BobBot: Conversation Analysis for Use in Chat Bots

Dustin Fink and Rupa Shankar

Introduction

In this day and age, chat bots are a phenomenon that we have all encountered, to varying degrees of success and frustration. They have been around since the 1960s, when the classic ELIZA bot was published. Human conversation is complex and difficult to mimic, so how do these bots manage it? The early bots, such as ELIZA, operated by recognizing cue words or phrases and simply outputting pre-programmed responses. But nowadays, with the advent of Siri for example, more modern AI and specifically NLP techniques work best in attempting to model conversational flow.

Task Definition

Write a predictor that given a conversation and a set of prompt sentences, chooses appropriate response sentences based on learned weights. Incorporate the predictor into a chat bot.

Implementation

Corpora

Switchboard Dialogue Act Corpus

We used the Switchboard Dialogue Act Corpus, provided on-line by Christopher Potts in Stanford Linguistics, and annotated by linguistic graduates at UC Boulder. The corpus consists of 1,115 transcripts of **discourses** between two callers (A and B) talking over the phone about a pre-provided topic. The discourse is broken down into **utterances**, a phrase, sentence, or set of phrase and sentences spoken by one caller. Each utterance is annotated with its speaker, parts of speech, tree parse, and other symbols to represent inflection and noise.

Most importantly, utterances are labeled with a single **dialogue act tag (DAT)**, representing the function of that utterance in the conversation. These dialogue act tags are the subject of the paper "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech" (Stocke et al.), which discuses an algorithm for automatic identification and tagging of utterances. In using the Switchboard Dialogue Act Corpus, we build on top of this algorithm to analyze conversational features.

Table: Sample Dialogue Act Tags

sd	Statement-non-opinion	32.13%	Me, I'm in the legal department
b	Backchannel	15.85%	Uh-huh.
sv	Statement-opinion	11.37%	I think it's great.
+	Multi-utterance	7.96%	*
%	Uninterpretable	7.11%	But, uh, yeah
aa	Agree/Accept	4.82%	That's exactly it.
ba	Appreciation	2.01%	I can imagine.
qy	Yes-No-Question	1.74%	Do you have any special training?
x	Non-verbal	1.62%	<Laughter>
ny	Yes answers	1.27%	Yes.
fc	Conventional-closing	1.06%	Well, it's been nice talking to you.
h	Hedge	0.57%	I don't know if I'm making any sense.
^2	Collaborative Completion	0.31%	-- who aren't contributing.

* Multi-utterances generally consist of stutters, stops and restarts, and continuations of interrupted speech.

General info:

Number of Transcripts:	1155
Average Utterance Length:	9.91 words
Average Turn Length:	2.05 utterances
Average Difference in Turn Lengths of a Pair	1.35 utterances
Number of Turns:	106619

Open Subtitles Corpus

We initially considered the Open Subtitles corpus, which is an online corpus of movie subtitles, and wrote scripts to procure the data. However, the disadvantage of this corpus was that it did not have speaker information nor dialogue act tags. Although we used the Stanford CoreNLP tools to parse the sentences and get their part-of-speech tags as well as parse trees, the parse information and raw sentence data was not contextual information with which to design useful features. As we describe later, the dialogue act tags and speaker information is crucial to the implementation of our chat bot, so it is a good thing that we used the Switchboard Dialogue Act Corpus.

Model and Method Used

Examples
The discourses from the Switchboard Dialogue Act corpus were partitioned into a set of training (n=850), dev (n=115), and test discourses (n=115). Each discourse was then converted into a set of examples. An **example** x consist of **prompt-response pair** (p, r), where each prompt and response is a turn. A **turn** is a set of consecutive utterances spoken by one speaker in a discourse, and the set of turns in a discourse is the set of alternating turns between two speakers.

Every discourse in the corpus was converted into a set of turns, and then set of positive examples and negative examples. All prompt-response pairs that were actual consecutive turns between two speakers in a discourse were used as positive examples. Negative examples consisted of a prompt-response pair, where the response was randomly selected from the set of turns in the same discourse.

