performance of our Fos against the previous state-of-the-art BERT HiGRU-f model and label count.

Invalid output sequences are compared with all possible labels using edit-distance and the closest match is used. In the event of a tie, the most frequent label in the training data is chosen.

3.5. Replication

All of the code written, data sets prepared, and experimental observations made in the course of this research are available on GitHub.[3]

4. Face-Only System

We begin by training the model with no additional information containing intentions. We will refer to this configuration as the Fos (Face-Only System).

Methodology We fine-tune Flan-T5-base on each of the five cross-validation folds with a batch size of 32 and two gradient accumulation steps. The AdamW optimizer is configured with a learning rate of 3e-4, weight decay of 0.01, and epsilon of 1e-8. As the cross-validation preparation does not contain a development set, early stopping is configured using micro F1 on the training data set with a patience of 3. In general, fine-tuning completed after roughly 15 to 20 epochs. These parameters were arrived at through a small run of hyperparameter tuning experiments. When predicting, generation is performed with a beam size of one.

Results Evaluation metrics, averaged across all folds, for the Fos are summarized in Table 2. This relatively straight-forward approach achieves a macro F1 of 0.63 and micro F1 of 0.73, improving on the previous state-of-the-art by 3 and 4 points, respectively. Among the labels, F1 correlates strongly ($r = 0.77$) with the number of examples found in the data. In other words, the minority classes are where this model struggles to find signal.

5. Adding Intention

We observe that every face act has an underlying intention which motivates it. A rational agent would not risk reprisals that result from threatening their interlocutor's face unless (1) they have a goal or intention which necessitates the face act or (2) other cultural factors such as power-distance protect them from such reprisals. As (2) is not the case in our corpus, we have reason to believe that providing the model with explicit information regarding agent intentions may improve performance.

5.1. Data

We represent information regarding intent using two well-known dialog act corpora with varying levels of granularity in distinguishing actions.

Meeting Recorder Dialog Act Corpus MRDA (Shriberg et al., 2004) consists of transcripts from 75 in-person meetings that are generally research oriented in nature, annotated for dialog acts using a tag set adapted from DAMSL (Core and Allen, 1997). It contains 108,202 utterances with a mean utterance length of 8 words. The annotations are provided in three levels of granularity. The "basic" level contains 6 tags, the "general" level contains 12 tags, and the full label set contains 52 tags. We utilize the basic tag set for our experiments.

Switchboard Dialog Act Corpus SWDA is a collection of short phone conversations in which callers are matched with a partner to discuss some provided general-interest topic (Stolcke et al., 2000b). It contains roughly 180,000 utterances with a mean length of 9.6 words across 1,155 unique conversations. They use the DAMSL dialog act tag set to annotate the transcripts with 41 dialog act labels.

5.2. Methodology

We experiment with two methods of integrating dialog act information into the model. In the first method, we augment the face act corpus with explicit dialog act tags in the text. For both MRDA and SWDA, we use the dialog act tagging system of He et al. (2021) to automatically augment the face act corpus and produce two new preparations of data for training. We will refer to the resulting text-augmented models as Ta-Mrda and Ta-Swda, respectively. Returning to the example in Section 3.3,

[3] https://github.com/cogstates/2024-lrec-coling-faceacts

	SPos+			HPos-			SNeg+			HNeg+			HNeg-		
	F1‡	Prec.†	Recall	F1	Prec.	Recall	F1	Prec.	Recall†	F1‡	Prec.†	Recall‡	F1	Prec.	Recall†
Fos	0.74	0.74	0.75	0.55	0.61	0.51	0.57	0.61	0.53	0.44	0.47	0.41	0.74	0.71	0.76
Ta-Swda	0.70	0.68	0.72	0.53	0.56	0.51	0.48	0.55	0.43	0.49	0.44	0.56	0.72	0.69	0.75
Ta-Mrda	0.70	0.70	0.69	0.53	0.58	0.49	0.51	0.59	0.45	0.51	0.47	0.55	0.72	0.71	0.73
Mtl-Swda	0.72	0.74	0.70	0.55	0.56	0.54	0.54	0.57	0.52	0.41	0.46	0.37	0.72	0.65	0.79
Mtl-Mrda	0.72	0.71	0.73	0.49	0.49	0.50	0.53	0.62	0.47	0.43	0.50	0.39	0.73	0.70	0.78

Table 3: Detailed evaluation results for all experiments on the five least common labels, excluding SPos-. Significant differences using the Friedman rank sum test are marked with † and ‡ for $\alpha = 0.10$ and 0.05 respectively.

