Towards Awareness of Human Relational Strategies in Virtual Agents

Ian Beaver, Cynthia Freeman
Verint Next IT
Spokane Valley, WA USA
{ian.beaver, cynthia.freeman}@verint.com

Abdullah Mueen
Department of Computer Science
University of New Mexico, USA
mueen@unm.edu

Abstract

As Intelligent Virtual Agents (IVAs) increase in adoption and further emulate human personalities, we are interested in how humans apply relational strategies to them compared to other humans in a service environment. Human-computer data from three live customer service IVAs was collected, and annotators marked all text that was deemed unnecessary to the determination of user intention as well as the presence of multiple intents. After merging the selections of multiple annotators, a second round of annotation determined the classes of relational language present in the unnecessary sections such as Greetings, Backstory, Justification, Gratitude, Rants, or Expressing Emotions. We compare the usage of such language in human-human service interactions. We show that removal of this language from task-based inputs has a positive effect by both an increase in confidence and improvement in responses, as evaluated by humans, demonstrating the need for IVAs to anticipate relational language injection. This work provides a methodology to identify relational segments and a baseline of human performance in this task as well as laying the groundwork for IVAs to reciprocate relational strategies in order to improve their believeability.

Introduction

Conversational AI is an active field of research involving techniques for software agents to engage in natural conversational interactions with humans (Ram et al. 2018). A popular application of conversational AI is Intelligent Personal Assistants such as Apple's Siri or Google Assistant, which are commonly used for answering questions and task optimization. Company-specific automated assistants, known as Intelligent Virtual Agents/Assistants or IVAs, are increasingly automating the first layer of technical and product support with a market value projected to reach $11.3 billion by 2024 (IMARC Group 2019). In these business domains the accuracy and efficiency of IVAs directly impacts customer satisfaction and company support costs.

To better assist humans, IVA designers strive to support human-like interactions. Take, for example, Amazon's Alexa Prize competition where student developers attempt to build IVAs that can carry on meaningful and engaging conversations for 20 minutes (Khatri et al. 2018). One psychological aspect of human conversation that has not yet been well addressed in conversational AI is the **relational strategies** (e.g. self-disclosure and justification) that humans employ when conversing with one another.

As IVAs become more human-like, we theorize that users will increasingly use relational strategies with IVAs similar to conversing with humans. There is a large body of work on development of trust between humans engaged in virtual dialog (Ballantyne 2004; Coppola, Hiltz, and Rotter 2004; Bickmore and Cassell 2001). In addition, there has been significant research into how humans communicate relationally when trying to perform tasks with other humans. Such communication may be to influence (e.g. compliance gaining (Littlejohn and Foss 2009; Wilson 2002) and affinity seeking (Bell and Daly 1984)) or to invoke reciprocal relational strategies (Mottet, Martin, and Myers 2004) by the other party. The overarching focus of such work is on how relational strategies contribute to trust and communication between human speakers. From this literature, we predict the types of strategies humans may employ with IVAs as they relate to them in an increasingly human manner.

In customer service and personal assistant domains, trust is necessary between the human agent and customer. The customer's issues must be viewed by the agent as legitimate for proper attention to be given. Likewise, customers must trust that the agent is capable of assisting them and will not mistreat their information. Current research shows that human-like virtual agents are associated with not only greater user trust but also trust resilience when the agent makes mistakes (de Visser et al. 2016; Bickmore and Picard 2005). To build trust with the agent, users may establish credibility through small talk, self-disclosure, and by providing justification of their requests (Bickmore and Cassell 2001). Handling such relational language directly has not yet been a focus of the conversational AI community.

In interactive question answering, such as dialogs with an IVA, understanding **user intent** is essential for the success of the IVA (Chai, Zhang, and Baldwin 2006). The intent can be defined as the interpretation of a user input that allows an agent to formulate the *best* response. However, when relational strategies are applied to IVAs, the additional language introduced can obfuscate the primary user intent. If not anticipated, such language can lead to confusion in the IVA and a degradation of user experience in the form of clarification questions and wrong information.

Example 1

I need a ticket to Boston this Saturday, my son is graduating!

In Example 1, the fact that the customer's son is graduating is unnecessary for determining the user's intent to purchase a ticket. By considering unnecessary background information when determining business intent, the IVA may incorrectly deduce that the customer is booking a ticket *for* his or her son instead. Thus, the identification of relational segments is a useful feature for an IVA; to our knowledge, no corpus of annotated relational segments exists to measure their incidence and study their effect on commercially deployed IVAs (Serban et al. 2018). Such a corpus could help IVAs better determine user intent and task-specific behavior in the presence of relational language.

This lack inspired us to create such a corpus. Within this corpus, we needed to not only identify the location of relational language but also label its type (Gratitude, Greetings, etc.) so that automated methods to determine the relational strategy in use can be explored in future work. For IVAs to become more human-like, determining which segments of a request are relational is necessary to allow these IVAs to both understand the user intent correctly and to include empathetic or reciprocal relational strategies. For example, an IVA responding to the input from Example 1 could set aside the segment on why the user was traveling to more precisely identify the business intent of booking a ticket for the user. Then during the response generation phase the IVA could parse the relational segment to determine a relational response to append to the business response such as "Congratulations! Where will you be leaving from?".

