Exploring Conversational Language Generation for Rich Content about Hotels

Marilyn A. Walker[1], **Albry Smither**[2], **Shereen Oraby**[1], **Vrindavan Harrison**[1], **Hadar Shemtov**[3]
University of California Santa Cruz[1], Google Content Studio[2], and Google Research[3]

Abstract

Dialogue systems for hotel and tourist information have typically simplified the richness of the domain, focusing system utterances on only a few selected attributes such as price, location and type of rooms. However, much more content is typically available for hotels, often as many as 50 distinct instantiated attributes for an individual entity. New methods are needed to use this content to generate natural dialogues for hotel information, and in general for any domain with such rich complex content. We describe three experiments aimed at collecting data that can inform an NLG for hotels dialogues, and show, not surprisingly, that the the sentences in the original written hotel descriptions provided on webpages for each hotel are stylistically not a very good match for conversational interaction. We quantify the stylistic features that characterize the differences between the original textual data and the collected dialogic data. We plan to use these in stylistic models for generation, and for scoring retrieved utterances for use in hotel dialogues.

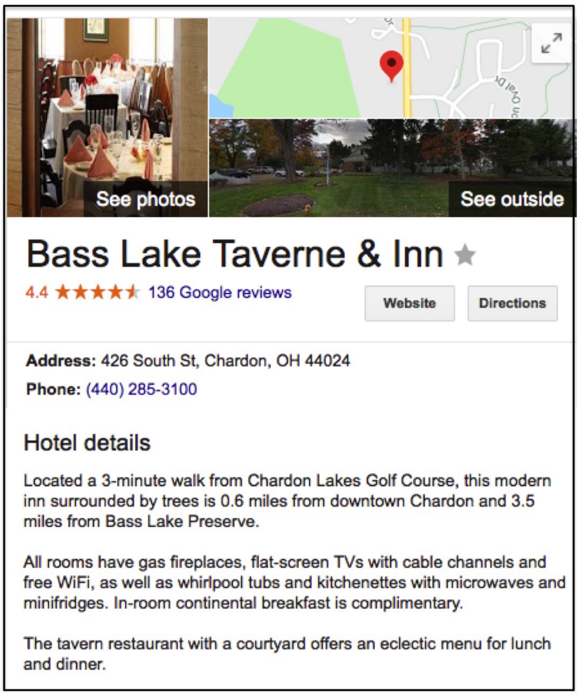
KEYWORDS: dialogue, conversation, natural language generation, hotels domain.

1. Introduction

Research and advanced development labs in both industry and academia are actively building a new generation of conversational assistants, to be deployed on mobile devices or on in-home smart speakers, such as Google Home. None of these conversational assistants can currently carry on a coherent multi-turn conversation in support of a complex decision task such as choosing a hotel, where there are many possible options and the user's choice may involve making trade-offs among complex personal preferences and the pros and cons of different options.

For example, consider the hotel description in the InfoBox in Figure 1, the search result for the typed query *"Tell me about Bass Lake Taverne"*. These descriptions are written by human writers within Google Content Studio and cover more than 200 thousand hotels worldwide. The descriptions are designed to provide travelers with quick, reliable and accurate information that they may need when making booking decisions, namely a hotel's amenities, property, and location. The writers implement many of the decisions that a dialogue system would have to make: they make decisions about content selection, content structuring, attribute groupings and the final realization of the content (Rambow and Korelsky, 1992). They access multiple sources of information, such as user reviews and the hotels' own web pages. The descriptions cannot be longer than 650 characters and are optimized for visual scanning. There is currently no method for delivering this content to users via a conversation other than reading the whole InfoBox aloud, or reading individual sections of it.

Structured data is also available for each hotel, which includes information about the setting of a hotel and its grounds, the feel of the hotel and its rooms, points of interest nearby, room features, and amenities such as restaurants and swimming pools. Sample structured data for the Bass Lake Taverne is in Figure 2.[1] The type of information available in the structured data varies a great deal according to the type of hotel: for specialized hotels it includes highly distinctive low-frequency attributes for look-and-feel such as "feels swanky" "historical rooms" or amenities such as "direct access to beach", "has hot tubs", or "ski-in, ski-out".



Bass Lake Taverne & Inn ★

4.4 ★★★★★ 136 Google reviews

Address: 426 South St, Chardon, OH 44024

Phone: (440) 285-3100

Hotel details

Located a 3-minute walk from Chardon Lakes Golf Course, this modern inn surrounded by trees is 0.6 miles from downtown Chardon and 3.5 miles from Bass Lake Preserve.

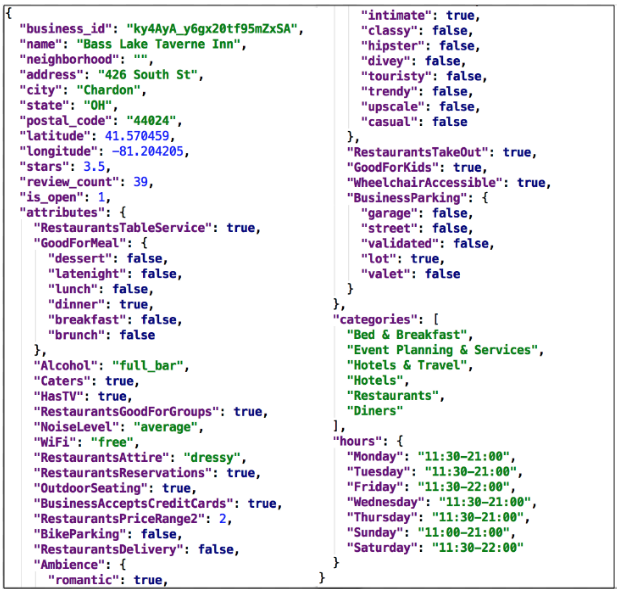
All rooms have gas fireplaces, flat-screen TVs with cable channels and free WiFi, as well as whirlpool tubs and kitchenettes with microwaves and minifridges. In-room continental breakfast is complimentary.