	F1		Precision		Recall	
	Micro	Macro	Micro	Macro	Micro	Macro
Fos	0.73	0.63	0.73	0.63	0.73	0.63
Ta-Swda	0.70	0.61	0.70	0.59	0.70	0.63
Ta-Mrda	0.70	0.60	0.70	0.60	0.70	0.60
Mtl-Swda	0.70	0.57	0.70	0.58	0.70	0.57
Mtl-Mrda	0.71	0.60	0.71	0.61	0.71	0.59
Dutt et al.	0.69	0.60	-	-	-	-

Table 4: Summary of model results.

the augmented input using SWDA tags is shown below.

```
[Input]
  ER: Are you interested in donating?
  (Yes-No-Question)
  EE: Possibly, I'm not sure.
  (Hedge)
  EE: I don't even know what the char-
  ity is.  (Statement-non-opinion)
[Output]
  sneg+
```

In the second method, we incorporate the dialog act data through traditional multi-task learning. As the dialog act data sets contain far more examples than the face act data set, we randomly sample 10% of conversations from both SWDA and MRDA. This results in training regimens with roughly 1:1 and 2:1 ratios of dialog acts to face acts, respectively. To indicate the desired task, each example input is prefixed accordingly with the task name (*dialog acts* or *face acts*) followed by a colon and new line. The data for both tasks is provided with the same three total turns of context as used for the Fos. We will refer to the resulting models as Mtl-Mrda and Mtl-Swda.

These experiments use the same training configuration, cross-validation folds, and hardware as we used in the case of Fos.

5.3. Results

The evaluation results for each of these model variants are summarized in Table 4. While models incorporating dialog act data generally outperform the baseline, they do not improve upon the Fos and, in fact, hamper performance across the board for these high-level measures. At first glance, this is a puzzling negative result. However, inspecting the performance on minority classes (see Table 3) we find several instances where SWDA, the more granular of the two, improves recall while MRDA, which uses a coarser tag set, improves precision.

To make sense of these distinctions we employ a set of non-parametric statistical tests. This minimizes the assumptions made about our data, as no distribution is required, while still allowing us to assign rankings to the variables under question. Holding the training regimen as a treatment variable we utilize the Friedman rank sum test to determine if this contributes significantly to predicting model F1, precision, and recall by considering the null hypothesis that all variations contribute equally. This, along with Kendell's W, identifies where the models are performing differently from each other and the effect size of that difference, respectively. Significant differences on the minority classes can be seen in Table 3. According to Cohen's interpretation guidelines for Kendall's W (0.1: small effect, 0.3: moderate effect, > 0.5: large effect) we find moderate to large effects in all metrics with significant differences. This confirms our suspicion that the dialog acts were indeed helping for minority labels.

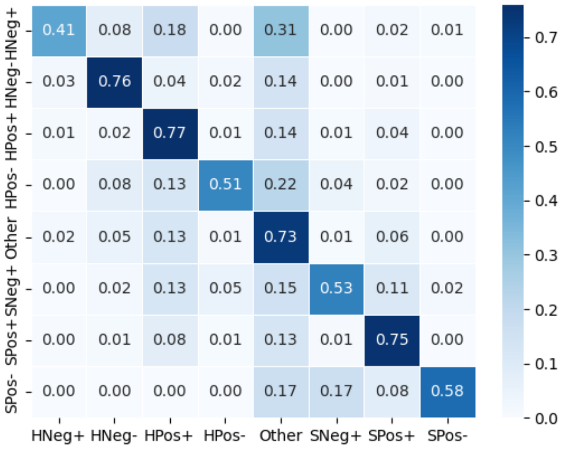

Figure 1: Confusion matrix for the Fos; the numbers show the fraction of time the tag on the x-axis is predicted instead of the gold label on the y-axis; the numbers in each row add up to 1.

	HNeg+	HNeg-	HPos+	HPos-	Other	SNeg+	SPos+	SPos-	Total
Both Happening (Same Part)	5	2	3	7	0	7	5	0	29
Both Happening (Diff. Part)	2	3	2	5	0	1	2	2	17
Gold Error (Correct)	3	0	3	0	18	4	5	0	33
Gold Error (Incorrect)	1	0	2	1	2	0	1	1	8
True for Previous	3	4	1	2	3	2	2	1	18
Predicted Other	10	13	11	5	0	8	7	1	55
No Idea	1	3	3	5	2	3	3	0	20
Total	25	25	25	25	25	25	25	5	180

Table 5: Fos error counts by gold label and error category.