Random Selection
For negative examples, evaluation, and responses in the chat bot, we rely a fair amount on randomization. While testing and for final results, we used multiple seeds and averaged the results together. Though not perfect, this technique gave us good approximations of unlikely conversation.

Stochastic Gradient Descent
We used stochastic gradient descent with a set of features defined over prompt-response pairs to find our weights. The weights were used both as a predictor of whether an example was positive or negative, and also to evaluate possible response pairings with a given prompt.

$$
\begin{aligned}
&Let\ y = 1\ for\ positive\ examples,\\
&\quad y = -1\ for\ negative\ examples\\
&R = training\ set\ of\ (example, y)\\
&w \leftarrow \{0, \dots, 0\}\\
&for\ i\ in\ 1, .., NUM\_ITERATIONS\\
&\quad for\ x, y\ in\ R\\
&\qquad w \leftarrow w - STEP * \nabla_w Loss_{Hinge}(\phi(x), y, w)
\end{aligned}
$$

Predictor
$f_w(x) = sign(\phi(x) \cdot w)$

Score Function
$s_w(x) = \phi(x) \cdot w$

Features

We started out by designing more general features that would capture as much information as possible about the dialogue act tags, utterance length, and turn length.

Note: We will often refer to the prompt as "Turn A" and the response as "Turn B."

Dialogue act tags

We designed four features that capture the dialogue act tags of the last two utterances of turn A and the first two utterances of turn B.

Specifically, they pair up like so:
last of A, first of B
2nd last of A, 2nd last of B
2nd last of A, first of B
last of A, 2nd last of B

Of our other two relevant features, one contains all four of these tags and the other contains the set of A's tags and the set of B's tags.

Utterance length

We designed five features that capture utterance length information. For each of the following pairs, we captured the length of both utterances:

last of A, first of B
2nd last of A, 2nd of B
3rd last of A, 3rd of B
4th last of A, 4th of B
5th last of A, 5th of B

We also captured the lengths of the last two utterances of A and the first two of B in one feature together.

Turn length

We captured both the number of utterances in A vs B as well as the number of total words in A vs B.

Parse Tree

Many of the utterances come with parse tree information, so if that exists and is a perfect match, we traverse the tree to find the subject/s of both turn A and turn B.

Interruption

We wanted the features to capture information relating to interruptions, since as we will describe later, we wanted the chatbot to be able to interrupt. Our heuristic for this purpose was to assume that interruptions only occurred with backchannel phrases (e.g "uh-huh") and collaborative completion phrases, which is where the other person interrupts and chimes in if they think they know what the speaker is going to say, or to clarify a point (e.g "It's just hard to work with people when - " "who aren't contributing"). If the first utterance in turn B was one of these types, then we recorded the type. We also recorded the type along with the set of turn A's dialogue act tags, in an attempt to capture which types of act tags are more likely to precede an interruption.

Pairwise Features

In designing these more specific pairwise positive features, we attempted to capture various aspects of human conversation by thinking about common pairs of phrases e.g a yes/no question should be followed by a yes/no answer, or an open question would likely be followed by an opinion statement. Here is our final list of pairwise features:

Yes/no question followed by a yes/no response
Yes/no question followed by a "maybe" response
Yes/no question followed by an "other answer"
A non-yes/no question followed by a statement of any kind
An open question followed by an opinion statement
A summarization followed by a backchannel
A summarization followed by a positive response
A collaborative completion followed by a backchannel
A collaborative completion followed by a "maybe" response
A collaborative completion followed by a rejection
An apology followed by a downplayer (e.g that's okay)

These types of phrases were determined by the existing dialogue act tags.

Chat bot

Several possibilities for the chat bot were considered, including having the bot choose a response from a large fixed set of responses and having it choose from candidates randomly generated from a probabilistic context-free grammar. In the end, in order to take advantage of the dialogue act tags in the corpus, we chose to have the chat bot select a discourse, and choose responses from that discourse. Since this chat bot focuses on aspects of conversational structure without much regard to semantics, this also allowed the bot to stay roughly on topic, considering that each discourse in the corpus focused on a specific topic.