The identification of relational strategies in a single conversational turn can be structured as a multi-intent detection problem. The user not only wants the task completed (the *primary* intent); they may also attempt to build credibility or some common ground with the IVA (the *secondary* intent). Multi-intent detection within dialog systems is still an emerging field (Khatri et al. 2018). A few methods exist such as (Xu and Sarikaya 2013) which uses multi-label learning and (Kim, Ryu, and Lee 2017) which employs a two-stage intent detection strategy. However, (Xu and Sarikaya 2013) provided no explanation of how data was annotated nor any mention of annotator agreement. In (Kim, Ryu, and Lee 2017), multi-intent data was fabricated by concatenating all combinations of single-intent sentences.

Furthermore, in the previous works, multi-intent detection is assumed to be used for separating multiple task-oriented intents within a single turn. In this work we show that is not always the case, the secondary intents may be relational in nature and must therefore be handled differently. Future dialog systems could include a component to track the current relational strategy in use and how to respond similar to how current dialog systems employ components for dialog state tracking to manage what the user wants from the system at each step (Henderson, Thomson, and Williams 2014).

In this paper, we provide several contributions. Most importantly, we create the first publicly available corpus with annotated relational segments. We propose an evaluation measure and set a baseline by comprehensive human annotation, ultimately confirming that the addition of relational language can obfuscate the user's intention to IVAs not designed to recognize it. Along with annotated relational segments, we create a corpus of human annotated real-world multi-intent requests to further research in multi-intent detection. We analyze human agreement in determining the presence of multiple intents so that future research on multi-intent detection can be evaluated in the light of prior human performance. Through these contributions, we hope to encourage further research and ultimately aid in the design of more believable IVAs and chatbots for customer service.

Data Collection

Verint - Next IT designs and builds IVAs on behalf of other companies and organizations, typically for customer service automation. This unique position allows access to a large number of IVA-human conversations that vary widely in scope and language domain. We selected IVAs for data collection based on the volume of conversations engaged in, the scope of knowledge, and the diversity of the customer base.

For diversity, we considered whether the target user base of the IVA was regional, national, or international and mapped the locations of the active user base to visually verify. We only considered IVAs that had a national or international target user base and did not appear to have a dominate regional clustering to ensure that conversations were well distributed across users from different regions. This was to control for relational styles that may differ between regions.

Human-computer data was collected from three live customer service IVAs in the language domains of airline, train travel, and telecommunications. The selected IVAs are implemented as mixed-initiative dialog systems, each understanding more than 1,000 unique user intentions. These IVAs use a symbolic ontological approach to natural language understanding, and all three were very mature, each having been deployed and continuously refined for four or more years. The IVAs have conversational interfaces exposed through company websites and mobile applications. In addition, the IVAs are multi-modal, accepting both speech and textual inputs, and also have human-like qualities with simulated personalities, interests, and an avatar.

A random sample of 2,000 conversations was taken from each domain. The samples originate from conversation logs during November 2015 for telecommunications and train travel and March 2013 for airline travel. There were 127,379 conversations available in the logs for the airline IVA. The telecommunications and train travel logs contained 837,370 and 694,764 conversations, respectively. The first user turn containing the problem statement was extracted. We focus on the initial turn as a user's first impression of an IVA is formed by its ability to respond accurately to his or her problem statement, and these impressions persist once formed (Madhavan, Wiegmann, and Lacson 2006). Therefore, it is imperative that any relational language present does not interfere with the IVA's understanding of the problem statement.

To comply with data use agreements of our customers, the data was fully scrubbed of all personally identifiable infor-

	Requests	Multi-Intent	Single Intent	Unnecessary	Avg. Length
TripAdvisor	2000	734	1266	94.1%	93.26
Telecom	2000	149	1851	77.3%	19.81
Airline	2000	157	1843	68.6%	21.64
Train	2000	201	1799	55.3%	20.07

Table 1: Dataset statistics. The Multi-Intent column represents the count of Requests where one or more annotators flagged it as containing more than one user intention. The Unnecessary column represents the percentage of Single Intent requests where one or more annotators selected *any* text as being unnecessary in determining user intent. Avg. Length is the number of words present in Requests, on average.

mation, mentions of the originating customer, and products that reveal their identity. Detailed sanitation steps followed and a data use agreement are included with the released data.

Finding a large mixed-initiative human-human customer service dataset for comparison with our human-computer dialogs proved difficult. Despite mentions of suitable data in (Vinyals and Le 2015) and (Roy et al. 2016), the authors did not release their data. A commonly used mixed-initiative dataset is the Ubuntu Dialogue Corpus (Lowe et al. 2017). The corpus originates from an Internet Relay Chat channel where users discuss issues relating to the Ubuntu operating system. However, for our experiments we desire to observe the effect of relational language on the existing IVA understanding in addition to measuring occurrence. To do this, we needed user intents that were very similar to those already handled by one of the selected IVAs. As the Ubuntu dataset is not in the domain of travel or telecommunications, we searched for publicly visible question and answering data in domains similar to those of the selected IVAs.

TripAdvisor.com is commonly used in literature as a source of travel-related data (Banerjee and Chua 2016; Valdivia, Luzón, and Herrera 2017). The TripAdvisor.com airline forum includes discussions of airlines and polices, flight pricing and comparisons, flight booking websites, airports, and general flying tips and suggestions. We observed that the intentions of requests posted by users were very similar to that of requests handled by our airline travel IVA. While a forum setting is a different type of interaction than chatting with a customer service representative (user behavior is expected to differ when the audience is not paid to respond), it was the best fit that we could obtain for our study. A random sample of 2,000 threads from the 62,736 present during August 2016 was taken, and the initial post containing the problem statement was extracted. We use **request** hereafter to refer to the complete text of an initial turn or post extracted as described.