This raises the question of what exactly the dialog act data was providing to the model and why did it succeed in some cases but fail to improve overall performance. In the subsequent sections, we perform an extensive error analysis which carefully examines our data and results to better disentangle the possible causes.

6. Error Analysis and Discussion

Our error analysis seeks to answer three questions. (1) What is the Fos doing? (2) Is there signal in the dialog act data being incorporated and if there is, (3) are our systems that are trained with dialog act information using that signal? We start by making some high level observations (§6.1), then perform a manual analysis of the output produced by Fos (§6.2), and finally discuss the contribution dialog act tags were making by comparing output between the systems (§6.3).

6.1. Basic Observations

As noted in Section 4 and shown in Table 2, Fos performance is highly correlated ($r = 0.77$) with the frequency of a tag. For the four tags that occur more than 1,000 times, we obtain 73% F-measure or more, and for three tags that occur between 250 and 350 times, we obtain F-measures below 56%. Using this information we divide the labels into three classes: The majority classes (Other, HPos+, SPos+, HNeg-), the minority classes (HPos-, HNeg+, SNeg+), and the extremely rare (SPos-). As discussed in Section 3.4, we disregard SPos- in our analysis as its infrequency causes highly volatile results.

The confusion matrix for Fos (see Figure 1) shows that, like in previous work (Dutt et al., 2020), the majority of misclassifications occur when no face act is predicted to be present (the Other column) though one is indeed there. As shown in Table 2, Other is the most frequent label in our corpus. The next largest source of error is HPos+, which is the second most common label in the data set. We conclude that the model is, to some extent, probability matching. Furthermore, the confusion matrix in Figure 1 reveals that the system may struggle with distinguishing positive and negative face, with HNeg+ being mistaken for HPos+ 18% of the time, and SNeg+ being mistaken for SPos+ 11% of the time.

6.2. Error Analysis on the Fos

We perform manual analysis of the output produced by our best performing system (Fos) by collecting 25 misclassified examples for each label with five randomly selected from each test fold for a total of up to 200 output-prediction pairs. As every test fold does not necessarily contain five incorrect predictions for SPos-, this produced 180 examples for study.

We sort these examples into the following seven error categories and consult the annotation guidelines and criteria to make the appropriate determinations.

1. **Both Happening (Same Part):** The predicted face act is also happening for that utterance in the same span of text (two face acts at once).
2. **Both Happening (Diff. Part):** The predicted face act is also happening for that utterance, but in a different span of the text.
3. **Gold Error (Correct):** The reference face act label is incorrect and the predicted face act label is correct.
4. **Gold Error (Incorrect):** The reference face act label is incorrect and the predicted face act label is also incorrect.
5. **True for Previous:** The predicted face act occurs for a previous utterance in the context window.
6. **Predicted Other:** None of the previous error categories apply and the system predicted Other.
7. **No Idea:** None of the previous error categories apply and we could not determine a specific reason for this (errorful) prediction.

The results of this analysis are summarized in Table 5. Overall, we find a gold error in 22.7% of the examples examined with 18.3% absolute of these errors being identified as instances where the sys-