The tavern restaurant with a courtyard offers an eclectic menu for lunch and dinner.

Figure 1: InfoBox Hotel Description for Bass Lake Taverne

Research on dialogue systems for hotel information has existed for many years, in some cases producing shared dialogue corpora that include hotel bookings (Devillers et al., 2004; Walker et al., 2002; Rudnicky et al., 1999; Villaneau and Antoine, 2009; Bonneau-Maynard et al., 2006; Hastie et al., 2002; Lemon et al., 2006). Historically, these systems have greatly simplified the richness of the domain, and supported highly restricted versions of the hotel booking task, by limiting the information that the system can talk about to a small number of attributes, such as location, number of stars, room type, and price. Data collection involved users being given specific tasks where they simply had to find a hotel in a particular location, rather than satisfy the complex preferences that users may have booking hotels. This reduction in content simplifies the decisions

[1] The publicly available Yelp dataset[2] has around 8,000 entries for US hotels, providing around 80 unique attributes.



```
{
  "business_id": "ky4AyA_y6gx20tf95mZxSA",
  "name": "Bass Lake Taverne Inn",
  "neighborhood": "",
  "address": "426 South St",
  "city": "Chardon",
  "state": "OH",
  "postal_code": "44024",
  "latitude": 41.570459,
  "longitude": -81.204205,
  "stars": 3.5,
  "review_count": 39,
  "is_open": 1,
  "attributes": {
    "RestaurantsTableService": true,
    "GoodForMeal": {
      "dessert": false,
      "latenight": false,
      "lunch": false,
      "dinner": true,
      "breakfast": false,
      "brunch": false
    },
    "Alcohol": "full_bar",
    "Caters": true,
    "HasTV": true,
    "RestaurantsGoodForGroups": true,
    "NoiseLevel": "average",
    "WiFi": "free",
    "RestaurantsAttire": "dressy",
    "RestaurantsReservations": true,
    "OutdoorSeating": true,
    "BusinessAcceptsCreditCards": true,
    "RestaurantsPriceRange2": 2,
    "BikeParking": false,
    "RestaurantsDelivery": false,
    "Ambience": {
      "romantic": true,
      "intimate": true,
      "classy": false,
      "hipster": false,
      "divey": false,
      "touristy": false,
      "trendy": false,
      "upscale": false,
      "casual": false
    },
    "RestaurantsTakeOut": true,
    "GoodForKids": true,
    "WheelchairAccessible": true,
    "BusinessParking": {
      "garage": false,
      "street": false,
      "validated": false,
      "lot": true,
      "valet": false
    }
  },
  "categories": [
    "Bed & Breakfast",
    "Event Planning & Services",
    "Hotels & Travel",
    "Hotels",
    "Restaurants",
    "Diners"
  ],
  "hours": {
    "Monday": "11:30-21:00",
    "Tuesday": "11:30-21:00",
    "Friday": "11:30-22:00",
    "Wednesday": "11:30-21:00",
    "Thursday": "11:30-21:00",
    "Sunday": "11:00-21:00",
    "Saturday": "11:30-22:00"
  }
}
```

Figure 2: Sample of Hotel Structured Data for "Bass Lake Taverne"

that a dialogue manager has to make, and it also reduces the complexity of the natural language generator, since a few pre-constructed templates may suffice to present the small number of attributes that the system knows about. It is also important to note that the challenges for the hotel domain are not unique. Dialogue systems for movies, weather reports, real estate and restaurant information also have access to rich content, and yet previous work and current conversational assistants reduce this content down to just a few attributes.

This paper takes several steps toward solving the challenging problem of building a conversational agent that can flexibly deliver richer content in the hotels domain. Section 2. first reviews possible methods that could be applied, and describes several types of data collection experiments that can inform an initial design. After motivating these data collection experiments, the rest of the paper describes them and their results (Section 3., Section 4., and Section 5.). Our results show, not surprisingly, that both the hotel utterances and the complete dialogues that we crowdsource are very different in style than the original written InfoBox hotel descriptions. We compare different data collection methods and quantify the stylistic features that characterize their differences. The resulting corpora are available at nlds.soe.ucsc.edu/hotels.

2. Background and Experimental Overview

Current methods for supporting dialogue about hotels revolve either around search or around using a structured dialogue flow. Neither of these methods on their own support fully natural dialogue, and there is not yet an architecture for conversational agents that flexibly combines unstructured information, such as that found in the InfoBox or in reviews or other textual forms, and structured information such as that in Figure 2.

Search methods could focus on the content in the current InfoBox, and carry out short (1-2 turn) conversations by applying compression techniques on sentences to make them more conversational (Andor et al., 2016; Krause et al., 2017b). For example, when asked "Tell me about Bass Lake Taverne", Google Home currently produces an utterance providing its location and how far it is from the user's location. When asked about hotels in a location, Google Home reads out parts of the information in the Infobox, but it does not engage in further dialogue that explores individual content items. Moreover, the well-known differences between written and oral language (Biber, 1991) means that selected spans from written descriptions may not sound natural when spoken in conversation, and techniques may be needed to adapt the utterance to the dialogic context. Our first experiment, described in Section 3. asks crowdworkers to (1) indicate which sentences in the InfoBox are most important, and (2) write dialogic paraphrases for the selected sentences in order to explore some of these issues.

Another approach is to train an end-to-end dialogue system for the hotels domain using a combination of simulation, reinforcement learning and neural generation methods (Nayak et al., 2017; Shah et al., 2018; Liu et al., 2017; Gašić et al., 2017). This requires first developing a user-system simulation to produce simulated dialogues, crowdsourcing utterances for each system and user turn, and then using the resulting data to (1) optimize the dialogue manager using reinforcement learning, (2) train the natural language understanding from the user utterances, and (3) train the natural language generation from the crowd-sourced system utterances. Currently however it is not clear how to build a user-system simulation for the hotels domain that would allow more of the relevant content to be exchanged, and there are no corpora available with example dialogue flows and generated utterances.