Annotation

From our four datasets of 2,000 requests each, we formed two equally-sized partitions of 4,000 requests with 1,000 pulled from every dataset. Each partition was assigned to four annotators; thus, all 8,000 requests had exactly four independent annotations. All eight annotators were employees of our company who volunteered to do the task in exchange for a $150 gift card. They worked on this task during company time; the gift card was in addition to their regular pay.

The annotators were instructed to read each request and mark *all* text that appeared to be additional to the user intention. The annotators were given very detailed instructions and were required to complete a tutorial demonstrating different types of relational language use before working on the actual dataset. As the data was to be publicly released, we ensured that the user intention was clear. If more than one user intention was observed, the annotator was instructed to flag it for removal. This was a design decision to simplify the problem of determining language necessary for identifying the user intention. Furthermore, as mentioned in the introduction, IVAs with the ability to respond to multiple intentions are not yet commonplace. Although flagged requests were not used for further analysis, they are included in the corpus to enable future research on multi-intent detection. After discarding all multi-intent requests, 6,759 requests remained. Per-dataset statistics are given in Table 1.

A request from the TripAdvisor data is given in Example 2 below. A annotator first read over the request and determined that the user intent was to gather suggestions on things to do in Atlanta during a long layover. The annotator then selected all of the text that they felt was not required to determine that intent. This unnecessary text in Example 2 is struck through. Each of the four annotators performed this task independently, and we discuss in the next sections how we compare their agreement and merged the annotations.

Annotators averaged 1 request per minute to perform this task on their assigned TripAdvisor requests and 4 per minute on their assigned requests from the three IVA datasets. We observed that each of the eight annotators required 29 hours on average to complete their 4,000 assigned requests.

Example 2
Original Request: Hi My daughter and I will have a 14 hour stopover from 20.20 on Sunday 7th August to 10.50 on Monday 8th August. Never been to Atlanta before. Any suggestions? Seems a very long time to be doing nothing. Thanks

Determine User Intent: *Things to do on layover in Atlanta*

Annotated Request: ~~Hi~~ My daughter and I will have a 14 hour stopover ~~from 20.20 on Sunday 7th August to 10.50 on Monday 8th August~~. Never been to Atlanta before. Any suggestions? ~~Seems a very long time to be doing nothing. Thanks~~

A: ~~Hi,~~ I need a new credit card[~~, my old one doesn't work any more.~~] Can you help?
B: ~~Hi~~, I need a new credit card, my old one doesn't work any more.[~~ Can you help?~~]

$$n = 77 \quad \sum_{i=1}^{77} e_i = 49 \quad align_{AB} = \frac{77-49}{77} = 0.364$$

Figure 1: Example alignment scoring between two fabricated annotations A and B. Struck through text was marked as unnecessary for intent determination. Positions with an alignment error are between "[" and "]".

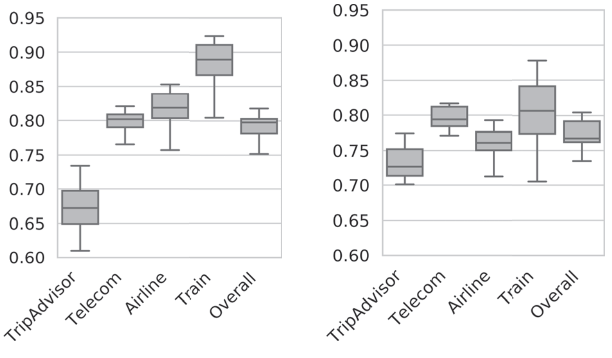
Annotation Alignment

To compare the raw agreement of annotations between two annotators, we use a simplification of the word error rate (WER) metric, a concept in speech recognition from hypothesis alignment to a reference transcript (Zechner and Waibel 2000). We modify this procedure as substitutions cannot occur. Annotators mark sequences of text as being unnecessary in determining user intention. When comparing annotations between two annotators, an error (e_i) is considered to be any character position i in the text where this binary determination does not match between them. $e_i = 1$ will represent either an insertion or deletion error from one annotation to the other. The alignment score can be calculated as:

$$align = \frac{n - \sum_{i=1}^{n} e_i}{n}$$

where n is the total number of characters. Thus, $align \in [0, 1]$ where 1 is perfect alignment. Annotators may or may not include whitespace and punctuation on the boundaries of their selections which can lead to variations in e_i. Therefore, when two selections overlap, we ignore such characters on the boundaries while determining e_i. Figure 1 shows a fabricated example of alignment between two annotations. In the first selection, disagreement on trailing whitespace and punctuation is ignored as it occurred within overlapping selections. Notice, however, that whitespace and punctuation count as alignment errors in the last selections as there is no overlap with the other annotator; therefore, there is no possibility of disagreement on the boundaries.