	SPos+	HPos+	SPos-	HPos-	SNeg+	HNeg+	HNeg-	Other
BackChannel	-0.02	0.01	-0.00	-0.01	-0.01	-0.01	-0.01	0.02
Disruption	0.03	0.01	-0.00	-0.00	0.02	-0.01	-0.04	-0.00
FloorGrabber	-0.01	-0.02	-0.00	-0.01	-0.01	-0.01	-0.01	0.04
Question	-0.19	-0.24	-0.02	0.02	-0.07	-0.05	0.49	0.09
Statement	0.14	0.20	0.02	-0.02	0.04	0.05	-0.38	-0.08
Acknowledge (Backchannel)	-0.03	0.00	-0.00	-0.02	-0.01	-0.00	-0.03	0.05
Action-directive	-0.01	-0.01	-0.00	0.00	-0.01	-0.01	0.04	-0.00
Affirmative non-yes answers	0.00	0.03	-0.00	-0.01	-0.01	-0.01	-0.01	-0.02
Agree/Accept	-0.02	0.15	-0.00	-0.02	-0.02	-0.02	-0.04	-0.07
Appreciation	-0.12	0.29	0.01	-0.05	-0.05	-0.05	-0.10	-0.06
Backchannel in question form	-0.01	-0.01	-0.00	-0.01	-0.01	-0.01	-0.00	0.03
Conventional-closing	-0.09	0.06	0.01	-0.04	-0.03	-0.03	-0.07	0.09
Conventional-opening	-0.01	-0.02	-0.00	-0.01	-0.00	-0.00	-0.01	0.04
Hedge	-0.01	-0.01	-0.00	-0.00	0.02	-0.00	-0.01	0.01
Hold before answer/agreement	-0.02	-0.01	-0.00	-0.01	-0.01	-0.01	-0.02	0.04
No answers	-0.02	0.00	-0.00	0.03	0.01	-0.01	-0.02	0.01
Non-verbal	-0.06	0.00	0.03	-0.02	0.02	0.00	-0.03	0.05
Other	-0.03	-0.05	-0.00	-0.01	-0.01	-0.01	-0.03	0.10
Other answers	-0.00	-0.01	-0.00	-0.00	-0.00	-0.00	-0.00	0.01
Response Acknowledgement	-0.01	-0.00	-0.00	-0.01	-0.01	-0.01	-0.01	0.02
Segment	0.07	0.01	-0.01	-0.00	0.02	0.02	-0.05	-0.04
Statement-non-opinion	0.30	-0.10	0.01	0.01	0.12	0.08	-0.21	-0.07
Statement-opinion	-0.03	0.13	-0.01	0.05	-0.03	0.02	-0.10	-0.05
Wh-Question	-0.14	-0.18	-0.01	0.02	-0.05	-0.05	0.35	0.07
Yes answers	0.01	0.01	-0.00	-0.01	-0.00	-0.01	-0.03	0.01
Yes-No-Question	-0.12	-0.16	-0.01	0.01	-0.05	-0.03	0.31	0.06

Table 6: The correlation coefficients for the dialog act labels produced by He et al. (2021)'s system for MRDA (top) and SWDA (bottom). These were incorporated into the training data for Ta-Mrda and Ta-Swda (§5.2).

tem output was correct. In another 25.5% of cases we identified criteria in the annotation guidelines for both the gold label *and* the predicted label (meaning that both labels are correct), with the "same" and "different" part categories occurring with roughly equal frequency. This is in stark contrast to the reported 2% frequency of multi-label annotations observed by Dutt et al. (2020). 10% of the time the system produced a prediction which is correct for a previous utterance in the context. This points to a systematic issue with this method of integrating the context and suggests performance could be improved by helping the model more accurately identify the exact utterance under question. Among the cases in which no discernible pattern was identified, 30.6% absolute involved the system predicting no face act to be present (i.e., Other), and 11.2% absolute involved another prediction. Thus, in summary, we have a gold error rate of 22.7%, and we find that in 43.8% of errors the predicted face act is actually correct (possibly among others).

We also break these counts down by gold label and error category in Table 5. Note that the first three rows of the table refer to predictions that are actually correct. The analysis reveals that the system's prediction is in fact correct for a majority of the cases with gold labels of Other, and about half the time for cases in which the gold label is HPos-, SNeg+, or SPos+. The worst performance is on HNeg-.

6.3. Contribution of Dialog Acts

We now turn to investigating the effects of integrating dialog acts into the model. We do so by first examining correlations found between face acts and dialog acts included in the text-augmented model data to determine, roughly, where signal might be. The results of this analysis are shown in Table 6 and qualitatively support our hypothesis that these concepts, face and intention, are intimately related. Inspecting the MRDA tags, there is a strong positive correlation ($r = 0.49$) between questions and HNeg- with a correspondingly negative correlation ($r = -0.38$) for statements. The trend continues across the various question categories for the SWDA tags though, critically, no correlation is observed for backchannels which take the form of a question. As impositions on negative face often take the form of requests or questions rather than statements, with the exception of commands (which are not frequent in this corpus), these correlations are expected. Furthermore, the trend is nicely reversed for HPos+, which sees

	All	HNeg+	FN to TP	TP to FN	FP to TN	TN to FP
Statement	80%	93%	97%	100%	78%	91%
Question	20%	7%	3%	0%	22%	9%
Count	-	305	64	20	50	101

Table 7: Distribution of MRDA dialog act tags Statement (containing also Disruption) and Question for different cases relating to predicting HNeg+, FN refers to false negatives, TP to true positives, and so on.

positive correlation ($r = 0.20$) with statements and negative ($r = -0.24$) with questions. The SWDA tags reveal this largely comes from the appreciation ($r = 0.29$) and agreement ($r = 0.15$) labels which are stereotypical examples of the HPos+ face act. Because the corpus contains a large number of utterances in which participants signal virtue (e.g. *I often make large donations*), the correlation between SPos+ and non-opinion statements ($r = 0.30$) makes sense.