Interaction with the bot begins with a welcome message, a few setup options, and then a selection of discourse. It prints out the first few turns of the discourse, assuming the user to be caller A, and the bot to be caller B. The user is then allowed to speak first, and then the bot responds.

A user is asked to interact with the bot as if they were discussing the given topic over the phone. The use types in one utterance at a time. Because we do not have the dialogue act tagging algorithm, the use is asked to input an act_tag for their utterance. Once they have finished speaking, they hit enter to finish their turn.

Their utterances are then formulated as a turn. The bot randomly generates a set of candidates (consisting of all or a given number of turns by caller B). For each candidate, the bot uses is as a response to the user's prompt, and gets a score for that prompt-response pair. The bot then prints out the highest scoring response candidate along with the score.

Evaluation metric

Due to the subjective and nuanced nature of conversation, we developed several evalution metrics.

Evaluation Metric 1 (primary): Predictor Analysis
The primary evaluation metric was the performance of the predictor on the set of examples. For the training, dev, and test sets, we evaluated the predictor on all corresponding examples (positive and negative), and calculated the resulting error rate

Evaluation Metric 2: Predictor Choice Analysis
This evaluation metric tested the ability of the predictor to choose between two responses given a prompt. For the training, dev, and test sets, we evaluated the predictor by giving it a prompt from a discourse, and a choice between the actual response and a randomly chosen response, and having it choose the one that scored higher. A tie was split 50-50, and the result was the percentage of correct answers.

Evaluation Metric 3: Predictor Guess Analysis
This evaluation metric was similar to Predictor Choice Analysis, except that given a prompt, it had to choose from all possible responses in the discourse.

Evaluation Metric 4: Chat Bot Analysis
Evaluation was also conducted subjectively through interaction with the chat bot. These results are more informal in nature. For further results, a user testing experiment could be conducted.

Testing and Revision

Corpora Cleaning

One challenge presented by the corpus was the existence of noise. **Noise** consisted of utterances with the act_tag 'b' (backchannel), '%' (uninterpretable) and 'x' (sound). Some turns consisted purely of noisy utterances, and we classified these a **noisy turns** (referenced in code as a **bad turn**). On average, 26% of turns in a discourse were noisy turns. This noise is an important part of conversation, especially backchannels, but they posed a problem in terms of defining positive and negative examples. Negative examples were generated randomly with the presumption that most random combinations would be unlikely. This worked well generally, with the exception of noisy turns, since noise could be paired with almost any kind of turn.

Table: Noise and Noisy Turns

'b'	Backchannel 15.85%	"Uh-huh"
'%'	Uninterprettable 7.11%	???
'x'	Noise 1.62%	<Laughter>

Number of Noisy Turns: 25,663 (out of 97,331 total turns)
Noisy Turn Density per Discourse : 26.37%

Preprocessing (not included in final runs)
One idea was to remove all noise during the processing of the transcripts. Testing this while still allowing backchannels was also tested.

Train error = 15.16%, dev error = 37.84%
Train Choosing Score: 80.79%
Test Choosing Score: 67.76%

(Just preprocessing out '%' and 'x')
Train error = 14.64%, dev error = 30.77%
Train Choosing Score: 80.81%
Test Choosing Score: 71.31%

Preprocessing out noise generally gave worse results on all evaluation metrics. This is due partly to the fact that our score improved when we did correctly identify noise, but also because by removing noise, we changed the natural state of the conversation. When comparing which examples the predictor succeeded/failed on with and without preprocessing, we found that preprocessing allowed us to better identify when collaborative completion was used as a response (59.17% vs. 55.93%), but worsened our ability to identify proper responses to a hedge (61.94% vs. 77.23%). Percentages for other categories were comparable. We decided not to preprocess out noise.

Negative Filtering (included in final runs)
Another idea was to remove all negative examples with noisy turns. This allowed us to still capture aspects of the use of backchannels and other noise, but by removing it from the negative examples, we did not penalize the use of noise in cases where it would be fine. This did bias the weights in favor of noise, which was fine for the evaluation metrics, but had to be taken into account in the chat bot (see Chat Bot Revisions)

Without Negative Filtering
Train error = 16.75%, dev error = 38.31%
Train Choosing Score: 79.49%
Test Choosing Score: 66.90%

Negative filtering maintained the integrity of the original discourses and provided for better evaluation results and chat bot use. We include negative filtering for all following discussion.