Another common similarity metric is the BLEU score. However, the BLEU score was also based on the WER metric, modified specifically to deal with language translation issues such as differences in word choice, word ordering, and allowing multiple reference translations (Papineni et al. 2002). As none of these situations can occur in our task, the BLEU score does not add any insight to the WER metric, and is more complex. In addition, we can look at the insertion and deletion errors used to create e_i and gain some intuition about *how* two annotators disagreed. In Figure 1, the insertion and deletion errors are 14 and 35 respectively, indicating that annotator A selected quite a bit of text B thought was necessary, but B also selected some text that A thought was necessary. If insertion error was 0 and any deletion errors existed, we would know that B selected a subset of A's selection. This additional insight is not given by the BLEU score, therefore we use the alignment score defined above.



(a) (b)

Figure 2: The distribution of average alignment scores between all four annotations per dataset is shown in **(a)**. We compute average alignment scores where all annotators agree that additional language is present in **(b)**.

The alignment score was calculated for every request between all four annotations and then averaged. For example, an alignment score was calculated for each request between annotator A and B, A and C, A and D. The same process was repeated between annotator B and C, B and D, then C and D. Finally, alignment scores between all unique pairs of annotators over all requests were averaged per dataset. The distribution of average scores per dataset is shown in Figure 2 **(a)**. It may appear that two annotators could inflate the dataset alignment score by simply making annotations infrequently thereby having a high alignment. However, as each request had four annotators, the average alignment score would decrease as those annotators would have large error compared to the other two. The per dataset alignment averages can, in fact, be higher if a dataset has a large number of requests where *no* annotator selected any text.

Therefore, it is interesting to remove the effect of these cases and compare the ability of annotators to agree on the selection boundaries given they both agree that selection is necessary. To measure this, we compute average alignment scores where both annotators agree that additional language is present, shown in Figure 2 **(b)**. Observe that although the Train dataset has the highest overall alignment in both cases, it is lower when the annotators both select text, indicating it has many cases where no annotators selected anything (which is in agreement with Table 1). In the case of TripAdvisor, it appears that there are a significant number of requests where one or more annotators do not select text, but the others do, lowering the overall alignment score in Figure 2 **(a)**.

Calculating alignment based on word-level instead of character-level agreement was also considered. For each word, if the annotator selected at least 50% of the word it was considered to be marked. This resolves situations where a annotator accidentally missed the first or last few characters of a word in their selection. However, this may introduce errors where two letter words have only one character

Table 2: Annotator agreement on if any text should be selected. For example, row 3 is the number of requests with selections by at least three annotators.

	TripAdvisor	Train	Airline	Telecom
κ	0.270	0.450	0.405	0.383
1	1192	995	1264	1431
2	1092	709	948	1154
3	863	458	644	795
4	534	205	292	410

Table 3: Annotator agreement on multi-intent detection. For example, row 3 is the number of requests flagged as containing multiple intentions by at least three annotators.

	TripAdvisor	Train	Airline	Telecom
κ	0.415	0.374	0.434	0.386
1	734	201	157	149
2	480	85	69	56
3	275	50	38	32
4	71	8	15	11

selected. In this case it is impossible to automatically decide if the annotator meant to select the word or not as always selecting such words will be susceptible to the same error.

Selected words were then used in place of selected characters in calculating the alignment scores between the annotators in the same manner as Figure 1. We discovered that the alignment scores were only 0.2% different on average across all datasets than the character level alignment scores shown in Figure 2. This indicates that annotators are rarely selecting partial words, and any disagreement is over *which* words to include in the selections. Therefore, in the released corpus and in this paper, we consider selections using absolute character position which retains the annotators' original selection boundaries. This choice allows for others to experiment with both word-level and character-level selection methods using our data.

Agreement Between Annotators

As it is difficult to determine how often all annotators agree additional language is present from alignment scores alone, we measured annotator agreement on the presence of additional language and multiple user intentions. For additional language presence, we calculated Fleiss' κ over the annotations where the classes compared (K) were if a annotator did or did not select text. As demonstrated in Table 2, regardless of domain, this is a subjective task. While there is moderate agreement in the Train and Airline sets, the TripAdvisor set, in particular, is lower in agreement which reinforces our previous observations in Figures 2 **(a)** and **(b)**. When the annotations of two fallible observers are compared, the κ measurement is very sensitive to the number of classes selected from and the variability of the probability of class membership (Feinstein and Cicchetti 1990; Guggenmoos-Holzmann 1993). Therefore, these values must be interpreted in light of the task. In our task, $K = 2$ (text selected or not), and the probability of selection varied per annotator and dataset. Under these conditions, according to the chart in (Bakeman et al. 1997), a κ between 0.27 and 0.45 suggests annotator accuracy between 85% to 90%, respectively. Therefore, despite the lower values for κ, the individual annotator annotations appear reliable and can be further improved when merged based on agreement as described later.

We did observe some situations where two annotators disagree on the real intent of the user thereby causing conflict in the selection of unnecessary text. Example 3 demonstrates how even humans sometimes struggle with determining the intention of written requests. Annotator R1 appears to believe that the primary intent of the user is to notify the agent about poor television reception, and the query about the outage in the area is out of curiosity. However, annotator R7 appears to believe the primary intent is to discover if a cable outage is present in the area, and the complaint about reception justifies the query. The effects of these disagreements on intent can be mitigated by merging the annotations based on the number of annotators who agreed on a selected character.