To build a simulation for such complex, rich content, we first need a model for how the dialogue manager (DM) should (1) order the content across turns, and (2) select and group the content in each individual turn. Our assumption is that the most important information should be presented earlier in the dialogue, so one way to do this is to apply methods for inducing a *ranking* on the content attributes. Previous work has developed a model of user preferences to solve this problem (Carenini and Moore, 2000), and shown that users prefer systems whose dialogue behaviors are based on such customized content selection and presentation (Stent et al., 2002; Polifroni et al., 2003; Walker et al., 2007). These preferences (ranking on attributes) can be acquired directly from the user, or can be inferred from their past behavior. Here we try two other methods. First, in Section 3., we ask Turkers to select the most important sentence from the InfoBox descriptions. We then tabulate which attributes are in the selected sentences, and use this to induce a ranking. After using this tabulation to collect additional conversational utterances generated from meaning representations (Section 4.), we carry out an additional experiment (Section 5.) where we collect whole dialogues simulating the exchange of information between a user and a conversational agent, given particular attributes to be communicated. We report how information is ordered and grouped across these dialogues.

An end-to-end training method also needs a corpus for training for the Natural Language Generator (NLG). Thus we also explore which crowdsourcing design yields the best conversational utterance data for training the NLG. Our first experiment yields conversationalized paraphrases that match the information in individual sentences in the original Infobox. Our second experiment (Section 4.) uses content selection preferences inferred from the paraphrase experiment and collects utterances generated to match meaning representations. Our third experiment (Section 5.), crowdsources whole dialogues for selected hotel attributes: the utterances collected using this method are sensitive to the context while the other two methods yield utterances that can be used out of context.

To measure how conversational our collected utterances are, we build on previous research that counts linguistic features that vary across different situations of language use (Biber, 1991), and tabulates the effect of variables like the mode of language as well as its setting. We use the linguistic features tabulated by the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et al., 2015). See Table 1. We select features to pay attention to using the counts provided with the LIWC manual that distinguish natural speech (Column 4) from articles in the New York Times (Column 5). Our hotel descriptions are not an exact genre match to the New York Times, but they are editorial in nature. For example, Table 1 shows that spoken conversation has shorter, more common words (Sixltr), more function words, fewer articles and prepositions, and more affective and social language.

Category	Abbrev	Examples	Speech	NYT
Summary Language Variables				
Words/sentence	WPS	-	-	21.9
Words >6 letters	Sixltr	-	10.4	23.4
Linguistic Dimensions				
Total function words	funct	it, to, no, very	56.9	42.4
Total pronouns	pronoun	I, them, itself	20.9	7.4
Personal pronouns	ppron	I, them, her	13.4	3.6
1st pers singular	i	I, me, mine	7.0	.6
2nd person	you	you, your, thou	4.0	.3
Impersonal pronouns	ipron	it, it's, those	7.5	3.8
Articles	article	a, an, the	4.3	9.1
Prepositions	prep	to, with, above	10.3	14.3
Auxiliary verbs	auxverb	am, will, have	12.0	5.1
Common Adverbs	adverb	very, really	7.7	2.8
Conjunctions	conj	and, but, whereas	6.2	4.9
Negations	negate	no, not, never	2.4	.6
Common verbs	verb	eat, come, carry	21.0	10.2
Psychological Processes				
Affective processes	affect	happy, cried	6.5	3.8
Social processes	social	mate, talk, they	10.4	7.6
Cognitive processes	cogproc	cause, know, ought	12.3	7.5
Other				
Affiliation	affiliation	friend, social	2.0	1.7
Present focus	focuspresent	today, is, now	15.3	5.1
Informal language	informal	-	7.1	0.3
Assent	assent	agree, OK, yes	3.3	0.1
Nonfluencies	nonflu	er, hm, umm	2.0	0.1
Fillers	filler	Imean, youknow	0.5	0.0

Table 1: LIWC Categories with Examples and Differences between Natural Speech and the New York Times

The experiments use Turkers with a high level of qualification and we ensure that Turkers make at least minimum wage on our tasks. For the paraphrase and single-turn HITs for properties and rooms, we ask for at least 90% approval rate and at least 100 (sometimes 500) HITs approved and we always do location restriction (English speaking locations). For the dialog HITs we paid 0.9 per HIT, and restricted Turkers to those with a 95% acceptance rate and at least 1000 HITs approved. We also elicited the dialogs over multiple rounds and excluded Turkers who had failed to include all 10 attributes on previous HITs.

We present a summary of all of our experiments in Table 3 and then discuss the relevant columns in each section. A scan of the whole table is highly informative however, because Biber (1991) makes the point that differences across language use situations are not dichotomous, i.e. there is not one kind of oral language and one kind of written language. Rather language variation occurs continuously and on a scale, so that language can be "more or less" conversational. The overall results in Table 3 demonstrates this scalar variation, with different methods resulting in more or less conversationalization of the content in each utterance.

3. Paraphrase Experiment

The overall goal of the Paraphrase experiment is to evaluate the differences between monologic and dialogic content that contain the same or similar information. These experiments are valuable because the original content is given in unordered lists that facilitate visual scanning, as opposed to a conversation in which the dialogue system needs to decide the order in which to present information and whether to leave some information out.

We ask Turkers to both select "the most important" content out of the hotel descriptions, and then to paraphrase that content in a conversational style. We use this data to induce an importance ranking on content and we also measure how the conversational paraphrases of that content differ from the original phrasing. We used a randomly selected set of 1,000 hotel descriptions from our corpus of 200K, with instructions to Turkers to:

- Select the sentence out of the description that has the most important information to provide in response to a user query to "tell me about HOTEL-NAME".
- Cut and paste that sentence into the "Selected Sentence" box.
- Rewrite your selected sentence so that it sounds conversational, as a turn in dialogue. You may need to reorder the content or convert your selected sentence to multiple sentences in order to make it sound natural.