Algorithm Revisions

Speed/Efficiency
For efficiency, we cached the feature vectors for each example for use in between iterations of the stochastic gradient descent. We also allowed saving of the generated weight vector if the train set and features did not change between uses of the chat bot. With this, the initial processing of the transcripts and the generation of the feature vectors took ~4 min., and from a user perspective, the response time of the bot is immediate, so time was not generally an issue.

Hyper-Parameter tuning
We tested different step sizes and numbers of iterations, as well as introducing a dampening factor to the step size in between iterations and a test for weight convergence.
STEP_SIZE = .5 and NUM_ITERATIONS = 40 provided the best and most balanced evaluation metric results between the train and dev sets.

Convergence was considered if the change in all feature weights was below a convergence threshold, and dampening was done by multiplying the step size by a dampening factor between iterations. With a conversion threshold of .01 and a dampening factor of .95, the weight vector did converge after 60 iterations with very similar results (train error up .5%, test error down .5%, Train choosing up 4%, test choosing up 4%). We elected to not use convergence and dampening.

Feature Testing

Our most general features, that fired for all or most turns, were the most effective in reducing our train and dev error rates. Unsurprisingly, the features utilizing dialogue act tags were the most effective, since they contain a lot of contextual information about an utterance. We thought that the order of tags in a turn would be useful, so considered two features that stored the tags of each turns as an ordered list. But, that increased both error rates by ~0.1-0.2% so didn’t use that. Similarly, capturing the consecutive pairs of tags in both turns increased our error as well. In retrospect, however, that makes sense since our chatbot deals with turns, not utterances or words, so the ordering within an individual turn is less important. That is likely why storing each turn’s tags as a set, which does not preserve the order, helped. As a whole, the dialogue act tag features reduced the train error by ~17.5% and the dev error by 9.9%.

Another feature relating to words within individual utterances that we tried was part-of-speech tags, which we attempted to capture as a set for each utterance as well as considering consecutive pairs. However, these features actively hurt our results for likely the same reasons mentioned above - that they contained information about particular words, not phrases, which was less helpful for the scope of our problem.

Utterance length was also effective although not as much as the dialogue act tags. Altogether, those features reduced the train error by ~2% and actually increased the dev error by ~0.6%. However, we decided that the decrease in train error was significant enough to warrant the slight dev error increase.

The length of a turn - both in terms of utterances as well as words - was effective but not as much as we thought it would be, considering the importance of a turn in how we have implemented our chatbot. These features reduced our train error by ~0.7% and increased our dev error by ~0.7%. But, we decided to keep them since it was making a difference, although small, for the train error and we thought that it was an important conceptual piece of information to capture.

Although we use an imperfect heuristic for interruptions, the inclusion of the interruption features reduced our train error by ~1.4% and did not affect the dev error, so it was certainly useful.

Including the subjects of the each turn did not affect our train error but increased our dev error by ~0.5%. This makes sense since it is possible to have conversations where both people are talking about the same thing as well as where both people are talking about different subjects although their conversation still stays on-topic. However, we thought that the increase in dev error was not significant, so kept these features nonetheless.

In contrast, considering pronouns in a turn - specifically keeping track of the presence of “I”s and “you”s - increased our train error by ~3%. Since all this feature did was mark the presence of these two pronouns without any contextual information about where in the turn nor where exactly in the utterance they occurred, it is not surprising that it did not help our results.

Although our pairwise features covered many different cases, they actually did not significantly affect our train and dev error rates. In retrospect, this makes sense because their specificity means that they did not fire very many times.

Chat Bot Revision

Noise

Because noise scored so highly, but only consisted of 26% of discourse, we decided to limit the uses of noise by the chat bot. Based on this statistic, we added noisy turn counter that did not allow a noisy turn to be used by the bot for its next three turns (allowing it noise every fourth turn).