Example 3

$R1$: Our tv reception is horrible. ~~is there an outage in my area?~~

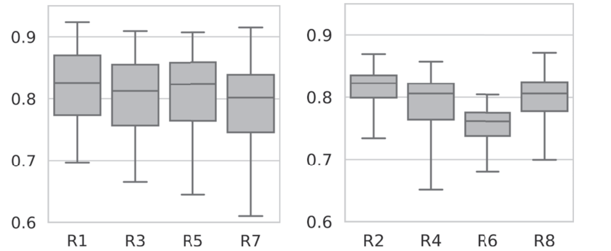
$R7$: ~~Our tv reception is horrible.~~ is there an outage in my area?

Next, we considered the annotators' determination of multiple intentions. A κ was calculated over how annotators flagged requests containing more than one user intention. As shown in Table 3, we see somewhat similar performance in this task as in the previous selection task. This table demonstrates the difficulty of multi-intent detection, even for humans. The domain does not seem to be a factor as κ is similar across datasets. It is apparent, however, that in the forum setting, users are much more likely to insert multiple intentions in a single request than in a chat setting. This task is also binary (multi-intent present or not), therefore $K = 2$, and the probability of flagging varied per annotator and dataset. These factors and κ values would suggest annotator accuracy between 87% to 90% using (Bakeman et al. 1997).

How one annotator compared to the others in their selections is another aspect to be considered. Figure 3 **(a)** compares how each annotator agreed with the other 3 in the first group. We can see that, overall, the mean is very close. However, annotator R7, in particular, had more variation in his or her selections. Similarly, Figure 3 **(b)** compares how each annotator agreed with the other 3 in the second group. In the second group, we see slightly more disagreement, particularly with annotator R6. This could be because he or she did not always interpret the user intention the same as others or because the annotator was more generous or conservative in selections compared to the others in the group.

Annotating Relational Content

To determine the use of relational strategies, a second round of manual analysis was performed. The four annotations per



(a) Alignment between group 1 annotators.

(b) Alignment between group 2 annotators.

Figure 3: Alignment scores between each annotator and the other three members of their group, averaged across the four datasets.

request were **merged** using the following strategy: for every character position in the request, if at least a threshold of two annotations contained that position, **highlight** it. The agreement of two is to mitigate confusion over the user intent as previously discussed. Once merged, highlighted sections were analyzed by the authors to determine the classes of language present. A comparison of relational annotation using all agreement levels is left for future works.

Each highlighted section was evaluated and given one or more of the following tags. We surveyed literature on relational communication theory to determine the set of relational tags to apply. In doing so we discovered there is little agreement in the field on a taxonomy of relational strategies. For example, 74 different taxonomies exist just in the subset of compliance gaining (Littlejohn and Foss 2009). To simplify the tagging task and prevent disagreement on nuances, we agreed to focus on high level relational behaviors that were easily spotted in customer service dialogs.

Greetings are a common relational strategy humans use to build rapport with other humans and machines (Lee, Kiesler, and Forlizzi 2010).

Backstory is a style of self-disclosure to create intimacy and is particularly important at the beginning of a relationship when communicants seek similarities (Littlejohn and Foss 2009). In Example 1, the customer included the *Backstory* that he or she is attending their child's graduation. This may be an attempt to build common ground with the agent or it may indicate the importance of the trip and motivate the agent to help the customer succeed.

Justification and excuses are forms of account giving and used to obtain credibility, deny responsibility, or argue why the agent should take action on the part of the customer (Littlejohn and Foss 2009). For instance, when trying to receive a refund, a customer may state the product was defective and therefore he or she should be refunded due to policy.

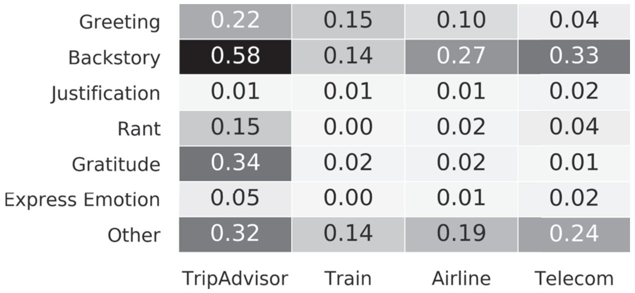
Gratitude, like greetings, are used by humans to also build rapport with humans and machines (Lee, Kiesler, and Forlizzi 2010).

Ranting is a means of expressing dissatisfaction when a customer feels frustrated, ignored, or misunderstood. In computer-mediated conversations, the non-verbal emotional cues present in face-to-face conversations are missing; thus, humans resort to such negative strategies to convey their emotions (Laflen and Fiorenza 2012). For tagging purposes, we define a *Rant* to encompass any excessive complaining or negative narrative.

Expressing emotions can be a means of showing displeasure when a customer feels a conversation is not making adequate progress or in reaction to an unexpected or disagreeable agent response. This can also indicate joking or other positive emotional expression. The tag *Express Emotion* is used as a catch-all for any emotional statement that is not covered by *Rant*. Examples would be: *"i love that!", "UGH!", "WHY???", "lol"*.