For each of the descriptions, three Turkers performed this HIT, yielding a total of 3,000 triples consisting of the original description, the selected sentence, and the human-generated dialogic paraphrases. For example, for the hotel description in Figure 3, two Turkers selected S1 and the other selected S4. These sentences have different content, so for each attribute realized we increase its count as part of our goal to induce a ranking indicating the importance of different attributes. The dialogic paraphrases the same Turkers produced are shown in Figure 4. The paraphrases contain fewer words per sentence, more use of anaphora, and more use of subjective phrases taking the listener's perspective such as *you will enjoy*.

S1 The elegant rooms, decorated in warm tones, feature high ceilings and lots of natural light, plus Turkish marble bathrooms, Bose sound systems, HDTVs and designer toiletries; some have views of the park.

S2 Suites include living rooms and soaking tubs; some have city views.

S3 Grand suites offer personal butler service.

S4 Open since 1930, this opulent landmark sits across the street from Central Park on New York's famed 5th Avenue.

Figure 3: An InfoBox description for the hotel *The Pierre, A Taj Hotel, New York*, split into sentences and labeled.

T1 This hotel's elegant rooms are decorated in warm tones. They feature high ceilings with lots of natural light. The rooms feature Turkish marble bathrooms, designer toiletries, high-definition televisions and Bose sound systems. Some rooms even offer views of the park.

T2 Located on 5th Avenue, this landmark hotel is located across the street from Central Park and dates back to 1930.

T3 Each room is elegantly decorated in warm tones. You will enjoy high ceilings and natural light. The bathrooms are done in Turkish marble and have designer toiletries. For entertainment, you will find HDTVs and Bose sound systems. There are views of Central Park from some rooms.

Figure 4: Turker generated paraphrases of the hotel description shown in Table 3. The Turkers T1 and T3 selected S1 as containing the most important information and Turker T2 selected S4.

attribute	F
locale_mountain	1.0
has_bed_wall_in_rooms	.67
has_wet_room	.67
feels_quaint	.61
has_crib	.50
feels_artsy	.44
is_whitewashed	.44
has_private_bathroom_outside_room	.44
feels_nautical	.42
has_luxury_bedding	.40
welcomes_children	.39
is_dating_from	.38
feels_retro	.38
all_inclusive	.34
has_casino	.33
has_heated_floor	.33
has_city_views	.33
has_boardwalk	.33
has_hammocks	.33
has_onsite_barbecue_area	.33

Table 2: Turker's Top 20 Attributes, shown with their frequency F of selection when given in the content.

Results. We build a ranked ordering of hotel attribute importance using the selected sentences from each hotel description. We count the number of times each attribute is realized within a sentence selected as being the most informative or relevant. We count the number of hotels for which each attribute applies. The attribute frequency F is given as the number of times an attribute is selected divided by the product of the number of hotels to which the attribute applies and the number of Turkers that were shown those hotel descriptions. Finally, the attributes are sorted and ranked by largest F.

Table 2 illustrates how the tabulation of the Turker's selected sentences provides information on the ranking of attributes that we can use in further experimentation. However, the frequencies reported are conditioned on the relevant attribute being available to select in the Infobox description, and many of the attributes are both low frequency and highly distinctive, e.g. the attribute `local_mountain`. A reliable importance ranking using this method would need a larger sample than 1000 hotels. It is also possible that attribute importance should be directly linked to how distinctive the attribute is, with less frequent attributes always mentioned earlier in the dialogue.

The first three columns of Table 3 summarize the stylistic differences between the original Infobox sentences and the collected paraphrases. Column 3 provides the p-values showing that many differences are statistically significant. Differences that indicate that the paraphrases are more similar to oral language (as in Speech, column 4 of Table 1), include the use of adverbs, words of affiliation, common verbs, and a reduced number of words per sentence. Expected differences that are not realized are in increases in affective and social language, reduced use of Articles and long words (SixLtr), greater use of conjunctions. So while this method improves the conversational style of the content realization, we will see that our other methods produce *more* conversational utterances. While this method is inexpensive and may not require such expert Turkers, the utterances collected may only be useful for systems that do not use structured data and so need paraphrases of the original Infobox data that is more conversational.

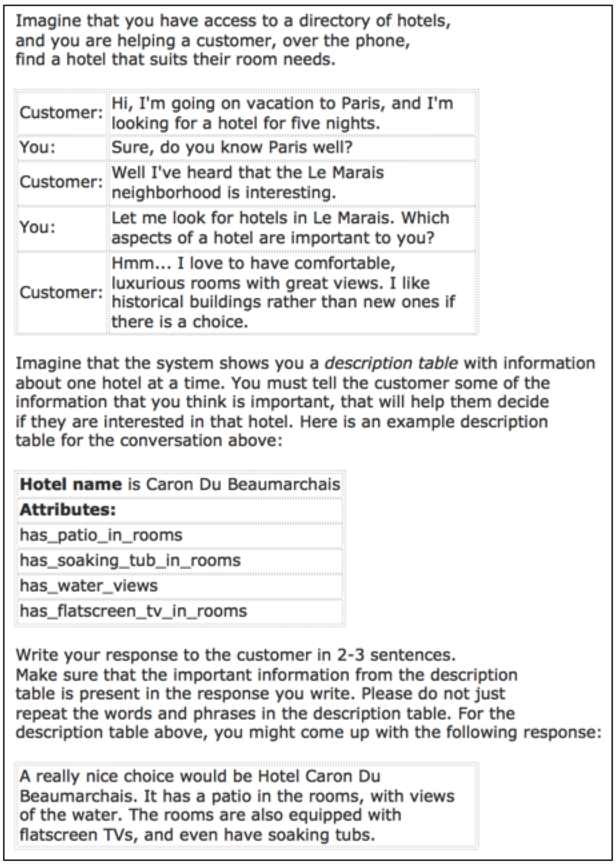
4. Generation from Meaning Representations

The second experiment aims to determine whether we get higher quality utterances if we ask crowdworkers to generate utterances directly from a meaning representation, in the context of a conversation, rather than by selecting from the original Infobox hotel descriptions. Utterances generated in this way should not be influenced by the original phrasing and sentence planning in the hotel descriptions.