Interruption

We also wanted to incorporate the possibility for interruption with backchannel or collaborative completion with respect to the given feature and interruption statistics. As the user enters utterances, the chat bot will treat the set of user utterances so far as a prompt, and see if any candidates starting with backchannel or collaborative completion score high enough to garner response. Since turns with backchannels scored very high, we also limited interruptions to only occurring after a given number of utterances were entered by the user, and added random variable to allow or not allow interruption. In practice, the bot tended not to interrupt, but interruptions that did occur were not intrusive fit well enough into conversation.

Greedy versus Probabilistic choice

The initial iteration of the chat bot had it greedily choose the candidate response with the greatest score. To allow for greater variety in discourse, we also implemented a probabilistic choice from the set of candidates. For all candidates with positive scores, we normalized their scores to use as probabilities, and chose based on that distribution.

To decide which form of computer choice to use, we ran the chatbot with both on the same discourses and tried to keep the conversations as similar as possible to holistically evaluate its performance. For the most part, the performance was the same. There were many instances where given a turn, the chatbot chose the same turn both times. But, there were a significant number of cases where the chatbot chose a better response with the probabilistic distribution. Consider the following example:

—— (discourse context)
Chatbot (qo): So how serious is the
You (x) : <Laughter>
Chatbot (+): subject of crime in your area?
You (sv): Well, needless to say, here in Washington DC, this is the war zone.
You (sd): Uh, DC around here stands for drug capital or death capital.
—— (chatting begins)
You (sv): Seriously, it's horrendous
You (sd): Every hour, someone dies from a drug related crime or gang violence.

With the probabilistic distribution:

Chatbot (b): Yeah
Chatbot (sv): there aren't that many

With the greedy algorithm:

Chatbot (b): Yeah

In this situation, the probabilistic discourse chose a more relevant response, as opposed to the more simplistic one. Because of these observations, as well as the fact that we wanted our chatbot to have more of a variety of responses to more accurately mimic human conversation, we chose to use the probabilistic distribution for computer response choice.

Evaluation Metric Revisions

Evaluation metrics were modified to account for negative filtering.
Evaluation Metric 1: Predictor Analysis
Other than the fact that the negative examples were different, no change was made.

Evaluation Metric 2: Predictor Choice Analysis
The random response was not allowed to be a noisy turn.

Evaluation Metric 3: Predictor Guess Analysis
No change.

Evaluation Metric 4: Chat Bot Analysis
See changes to chat bot above.

Final Results

The following results were with the same training set, dev set, and an unused test set. We include a baseline (minimal features, essentially random), and an oracle (human attempt at similar task) for comparison.

Table: Results

Eval 1: Predictor Error

	Train	Dev	Test
Baseline	50.00%	--	--
Bot	11.95%	22.17%	21.01%
Oracle	33.00%	--	--

Eval 2: Choose Correctness

	Train	Dev	Test
Baseline	50.00%	--	--
Bot	82.59%	75.82%	76.59%
Oracle	87.00%	--	--

Eval 3: Guess Correctness

	Train	Dev	Test
Baseline	1.01%	--	--
Bot	6.57%	--	4.82%
Oracle	--	--	--

Weights Analysis

Noise-related Weights

As previously mentioned, noise consists of acknowledgements, abandoned, and nonverbal statements.

Highest weights:

3.5 - the DAT sets: {non-opinion statement}, {acknowledgement}
3.2 - the DAT sets: {abandoned}, {multi-utterance}
3.15 - the DAT sets: {opinion statement}, {acknowledgement, abandoned}
3.1 - the DAT sets: {opinion statement}, {acknowledgement}
2.9 - the DAT sets: {non-opinion statement}, {acknowledgement, abandoned}

#2 aside, these all make sense considering how telephone conversations tend to work. There are lots of acknowledgements, such as “uh-huh”, “yeah”, and “right”. More specifically, it is most likely for acknowledgements to come after non-opinion and opinion statements.