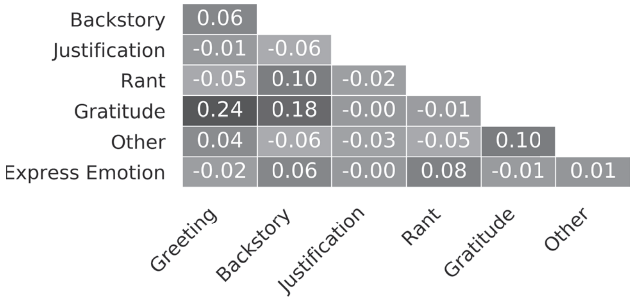
The **Other** tag indicates that some or all of the selection does not contain any relational language. This is commonly a restatement of the primary intent or facts that annotators marked as unnecessary for the primary intent. The dates and times in Example 2 would be tagged as *Other* as they were selected but are not relational content.

Analysis of Relational Tags



	TripAdvisor	Train	Airline	Telecom
Greeting	0.22	0.15	0.10	0.04
Backstory	0.58	0.14	0.27	0.33
Justification	0.01	0.01	0.01	0.02
Rant	0.15	0.00	0.02	0.04
Gratitude	0.34	0.02	0.02	0.01
Express Emotion	0.05	0.00	0.01	0.02
Other	0.32	0.14	0.19	0.24

Figure 4: Incidence of relational language per dataset. An incidence of 0.5 means the tag is present in 50% of all Single Intent requests in Table 1.



	Greeting	Backstory	Justification	Rant	Gratitude	Other
Backstory	0.06					
Justification	-0.01	-0.06				
Rant	-0.05	0.10	-0.02			
Gratitude	0.24	0.18	-0.00	-0.01		
Other	0.04	-0.06	-0.03	-0.05	0.10	
Express Emotion	-0.02	0.06	-0.00	0.08	-0.01	0.01

Figure 5: Pearson coefficients of tag correlation across datasets.

As shown in Figure 4, we see that backstory and gratitude are common in human-to-human forum posts. This analysis was done only on the problem statement (first post or user turn), not the entire conversation. While it appears that humans do not thank IVAs, this is more likely an artifact of the forum format versus a conversation format. In a forum post it is common for the initial poster to thank the readers in advance for any help they may provide. However, in a live conversation setting, gratitude is generally reserved for after help is actually provided. Therefore, we cannot draw the conclusion from this table that humans thank other humans more than IVAs as humans may actually be thanking the IVAs at the end of the conversation with similar frequency.

We can say that both Airline and Telecom IVAs also have a significant amount of backstory. For both domains nearly 1 out of every 3 requests contained backstory. Although minimal, ranting and justification were also present in Telecom more than the other IVAs. This could be because the Telecom IVA has more of a product and service support role than the travel IVAs, which focus more on planning travel. The Train dataset appeared to contain many greetings while having the least amount of other relational language. It is difficult to speculate why without deeper analysis of the user demographic, the presentation of the IVA on the website, and the IVA knowledge base.

The correlation between tags is shown in Figure 5. When greetings are present, it appears that there is a likelihood there will also be gratitude expressed which agrees with the findings in (Lee, Kiesler, and Forlizzi 2010) and (Makatchev, Lee, and Simmons 2009). Also interesting is the apparent correlation between backstory and gratitude. Those that give background on themselves and their situations appear more likely to thank the listener. Ranting appears to be slightly negatively correlated with greetings, which is understandable assuming frustrated individuals are not as interested in building rapport as they are venting their frustrations.

Relational Content and IVA Understanding

To measure the effect of relational language on IVA performance and determine what level of annotator agreement is acceptable, we first constructed highlights for the 6,759 requests using all four levels of annotator agreement. Next, four *cleaned* requests were generated from each original request by removing the highlighted portion for each threshold of annotator agreement resulting in 27,036 requests with various amounts of relational language removed.

Every unaltered request was fed through its originating IVA, and the intent confidence score and response was recorded. We then fed each of the four cleaned versions to the IVA and recorded the confidence and response. The TripAdvisor data was fed through the Airline IVA as it provided the most similar domain. This was also a test to see if lengthy human-to-human forum posts could be condensed and fed into an existing IVA to generate acceptable responses.

The three commercial IVAs evaluated use a symbolic language model for natural language understanding in which a simple confidence metric is calculated by the percentage of input consumed by known vocabulary within the patterns of the responding intent. Text consumed by the wild card character (.) and generic placeholders (e.g. \s or \w) are not considered for scoring. Therefore, the more unique words matched in the input, the higher the intent confidence.

We are interested in how the confidence metric changes as language is removed. We would expect that if highlighted sections contained language unnecessary to determine the intent, confidence scores should increase as this language is removed. If the confidence score is static or decreases, it would indicate that vocabulary necessary for determining the intent has been removed. In Figure 6 we see that when sections are removed, the confidence scores increase for all domains. This shows that text marked by annotators is indeed unnecessary to determine intent, even at lower annotator agreement levels, and may cause confusion in the IVAs.

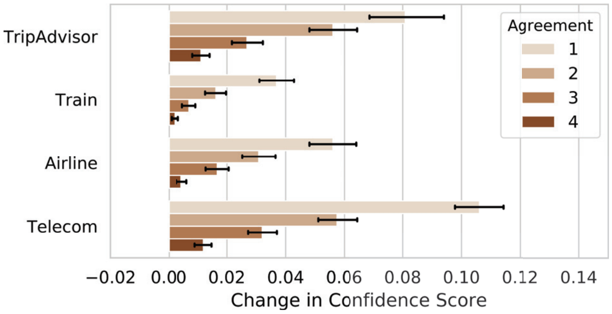


Figure 6: Change in confidence score (%) when highlighted text is removed by differing thresholds of annotator agreement. Black bars indicate 95% confidence intervals.