Instructions for our second experiment are shown in Figure 5. Here we give Turkers specific content tables and ask them to generate utterances that realize that content. Note that the original hotel descriptions, as illustrated in Figure 1 consists of three blocks of content, property, rooms and amenities. For each hotel in a random selection of 200 hotels from the paraphrase experiment, we selected content for both rooms (4 attributes) and properties (6 attributes) by picking the attributes with the highest scores (as illustrated for a small set of attributes in Table 2). Thus hotel has two unique content tables assigned to it, one pertaining to the hotel's rooms, and the other for the hotel grounds. Each hotel content table is given to three Turkers which results

Category	InfoBox	Paraphrase	p-val	Props+Rooms	Dialogues	p-val Props+Rooms vs. Para	p-val Props+Rooms vs. Dial
Impersonal Pronouns	0.97	3.80	0.00	3.36	5.19	0.00	0.00
Adverbs	0.97	3.41	0.00	3.57	6.25	0.12	0.00
Affective Processes	4.98	4.81	0.14	8.09	8.55	0.00	0.26
Articles	8.08	9.06	0.00	11.54	7.62	0.00	0.00
Assent	0.02	0.04	0.00	0.07	1.13	0.11	0.00
Auxiliary Verbs	1.69	6.12	0.00	8.02	11.81	0.00	0.00
Common Verbs	3.64	7.94	0.00	10.97	15.07	0.00	0.00
Conjunctions	8.07	8.13	0.54	7.33	6.52	0.00	0.00
First Person Singular	0.01	0.02	0.00	0.41	3.41	0.00	0.00
Negations	0.03	0.07	0.00	0.27	0.44	0.00	0.00
Personal Pronouns	0.06	1.15	0.00	3.87	10.17	0.00	0.00
Second Person	2.31	0.45	0.00	2.43	5.63	0.00	0.00
Six Letter Words	4.81	19.15	0.00	20.74	15.50	0.00	0.00
Social Processes	16.24	5.63	0.00	8.53	14.66	0.00	0.00
Total Pronouns	1.03	4.94	0.00	7.23	15.36	0.00	0.00
Words Per Sentence	22.86	14.69	0.00	14.52	10.90	0.00	0.00
Affiliation	1.18	0.95	0.00	1.13	5.97	0.00	0.00
Cognitive Processes	2.58	3.11	0.00	9.18	10.47	0.00	0.00
Focus present	3.64	7.91	0.00	9.35	14.40	0.00	0.00
Function	26.79	37.62	0.00	44.58	53.08	0.00	0.00
Informal	0.42	0.35	0.03	0.51	1.76	0.00	0.00
nonflu	0.37	0.28	0.00	0.42	0.63	0.00	0.00
prep	9.65	8.50	0.00	9.83	9.88	0.00	0.72

Table 3: Conversational LIWC features across all Utterance Types/Data Collection Methods



Imagine that you have access to a directory of hotels, and you are helping a customer, over the phone, find a hotel that suits their room needs.

Customer:	Hi, I'm going on vacation to Paris, and I'm looking for a hotel for five nights.
You:	Sure, do you know Paris well?
Customer:	Well I've heard that the Le Marais neighborhood is interesting.
You:	Let me look for hotels in Le Marais. Which aspects of a hotel are important to you?
Customer:	Hmm... I love to have comfortable, luxurious rooms with great views. I like historical buildings rather than new ones if there is a choice.

Imagine that the system shows you a *description table* with information about one hotel at a time. You must tell the customer some of the information that you think is important, that will help them decide if they are interested in that hotel. Here is an example description table for the conversation above:

Hotel name is Caron Du Beaumarchais
Attributes:
has_patio_in_rooms
has_soaking_tub_in_rooms
has_water_views
has_flatscreen_tv_in_rooms

Write your response to the customer in 2-3 sentences. Make sure that the important information from the description table is present in the response you write. Please do not just repeat the words and phrases in the description table. For the description table above, you might come up with the following response:

A really nice choice would be Hotel Caron Du Beaumarchais. It has a patio in the rooms, with views of the water. The rooms are also equipped with flatscreen TVs, and even have soaking tubs.

Figure 5: Instructions for Room Attributes HIT

Prop	A good choice is 1 Hotel South Beach in Miami Beach. It's luxurious, lively, upscale, and chic, with beach access and a bar onsite.
Prop	I think that 1 Hotel South Beach will meet your needs. It's a chic luxury hotel with beach access and a bar. Very lively.
Room	One of the excellent hotels Miami Beach has to offer is the 1 Hotel South Beach. The upgraded rooms are full featured, including a kitchen and a desk for work. Each room also has a balcony.
Room	The 1 Hotel South Beach doesn't mess around. When you come to stay here you won't want to leave. Each upgraded room features a sunny balcony and personal kitchen. You also can expect to find a lovely writing desk for your all correspondence needs.

Figure 6: Example utterances generated by Turkers in the second experiment. Turkers were given specific content tables from which to generate dialogue utterances that realize that content.

in a total of 1,200 utterances collected. Turkers were instructed to create utterances as conversational as possible.