Lowest weights:
-3.7 - the DAT sets: {multi-utterance, non-opinion statement, abandoned, opinion statement, non-opinion statement with quoted material}, {agreement}
-2.65 - the DAT sets: {acknowledge, abandoned, opinion statement}, {acknowledge, non-opinion statement}
-2.5 - the DAT sets: {agreement, opinion statement}, {acknowledgement, abandoned, statement}
-2.5 - the DAT sets: {open question}, {abandoned, opinion statement}
-2.45 - the DAT sets: {agreement, abandoned, non-opinion statement}, {action-directive}

#1 through #3 are relatively specific, particularly #1, so it is not surprising that these are the most hurtful negative weights with respect to noise. However, #4 makes sense since it is more likely that people would answer an open question straight away with their own opinion as opposed to doing so after cutting themselves off. Similarly, an action-directive phrase is unlikely to follow a non-opinion statement.

Utterance and Turn Length Related Weights

Highest weights:
1) 3.15 - utterance length: 2nd last of A = 7, last of A = 5, first of B = 6, 2nd of B = 10
2) 2.9 - turn length in words: 5, 95
3) 2.85 - utterance length: 2nd last of A = 16, last of A = 7, first of B = 12, 2nd of B = 6
4) 2.75 - turn length in words: 8, 114
5) 2.7 - utterance length: 2nd last of A = 2, last of A = 9, first of B = 15, 2nd of B = 11

As they contain more information, it makes sense that some of the features with the highest weights contain the four pieces of utterance length information as opposed to just the two. That being said, #2 and #4 are also interesting to note since they are two statistics without much other context, but turn out to have quite high weights.

Lowest weights:
1) -3.0 - utterance length: 2nd last of A = 2, last of A = 4, first of B = 2, 2nd of B = 4
2) -2.7 - utterance length: last of A = 51, first of B = 3
3) -2.65 - utterance length: 2nd last of A = 8, last of A = 13, first of B = 2, 2nd of B = 2
4) -2.6 - turn length in words: 4, 93

The fact that the four piece utterance length features contain more information means that there exist more four piece features. Thus some will be more specific and less generally applicable than others so it makes sense that the lowest weight features consist of mostly four piece utterance length features. #4 was also less surprising since a very short turn (but not a one word acknowledgement) is unlikely to follow or precede a long turn. Let us define an interjection to be a turn that comes in between two long turns and that is not a one word acknowledgement. Considering that 4, 93 had a low weight but 5, 95 and 8, 114 had high weights, if we assume that perhaps the 4/5/8 word responses came after another long turn, it is possible that there is a threshold for the minimum length of an interjection.

Dialogue Act Tag Related Weights

Highest weights:
1) 2.75 - DAT sets: {multi-utterance}, {non-opinion statement with quotation, opinion statement, non-opinion statement}
2) 2.7 - DAT sets: {multi-utterance, opinion-statement}, {repeated acknowledgement}

3) 2.6 - DAT sets: {non-opinion statement, other answer}, {response acknowledgement}
4) 2.5 - DAT sets: {yes-answer, elaborate response to yes/no question}
5) 2.45 - DAT sets: {multi-utterance, non-opinion statement}, {agreement, opinion statement, repeated agreement}

Overall, it makes sense that the sets of DAT tags are the highest weight DAT-related feature class. Each individual DAT tag contains a lot of information, so all of them together reveals even more about a particular turn, and both sets at once provide a lot of context about the ways in which people have conversations. In general, phone conversations tend to include lots of acknowledgements (e.g "uh-huh") and agreements (e.g "so true"), and it makes sense that these follow combinations of opinion and non-opinion statements in #2, #3, and #5. #4 makes sense as well since often people answer a yes/no question with an immediate yes/no answer and then a follow-up explanation. It is hard to speculate about #1 since the multi-utterance DAT is not well documented, but people often quote others in phone conversations and then give their opinions on them.