This all indicates that the dialog act data included in the text-augmented models, and learned from the same data set used in training the multi-task models, does contain information with signal. One might conclude that (1) dialog acts were not providing very much orthogonal information in training (i.e. Fos already learned to distinguish these) or, if it was, perhaps (2) these methods of integration were not effective for this task.

To untangle these options, we carefully examine examples where the gold label or output from Fos, Ta-Mrda, or Ta-Swda contained HNeg+. This label indicates that an utterance aims to raise the hearer's negative face, meaning that the speaker is trying to provide or point out options for action to the hearer (rather than restrict them). We focus on this label for analysis because it is an outlier in that this is the only label for which the Fos system did not obtain the best F1 result (see Table 3), meaning that dialog acts actually helped its performance. We use the MRDA dialog acts to investigate this phenomenon, partly because the MRDA tag set is simpler than the SWDA tag set, and partly because the Ta-Mrda system obtained the best results of all five systems. There are 305 instances of HNeg+ in the corpus. We investigate the distribution of dialog act tags. In doing this, we collapse Statement and Disruption, as we could find no meaningful distinction in this written dialog corpus. Since the other tags are exceedingly rare, we restrict our analysis to the distinction between Statement (to which we have added Disruption) and Question. The results are in Table 7.

In the first two columns, we see that overall, 20% of turns are questions, but for the 305 turns labeled HNeg+, only 7% are, which is consistent with the semantics of the tag. However, as we could see in Table 6, the correlation is not strong (5% absolute for both dialog act tags). Despite the rather weak correlation, the presence of the tag increases the F1 measure from a low 44% (the lowest of any tag) for the Fos system to 51% for the Ta-Mrda system, entirely by increasing recall from 41% to 55%. To understand why, we can look at the cases in which the MRDA tags in the input change the prediction. There are four cases: the predicted tag is HNeg+ (Positive) or some other tag (Negative), and the prediction is correct (True) or not (False). This gives us four possible shifts as we go from the Fos system to the Ta-Mrda system which uses the dialog act tags. The correct shifts from False Negative to True Positives have 97% Statements, and the correct shift from False Positives to True Negatives have 22% Questions – as expected. In terms of the incorrect shifts, the Ta-Mrda system does not change a True Positive to a False Negative frequently, and we disregard this case. The Ta-Mrda system does change a True Negative to a False Positive very frequently (101 cases), but here the distribution of Statements and Questions is very close to that of HNeg+ in general, so the dialog act label does not contribute to this error class. In conclusion, we can see that simply adding a very simple dialog act distinction helps the classifier for a low-frequency tag in the way we expect, increasing TPs and decreasing FPs.

While it may seem odd that for the three face act tags for which the correlations are strongest (SPos+, HPos+, and HNeg-) we do not see an improvement from adding the tags, we note that these three face acts are also the most common in our corpus, and we assume that the Fos system has enough data and can derive the face act tag from the lexical information on its own.

7. Conclusion and Future Work

We have presented a new study on the face act corpus of Dutt et al. (2020) and use a generative approach to obtain state-of-the-art results. The model is then augmented to investigate the role of communicative intention in determining face acts. Through several methods of analysis we find evidence that there is a close relationship between dialog acts and face acts. Despite showing some improvement on minority labels (§ 5.3) and correlations between dialog acts and face acts (§ 6.3), our augmented models see an overall decline in performance when incorporating dialog acts. Our error analysis finds issues with the annotation consistency (Other in particular) and, more importantly, points to methods of improving future annotation efforts.