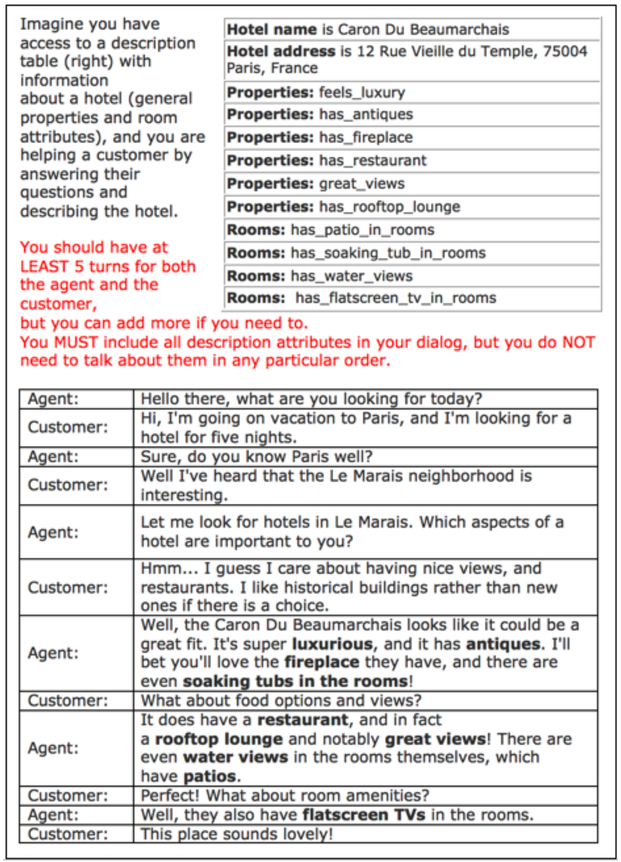
Results. Sample utterances for both properties and rooms are shown in Figure 6.

Column 5 in Table 3 shows the frequencies of LIWC's conversationalization features for the utterances collected in this experiment, and Column 7 reports statistical significance (p-values) for comparing these collected utterances to the paraphrases collected in Experiment 1, using an unpaired t-test on the two datasets. We can see that some of the attributes that indicate conversationalization indicate that this method yields more conversational utterances: there are significantly more more auxiliary verbs and common verbs. There is a greater use of first person and second person pronouns, as well as words indicating affective, social and cognitive processes. Counts of function words and focusing on the present are also higher as would be expected of more conversational language.

5. Dialogue Collection Experiment

The final data collection experiment focuses on utterance generation in an explicitly dialogic setting. In order to collect dialogues about our hotel attributes, we employ a technique called "self-dialogue" collection, which to our knowledge was pioneered by Krause et al. (2017a), who claim that the results are surprisingly natural. We ask individual Turkers to write a full dialogue between an agent and a customer, where the Turker writes both sides of the dialogue. The customer is looking for a hotel for a trip, and the agent has access to a description table with a list of 10 attributes for a single hotel. The agent is tasked with describing the hotel to the customer. Figure 7 shows our HIT instructions that provided a sample dialogue as part of the instructions to the Turkers.

This experiment utilizes 74 unique hotels, a subset of those used in the property and room experiments above (Section 4.), where we have both 6 property attributes, and 4 room attributes. We aimed to collect 3 dialogues per hotel (from unique Turkers), but due to some Turkers failing to follow instructions, the final corpus consists of 205 dialogues (comprised of 58 hotels with 3 dialogues each, 15 hotels with 2 dialogs each, and 1 hotel with only 1 dialogue).



Imagine you have access to a description table (right) with information about a hotel (general properties and room attributes), and you are helping a customer by answering their questions and describing the hotel.

Hotel name is Caron Du Beaumarchais
Hotel address is 12 Rue Vieille du Temple, 75004 Paris, France
Properties: feels_luxury
Properties: has_antiques
Properties: has_fireplace
Properties: has_restaurant
Properties: great_views
Properties: has_rooftop_lounge
Rooms: has_patio_in_rooms
Rooms: has_soaking_tub_in_rooms
Rooms: has_water_views
Rooms: has_flatscreen_tv_in_rooms

You should have at LEAST 5 turns for both the agent and the customer,
but you can add more if you need to.
You MUST include all description attributes in your dialog, but you do NOT need to talk about them in any particular order.

Agent:	Hello there, what are you looking for today?
Customer:	Hi, I'm going on vacation to Paris, and I'm looking for a hotel for five nights.
Agent:	Sure, do you know Paris well?
Customer:	Well I've heard that the Le Marais neighborhood is interesting.
Agent:	Let me look for hotels in Le Marais. Which aspects of a hotel are important to you?
Customer:	Hmm... I guess I care about having nice views, and restaurants. I like historical buildings rather than new ones if there is a choice.
Agent:	Well, the Caron Du Beaumarchais looks like it could be a great fit. It's super **luxurious**, and it has **antiques**. I'll bet you'll love the **fireplace** they have, and there are even **soaking tubs in the rooms**!
Customer:	What about food options and views?
Agent:	It does have a **restaurant**, and in fact a **rooftop lounge** and notably **great views**! There are even **water views** in the rooms themselves, which have **patios**.
Customer:	Perfect! What about room amenities?
Agent:	Well, they also have **flatscreen TVs** in the rooms.
Customer:	This place sounds lovely!

Figure 7: Instructions for Hotel Dialog HIT

Figure 4 and Figure 5 provide sample dialogues from the corpus, with the 10 required attributes shown in bold, and the agent (A) and customer (C) turns shown with their respective turn numbers. In Dialogue 1, we see an example of a creative dialogue where the Turker designs a situational context for the dialogue where the customer is looking for a hotel for a bachelorette party weekend, and has specific requirements about flooring and amenities. We note that in this dialogue, the agent only begins to discuss the hotel in their third turn. In Dialogue 2, we see a much more basic dialogue, where the agent begins to list properties and room attributes earlier on the in dialogue (at Agent Turn 2), and the full list of attributes is exhausted halfway through the conversation, at Agent Turn 3.