Lowest weights:
1) -2.4 - DAT sets: {rhetorical question, opinion statement}, {agreement, repeated agreement}
2) -2.35 - DAT sets: {open question, other, non-opinion statement}, {non-opinion statement}
3) -2.2 - DAT sets: {collaborative completion}, {agreement, opinion statement}
4) -2.15 - DAT: last of A: yes/no question, first of B: agreement
5) -2.15 - DAT sets: {appreciation}, {multi-utterance, opinion statement}

Again, although the sets of DAT tags are the highest weight DAT-related feature class, they are also the lowest weight class for the same reason - more information means that there are more possibilities for DAT tag sets, which means that some sets will be very specific and non-generalizable. #1 makes sense since people rarely say so many consecutive agreement phrases. #2 also makes sense since it is unlikely that open questions would be followed by non-opinion statements. #3 and #5 are more specific, but #4 is quite logical since yes/no questions tend to be followed by yes/no answers as opposed to blind agreement.

Interruption-Related Weights

Highest weights:
1) 1.2 - Pre acknowledgment interruption: turn length (utterances) = 10
2) 1.2 - Pre acknowledgment DAT set: {declarative yes/no question, non-opinion statement}
3) 1.1 - Pre acknowledgment DAT set: {hold, no-answers}
4) 1.05 - Pre acknowledgment DAT set: {yes-answer, elaborate response to yes/no question, repeated yes-answer, non-opinion statement}

We can see that the feature class with both the highest and lowest weights is the pre acknowledgement DAT set. We have already discussed why the plethora of information in a DAT set makes it likely to result in very high and very low weights. As for pre acknowledgment vs pre collaborative completion, our two types of interruptions, these high impact features reflect our corpus; acknowledgments are 15.85% of it while collaborative completions are only 0.31%.

That being said, it is surprising to note that one of the highest weight features, #1, is actually a turn length in utterances. One can imagine, though, that a speaker is more likely to be interrupted after a very long turn. Looking at #2 through #4, the combination of DATs in the sets are logical.

Lowest weights:
1) -1.15 - Pre acknowledgment DAT set: {hold, non-opinion statement}
2) -1.1 - Pre acknowledgement DAT set: {acknowledgement in question form, non-opinion statement, opinion statement}
3) -1.1 - Pre acknowledgement DAT set: {conventional closing}
4) -1.05 - Pre acknowledgement DAT set: {multi-utterance}
5) -1.0 - Pre acknowledgment DAT set: {open question, other answer}

Looking at these weights, it seems that there are a few specific combinations of utterances that are less likely to be interrupted. #3 seems logical, but the others seem rather random.

Subject-related Weights

Highest weights:
1) 0.5 - B subject = I
2) 0.4 - A subject = it
3) 0.3 - A subject = you
4) 0.3 - B subject = they
5) 0.3 - B subject = you

Lowest weights:
1) -0.95 - A subject = I
2) -0.1 - A subject = that
3) 0.0 - B subject = somebody
4) 0.0 - A subject = that
5) 0.0 - B subject = Texas

It is certainly surprising that turn A having a subject of "I" has a negative weight; both the highest and lowest weights seem rather arbitrary, which perhaps explains why the subjects barely affected our train and dev error rates.

Pairwise Feature Weights

Although the pairwise features did not affect our train nor dev error rates, we wanted to examine their weights to consider how well the features were designed.

Positive weights:
1) 2.05 - Summarization followed by acknowledge
2) 1.55 - Apology followed by downplayer
3) 1.25 - Yes/no question followed by an 'other-answer'
4) 1.15 - Collaborative completion
5) 0.4 - Yes/no question followed by a yes/no response
6) 0.35 - Open question followed by an opinion
7) 0.35 - Wh-question followed by a non-opinion statement

Negative weights:
1) -0.45 - Collaborative completion followed by a "maybe"
2) -0.3 - Non yes/no question followed by a statement of any kind

It looks like most of them had positive weights, which is promising, except for two. The second negative weight is surprising since it seems like a very general feature. However, perhaps it was too general.