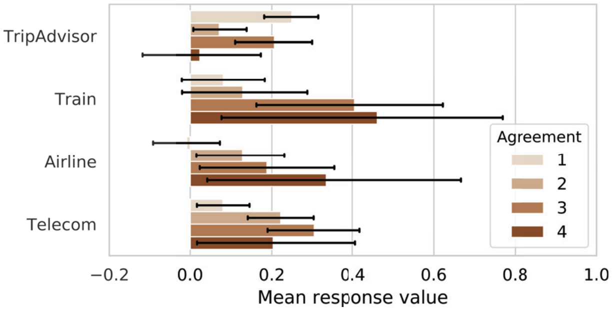


Figure 7: Results of the A-B test on IVA response to original request versus cleaned request by annotator agreement level.

In addition to intent confidence, we measured the effect of relational language removal on overall customer experience. An A-B test was conducted where four annotators were shown the user's original request along with the IVA response from the original request and the IVA response from a cleaned request. They were asked to determine which, if any, response they believed better addressed the original request. If the original IVA response was preferred, it was assigned the value -1. If the response to the cleaned request was preferred, it was assigned the value 1. Finally, if neither response even remotely addressed the user's request or if both responses were comparable, it was given the value 0.

This A-B test was done only on responses that changed as a result of the cleaned request (3,588 IVA responses changed out of the 27,036 total responses). The result of this analysis is shown in Figure 7. Note that the lower bound is -1, indicating the original IVA response is preferred. The IVA response to the cleaned request is preferred as made evident by the significantly positive skew. 95% confidence intervals are included, and although they may seem large, this is expected; recall that a 0 was assigned if both IVA responses address the user request comparably or neither did. In 10 of the 16 cases, the skew is towards the cleaned response within the 95% confidence interval.

This is further evidence that the current usage of relational language has a measurable negative effect on live commercial IVAs. TripAdvisor is an interesting exception, especially when the threshold is 4. However, this can be somewhat expected as it is a human-to-human forum where user inputs are significantly longer, and primary intent can be difficult to identify even for a human.

Although, in general, the removal of language is beneficial, how *much* removal? This is another question addressed in Figure 7. The higher the threshold, the more annotators need to agree on the removal of the same segment of text. Thus, although language may still be removed, less language is removed as threshold increases due to low kappa (see previous section on agreement). In effect, the higher thresholds may remove less unneeded language but the language that *is* removed is more likely to be actually unnecessary. However, using a threshold of 4 seems to have limited improvement over 3 due to the annotator disagreement.

Discussion and Future Works

The eight annotators were familiar with IVAs and the task of intent determination but did not work directly on any of the IVAs used in the study. They were therefore skilled in the field but not experts in the language domains. As all were given training before working on the data and the authors monitored and answered all questions during the task, we are confident that the annotators understood the task well. The reviewer disagreement on the selection of text is not surprising given it can be difficult to know exactly what a user's intent was from a request in isolation.

For example, in the following request: *"i am having a problem finding web sites that have multi city flights. is it better to book one way from each airport? thanks"*; all four annotators selected *"thanks"*, three annotators selected *"i am having a problem finding web sites that have multi city flights."* but one selected *"is it better to book one way from each airport?"* instead. This indicates confusion on if the user's real intent was to find a web site that offers multi city flights or to get an answer to the question about booking one way. Both are valid assumptions and different text would be selected based on the interpretation.

Therefore, if this annotation task is repeated, we would suggest showing the annotators the request within the context of the full conversation. This will slow down the already slow annotation process, but it is clear that seeing the resolution would help annotators agree on what the true intent was, and therefore reduce the disagreement on text selection.

Despite these disagreements, a majority consensus can be reached as previously described creating a gold standard in relational language separation. There are several uses for this corpora. First, the relational segments can be used to investigate means to create parsers to identify relational segments in task-oriented chat settings. If relational strategies are viewed as a secondary intent, work in multi-intent detection such as (Kim, Ryu, and Lee 2017) may be applied to separate the primary intent in order to improve intent classification accuracy. Secondly, promising work has been done with transfer learning for emotion recognition using small datasets in image processing (Ng et al. 2015). Similar transfer learning approaches may use the tagged relational segments to create relational strategy classifiers, thereby allowing IVAs to create reciprocal strategies to provide a more human-like conversation.

Finally, as multi-intent detection is still an unsolved problem (Khatri et al. 2018), the 1,240 multi-intent requests can be used as a multi-domain real-world evaluation set for developing multi-intent classifiers. The intent boundaries were not annotated in this study, but this task is a focus of future work. At the least, this study has recorded a baseline human performance in the task of multi-intent detection.

In the future, we wish to repeat this exercise giving the annotators the resolution as well as the request to improve agreement, and find a more similar human-human task-oriented chat dataset that would be more comparable in communication style than the forum format of Tripadvisor.

Conclusion

In this work we have collected a corpus of human-computer and human-human task-based inputs and defined a methodology to annotate the relational segments within them. We have established a baseline of human agreement on minimal language for intent determination and multi-intent detection, which is missing in literature. Through analysis of this corpus we have shown that users of commercial IVAs are already applying relational strategies to them. For instance, 1 of 4 inputs in Airline and 1 of 3 in Telecom contained Backstory, an important similarity-seeking strategy. Justification was used equally across the human-human and human-computer datasets. It is our prediction that these strategies will increase as IVAs become more ubiquitous and human-like. We have also shown simplifying inputs before determining user intention as in our experiment can increase intent classification accuracy.