A1	Hi! How can I help you today?
C1	I am planning a trip to New York for a bachelorette party weekend and need help finding a hotel.
A2	OK, what will you girls be planning to do?
C2	We're going to a Broadway show but other than that just going to dinner and hitting some bars.
A3	OK, I think the Hotel Indigo in the Chelsea section would be great! It's **upscale** and has a great **hip** and **contemporary** vibe with that **buzzy** New York City energy feel.
C3	That sounds like what we're looking for. I know this is a weird question but one of the girls sometimes has an allergy to carpet, is there any type of option for a non carpeted room?
A4	Actually, this hotel has **hardwood floors in the rooms**.
C4	Great! I think we may be bringing some snacks and maybe some of our own alcohol. Can we arrange for a fridge or do they just have ice buckets?
A5	The rooms have **mini bars** as well as **coffee** if you girls need some help waking up for your time out. There is also a **bar** on site so you can start the party before you even head out.
C5	Great! One more question, one of the girls does need to keep in touch with work. Do you offer WiFi?
A6	The hotel has **desks in every room** and offers a **business center** if she needs anything like a printer or a desktop computer.
C6	I think we'll go ahead and book this. It sounds perfect!

Table 4: Situational Context for Content Hotel Dialogue

A1	Good evening, how can i help you?
C1	I am looking for a good hotel to have a business conference in the Brooklyn area.
A2	Sure, let me see what i can find. Hotel Indigo Brooklyn may be just what you are looking for. It has a **hip** feel with an **onsite bar**, **Business center**, **restaurant**, **free wifi**. Its got it all.
C2	That sounds excellent. What room amenities are offered?
A3	There is **coffee** in the rooms and a **mini fridge**. All the rooms have been recently **upgraded** and did i mention it has a **fitness room**? I has full **room service** as well.
C3	Wow, that sounds great.Whats the address? I need to make sure its in the right area for me.
A4	Sure, its 229 Duffield Street, Brooklyn, NY 11201, USA.
C4	Thanks, thats just the right spot. Go ahead and make me a reservation for next Tuesday.
A5	excellent! Its done!
C5	Thanks you have been extremely helpful!

Table 5: Straightforward Attribute Listing Hotel Dialogue

Results. We begin by analyzing information that both the dialogue manager and the natural language generator would need to know, namely how frequently attributes are grouped in a single turn in our collected dialogues, by counting the number of times certain keywords are mentioned related to the attributes in the dialogues. Table 6 shows attributes groups that occur at least 4 times in the dataset, showing the group of attributes and the frequency count. We note that the attributes within the groups are generally either: 1) semantically similar, e.g. "modern" and "contemporary"; 2) describe the same aspect, e.g. "feels elegant" and "feels upscale"; or 3) describe the same general attribute, e.g. "has breakfast buffet", "has free breakfast", and "has free breakfast buffet". It is interesting to note that the semantic similarity is not always completely obvious (for example, "has balcony in rooms" and "has fireplace" may be used to emphasize more luxurious amenities that are a rare find).

Attribute Group	Count
(has_business_center, has_meeting_rooms)	13
(has_bar_onsite, has_restaurant)	9
(feels_contemporary, feels_modern)	9
(has_bar_onsite, has_business_center)	8
(has_business_center, has_desk_in_rooms)	7
(feels_casual, feels_contemporary, feels_modern)	6
(feels_modern, has_business_center)	6
(has_business_center, has_convention_center)	6
(feels_elegant, feels_upscale)	4
(has_bar_onsite, has_bar_poolside)	4
(has_microwave_in_rooms, has_minifridge_in_rooms)	4
(feels_contemporary, feels_elegant, feels_modern)	4
(feels_contemporary, feels_upscale)	4
(has_balcony_in_rooms, has_fireplace)	4
(feels_chic, feels_upscale)	4
(has_business_center, has_desk_in_rooms, has_wi_fi_free)	4
(has_breakfast_buffet, has_free_breakfast, has_free_breakfast_buffet)	4
(has_coffee_in_rooms, has_desk_in_rooms, has_microwave_in_rooms, has_minifridge_in_rooms)	4

Table 6: Attributes Frequently Grouped in a Single Turn

Our assumption is that more important attributes should be presented earlier in the dialogue, and that a user-system dialogue simulation system design (Shah et al., 2018; Liu et al., 2017; Gašić et al., 2017) would require such information to be available. Thus, in order to provide more information on the importance of particular attributes, we analyze where in the conversation (i.e. first or second half) certain types of attributes are mentioned. For example, we observe that attributes describing the "feel", such as "feels chic" or "feels upscale", are mentioned around 700 times, and that for 80% of those times they appear in the first half of the conversation as opposed to the second half, showing that they are often used as general hotel descriptors before diving into detailed attributes. Attributes describing room amenities on the other hand, such as "has kitchen in rooms" or "has minifridge", were mentioned around 530 times, with a more even distribution of 53% in the first half of the conversation, and 47% in the second half.

We also observe that most attributes are first introduced into the conversation by the agent, but that a small number of attributes are more frequently first introduced by the customer, specifically: *has_swimming_pool_indoor, popular_with_business_travelers, has_onsite_laundry, welcomes_families, has_convention_center, has_ocean_view, has_free_breakfast_buffet, has_swimming_pool_saltwater*.

Next, we compare our collected dialogues to the single-turn dialogue descriptions described in Section 4.). Specifically, we focus on the "agent" turns of our dialogues, as they are more directly comparable to the property and room turns. Table 7 describes the average number of turns, number of sentences per turn, words per turn, and attributes per turn across the property, room, and agent dialogue turns. We note that the average number of sentences, words, and attributes per turn for our property and room descriptions are higher in general than the agent turns in our dialogues, because the dialogues allow the agent to distribute the required content across multiple turns.

	Properties	Rooms	Dialogues
Number of turns	600	600	1227
Sentences per turn			
Average	2.80	2.55	1.80
Words per turn			
Average	41.37	39.81	21.45
Attributes per turn			
Average	6	4	1.62

Table 7: Comparing Property, Room, and Agent Dialogue Turns

Column 6 (Dialogues) of Table 3 reports the frequencies for conversational features in the collected data, with p-values in Column 8 comparing the dialogic utterances to the property+room utterances collected in Experiment 2 (Section 4.. The dialogic data collection results in utterances that are more conversational according to these counts, with higher use of impersonal pronouns and adverbs, auxiliary verbs and common verbs, and first person and second person pronouns. We also see increases in words indicative of affective, social and cognitive processes, more informal language, and reduced use of Six Letter words, fewer words per sentence and greater use of language focused on the present. Thus these utterances are clearly much more conversational, and provide information on attribute ordering across turns as well as possible ways of grouping attributes. The utterances collected in this way might also be useful for template induction, especially if the induced templates could be indexed for appropriate use in context.