We observe that some of the theoretical study that was done when developing dialog act representations is very applicable to face acts. Early work on speech act theory also limited utterances to a single label but this was, over time, identified as a serious flaw by several researchers (Cohen and Levesque, 1987; Hancher, 1979). As a result, DAMSL (Core and Allen, 1997), now the most commonly adapted methodology for dialog act annotations, made supporting multiple labels a primary objective in its design. Future annotations of face acts should do the same and the frequency that multiple act utterances were found in our error analysis supports this recommendation.

Face acts are an important part of language use, and while Dutt et al. (2020) have made a major contribution, there has been little work with this corpus, and future work will require thinking hard about the data. We hope this paper will allow other researchers to use the corpus in constructive ways while being aware of the nature of the data in more detail. Despite their similarities from a computational perspective, it does not seem to be as simple as lifting the approaches used for dialog acts when modeling face acts. In the future, we plan to develop an annotation which incorporates additional aspects of politeness theory and to label a corpus with in-house trained annotators. Once we have such a corpus, we predict that we will be able to exploit the double annotation of face acts and dialog acts in machine learning more effectively and obtain a much deeper understanding of how intention and modeling of the audience interact.

Ethics Statement

The experiments for this work were performed using computational resources that are not, in general, freely available. In part due to these computational requirements, but also a result of minimal data, we were not able to evaluate the techniques on additional languages and acknowledge the limitations this places on extending our results to other cultures. We also note along similar lines that while Brown and Levinson (1987) claim their theory of politeness to be culturally universal, this claim has been contested – most notably for eastern cultures (Al-Duleimi et al., 2016; Purkarthofer and Flubacher, 2022). As discussed in detail above, taking utterances to have a single face act or intent is a critically limiting assumption which lends some uncertainty to our conclusions.

Despite a detailed analysis of the errors, we cannot verify the safety of this system in any user-oriented context and therefore do not recommend such uses without further study. While we do not produce any data sets directly from human annotations, we do use several which were, to the best of our knowledge, compiled ethically. As the primary object of study in this work is the relationship between face and intention, we do not anticipate broad risks to its application.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation (NSF) under No. 2125295 (NRT-HDR: Detecting and Addressing Bias in Data, Humans, and Institutions) as well as by funding from the Defense Advanced Research Projects Agency (DARPA) under the CCU (No. HR001120C0037, PR No. HR0011154158, No. HR001122C0034) program. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or DARPA.

We thank both the Institute for Advanced Computational Science and the Institute for AI-Driven Discovery and Innovation at Stony Brook for access to the computing resources needed for this work. These resources were made possible by NSF grant No. 1531492 (SeaWulf HPC cluster maintained by Research Computing and Cyberinfrastructure) and NSF grant No. 1919752 (Major Research Infrastructure program), respectively.

We would also like to thank our anonymous reviewers for their perceptive comments, which improved this work, and Shyne Choi for her participation in several conversations which encouraged this study.

8. Bibliographical References

Hutheifa Y. Al-Duleimi, Sabariah Hj Md Rashid, and Ain Nadzimah Abdullah. 2016. A critical review of prominent theories of politeness. *Advances in Language and Literary Studies*, 7:262–270.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

J. L. Austin. 1962. *How to do things with words*. Oxford University Press.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Philip R Cohen and Hector J Levesque. 1987. *Rational interaction as the basis for communication*, volume 87. CSLI Stanford.

Mark G. Core and James F. Allen. 1997. Coding Dialogs with the DAMSL Annotation Scheme.

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.

Ritam Dutt, Rishabh Joshi, and Carolyn Rose. 2020. Keeping up appearances: Computational modeling of face acts in persuasion oriented discussions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7473–7485, Online. Association for Computational Linguistics.

Herbert Paul Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and semantics, vol 3*. Academic Press, New York.

Michael Hancher. 1979. The classification of cooperative illocutionary acts. *Language in society*, 8(1):1–14.

Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani. 2021. Speaker turn modeling for dialogue act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2150–2157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Murzaku, Peter Zeng, Magdalena Markowska, and Owen Rambow. 2022. Re-examining FactBank: Predicting the author's presentation of factuality. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 786–796, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Judith Purkarthofer and Mi-Cha Flubacher. 2022. *Speaking Subjects in Multilingualism Research: Biographical and Speaker-centred Approaches*, volume 7. Channel View Publications.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000a. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26:339–373.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000b. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Marilyn A. Walker, Janet E. Cahn, and Stephen J. Whittaker. 1997. Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the First International Conference on Autonomous Agents*, AGENTS '97, page 96–105, New York, NY, USA. Association for Computing Machinery.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.