Once intent determination is made, further classifying relational segments to explicitly identify relational content will allow future IVAs to respond with relational strategies of their own. Such an approach may greatly improve the relational abilities and believability of IVAs. By providing this methodology and data[1] to the community, we aim to contribute to the development of more relational and, therefore, more human-like IVAs and chatbots.

References

Bakeman, R.; McArthur, D.; Quera, V.; and Robinson, B. F. 1997. Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods* 2(4):357.

Ballantyne, D. 2004. Dialogue and its role in the development of relationship specific knowledge. *Journal of Business & Industrial Marketing* 19(2):114–123.

[1] http://s3-us-west-2.amazonaws.com/nextit-public/rsics.html

Banerjee, S., and Chua, A. Y. 2016. In search of patterns among travellers' hotel ratings in tripadvisor. *Tourism Management* 53:125–131.

Bell, R. A., and Daly, J. A. 1984. The affinity-seeking function of communication. *Communications Monographs* 51(2):91–115.

Bickmore, T., and Cassell, J. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 396–403. ACM.

Bickmore, T. W., and Picard, R. W. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12(2):293–327.

Chai, J. Y.; Zhang, C.; and Baldwin, T. 2006. Towards conversational qa: automatic identification of problematic situations and user intent. In *Proceedings of the COLING/ACL on Main conference poster sessions*, 57–64. Association for Computational Linguistics.

Coppola, N. W.; Hiltz, S. R.; and Rotter, N. G. 2004. Building trust in virtual teams. *IEEE transactions on professional communication* 47(2):95–104.

de Visser, E. J.; Monfort, S. S.; McKendrick, R.; Smith, M. A.; McKnight, P. E.; Krueger, F.; and Parasuraman, R. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22(3):331.

Feinstein, A. R., and Cicchetti, D. V. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology* 43(6):543–549.

Guggenmoos-Holzmann, I. 1993. How reliable are change-corrected measures of agreement? *Statistics in Medicine* 12(23):2191–2205.

Henderson, M.; Thomson, B.; and Williams, J. D. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 263–272.

IMARC Group. 2019. Intelligent virtual assistant market: Global industry trends, share, size, growth, opportunity and forecast 2019-2024. Technical report.

Khatri, C.; Hedayatnia, B.; Venkatesh, A.; Nunn, J.; Pan, Y.; Liu, Q.; Song, H.; Gottardi, A.; Kwatra, S.; Pancholi, S.; et al. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv preprint arXiv:1812.10757*.

Kim, B.; Ryu, S.; and Lee, G. G. 2017. Two-stage multi-intent detection for spoken language understanding. *Multimedia Tools and Applications* 76(9):11377–11390.

Laflen, A., and Fiorenza, B. 2012. "okay, my rant is over": The language of emotion in computer-mediated communication. *Computers and Composition* 29(4):296–308.

Lee, M. K.; Kiesler, S.; and Forlizzi, J. 2010. Receptionist or information kiosk: How do people talk with a robot? In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 31–40. ACM.

Littlejohn, S. W., and Foss, K. A. 2009. *Encyclopedia of communication theory*, volume 1. Sage.

Lowe, R. T.; Pow, N.; Serban, I. V.; Charlin, L.; Liu, C.-W.; and Pineau, J. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse* 8(1):31–65.

Madhavan, P.; Wiegmann, D. A.; and Lacson, F. C. 2006. Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48(2):241–256.

Makatchev, M.; Lee, M. K.; and Simmons, R. 2009. Relating initial turns of human-robot dialogues to discourse. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 321–322. ACM.

Mottet, T. P.; Martin, M. M.; and Myers, S. A. 2004. Relationships among perceived instructor verbal approach and avoidance relational strategies and students' motives for communicating with their instructors. *Communication Education* 53(1).

Ng, H.-W.; Nguyen, V. D.; Vonikakis, V.; and Winkler, S. 2015. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 443–449. ACM.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Ram, A.; Prasad, R.; Khatri, C.; Venkatesh, A.; Gabriel, R.; Liu, Q.; Nunn, J.; Hedayatnia, B.; Cheng, M.; Nagar, A.; et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Roy, S.; Mariappan, R.; Dandapat, S.; Srivastava, S.; Galhotra, S.; and Peddamuthu, B. 2016. Qart: A system for real-time holistic quality assurance for contact center dialogues. In *AAAI*, 3768–3775.

Serban, I. V.; Lowe, R.; Henderson, P.; Charlin, L.; and Pineau, J. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse* 9(1):1–49.

Valdivia, A.; Luzón, M. V.; and Herrera, F. 2017. Sentiment analysis in tripadvisor. *IEEE Intelligent Systems* 32(4):72–77.

Vinyals, O., and Le, Q. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Wilson, S. R. 2002. *Seeking and resisting compliance: Why people say what they do when trying to influence others*. Sage Publications.

Xu, P., and Sarikaya, R. 2013. Exploiting shared information for multi-intent natural language sentence classification. In *INTERSPEECH*, 3785–3789.

Zechner, K., and Waibel, A. 2000. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, 186–193. Association for Computational Linguistics.