6. Conclusion and Future Work

This paper presents a new corpus that contributes to defining the requirements and provide training data for a conversational agent that can talk about all the rich content available in the hotel domain. All of the data we collect in all of the experiments is available at `nlds.soe.ucsc.edu/hotels`. After completing three different types of data collection, we posit that the self-dialogue collection might produce the best utterances but at the highest cost, with the most challenges for direct re-use. The generation from meaning representations produces fairly high quality utterances, but they are not sensitive to the context, and our results from the dialogic collection suggest that it might be useful to collect additional utterances using this method that sample different combinations of attributes, and select fewer attributes for each turn. In future work, we plan to use these results in three different ways. First, we can train a "conversational style ranker" based on the data we collected, so that it can retrieve pre-existing utterances that have good conversational properties. The features that this ranker will use are the linguistic features we have identified so far, as well as new features we plan to develop related to context. Second, we will experiment directly using the collected utterances in a dialogue system, first by templatizing them by removing specific instantiations of attributes, and then indexing them for their uses in particular contexts.

7. References

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2442–2452.

Biber, D. (1991). *Variation across speech and writing*. Cambridge Univ Pr.

Bonneau-Maynard, H., Ayache, C., Bechet, F., Denis, A., Kuhn, A., Lefèvre, F., Mostefa, D., Quignard, M., Rosset, S., Servan, C., et al. (2006). Results of the french evalda-media evaluation campaign for literal understanding. In *The fifth international conference on Language Resources and Evaluation (LREC 2006)*.

Carenini, G. and Moore, J. (2000). A strategy for generating evaluative arguments. In *Proc. of the 1st International Conference on Natural Language Generation (INLG-00)*, Mitzpe Ramon, Israel.

Devillers, L., Maynard, H., Rosset, S., Paroubek, P., McTait, K., Mostefa, D., Choukri, K., Charnay, L., Bousquet, C., Vigouroux, N., et al. (2004). The french media/evalda project: the evaluation of the understanding capability of spoken language dialogue systems. In *LREC*.

Gašić, M., Hakkani-Tür, D., and Celikyilmaz, A. (2017). Spoken language understanding and interaction: machine learning for human-like conversational systems. *Computer Speech and Language 46*, pages 249 – 251.

Hastie, H. W., Prasad, R., and Walker, M. A. (2002). Automatic evaluation: Using a date dialogue act tagger for user satisfaction and task completion prediction. In *LREC*.

Kittredge, R., Korelsky, T., and Rambow, O. (1991). On the need for domain communication knowledge. *Computational Intelligence*, 7(4):305–314.

Krause, B., Damonte, M., Dobre, M., Duma, D., Fainberg, J., Fancellu, F., Kahembwe, E., Cheng, J., and Webber, B. (2017a). Edina: Building an open domain socialbot with self-dialogues. *arXiv preprint arXiv:1709.09816*.

Krause, S., Kozhevnikov, M., Malmi, E., and Pighin, D. (2017b). Redundancy localization for the conversationalization of unstructured responses. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 115–126.

Lavoie, B. and Rambow, O. (1997). A fast and portable realizer for text generation systems. In *Proc. of the Third Conference on Applied Natural Language Processing, ANLP97*, pages 265–268.

Lemon, O., Georgila, K., and Henderson, J. (2006). Evaluating effectiveness and portability of reinforcement learned dialogue strategies with real users: the talk towninfo evaluation. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 178–181. IEEE.

Liu, B., Tur, G., Hakkani-Tur, D., Shah, P., and Heck, L. (2017). End-to-end optimization of task-oriented dialogue model with deep reinforcement learning. *arXiv preprint arXiv:1711.10712*.

Nayak, N., Hakkani-Tur, D., Walker, M., and Heck, L. (2017). To plan or not to plan? discourse planning in slot-value informed sequence to sequence models for language generation. In *Proc. of Interspeech 2017*.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Polifroni, J., Chung, G., and Seneff, S. (2003). Towards the automatic generation of mixed-initiative dialogue systems from web content. In *EUROSPEECH, European Conference on Speech Processing*.

Rambow, O. and Korelsky, T. (1992). Applied text generation. In *Proc. of the Third Conference on Applied Natural Language Processing, ANLP92*, pages 40–47.

Rudnicky, A., Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K., Xu, W., and Oh, A. (1999). Creating natural dialogs in the carnegie mellon communicator system. In *Eurospeech*, pages 1531–1534.

Shah, P., Hakkani-Tür, D., and Heck, L. (2016). Interactive reinforcement learning for task-oriented dialogue management. In *NIPS 2016 Deep Learning for Action and Interaction Workshop*.

Shah, P., Hakkani-Tür, D., Tür, G., Rastogi, A., Bapna, A., Nayak, N., and Heck, L. (2018). Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Stent, A., Walker, M., Whittaker, S., and Maloor, P. (2002). User-tailored generation for spoken dialogue: An experiment. In *ICSLP*.

Villaneau, J. and Antoine, J.-Y. (2009). Deeper spoken language understanding for man-machine dialogue on broader application domains: a logical alternative to concept spotting. In *Proceedings of the 2nd Workshop on Semantic Representation of Spoken Language*, pages 50–57. Association for Computational Linguistics.

Walker, M., Rudnicky, A., Aberdeen, J., Bratt, E., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Prasad, R., Roukos, S., Sanders, G., Seneff, S., and Stallard, D. (2002). DARPA communicator evaluation: Progress from 2000 to 2001. In *ICSLP*.

Walker, M. A., Stent, A., Mairesse, F., and Prasad, R. (2007). Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research (JAIR)*, 30:413–456.