Overall Weights

Highest weights:

1) 3.5 - the DAT sets: {non-opinion statement}, {acknowledgement}
2) 3.2 - the DAT sets: {abandoned}, {multi-utterance}
3) 3.15 - the DAT sets: {opinion statement}, {acknowledgement, abandoned}
4) 3.15 - utterance length: 2nd last of A = 7, last of A = 5, first of B = 6, 2nd of B = 10
5) 3.1 - the DAT sets: {opinion statement}, {acknowledgement}

Considering the number of statements in our corpus - 32.1% of utterances are non-opinion statements and 11.37% are opinion statements - it is not surprising that the best features include these two statement types.

Lowest weights:

1) -3.7 - the DAT sets: {multi-utterance, non-opinion statement, abandoned, opinion statement, non-opinion statement with quoted material}, {agreement}
2) -3.0 - utterance length: 2nd last of A = 2, last of A = 4, first of B = 2, 2nd of B = 4
3) -2.85 - turn length in words: A = 50, B = 32
4) -2.7 - utterance length in words: A = 51, B = 3

As for #2 and #3, although utterance and turn lengths were helpful for our train and test errors, they contain less contextual information than some of our other features, such as those involving the dialogue act tags or parse tree information, so it makes sense that these characteristics inform our worst features.

Linguistic Discussion

For linguistic discussion, we investigate both the results of the predictor on the positive and negative examples, as well as a series of conversations held with the chat bot.

Discussion of Examples

In addition to the results given in evaluation metrics 1-3, we give a breakdown of predictor success on subsets of the examples. With each subset, we give a quantitative analysis of how the predictor preformed, a short discussion, and a few examples.

Questions in Prompt
Correct positive examples: 74.73%
Average score: .72
Correct negative examples: 79.63%
Average score: -1.23

The predictor did generally well with respect to questions. It approved of yes-no answers to yes-no questions, statements of fact to open questions, etc. It disapproved of multi-utterances and collaborative completions following questions, since those are for carrying on previous discussion, and questions redirect conversation. Features did not do a great job of dealing with cases in which a statement and a multi-utterance both existed, and of course did not deal with semantics.

True Positive
Prompt: Yeah did you go to Baylor?
Response No
I went to UT at Austin
Uh-huh

True Negative
Prompt: Do have air conditioning in your car
Response: --and not, not done anything

False Negative
Prompt: That could be very helpful
Response: A lot of fun at the moment
Not really a good example of a question

False Positive
Prompt: Well what else do you like?
Response: --and I notice they have like the, you know, The Louisiana Purchase
Razzoo's has opened up down, you know, off of, um, oh close to six thirty-five/
Response has statement and multi-utterance in response to open question. Length also led this to be positive

Noise in Response
Correct positive examples: 82.82%
Average score: 2.49
Correct negative examples: 78.76%
Average score: -1.07

Noise is almost always okay. Really came down to interaction with length features. Certain lengths of utterances and turns were typical to backchannel and some were not, but there was not a very clear pattern.

(See appendix for examples used in chat bot)

Collab in Response
Correct positive examples: 36.17%
Average score: -0.24
Correct negative examples: 75.68%
Average score: -0.63

Collaborative completion was tricky to capture. It was generally found to come after statements, and turns of length 3 utterances, but this was also many cases. It did notice when not to use collaborative completion, such as after a command or closing. It did make mistakes however, as sometimes this does occur.

FOUND: True Positive
Prompt: --and I'll pay you or whatever
but, uh, there's been some magazines I've looked at
and it's a real turn off to see every
Response: Every page

FOUND: True Negative
Prompt: Well, have a good day
Response: They may- <faint>

FOUND: False Negative
Prompt: You're just going to have to imagine what you would like as benefits <laughter>.

Response: <Laughter> Or what I'm looking for.
This was a collab in response to a command. Examining conjunctions, and the use of phrases vs. full sentences would likely help.

Agreement in Response
Correct identification of positive examples: 53.92% Average score: .02
Correct identification of negative examples: 72.28% Average score: -0.79

The predictor caught the pattern of opinion, acknowledgement, opinion, but the weights were not strong enough.

True Positive
Prompt: I do not want you here.
But there is just such a surge of southeast Asians.
It has affected the culture of this country.
Response: Oh, yeah.
It's ab --

True Negative
Prompt: I think they can do the plastics like the liter bottles –
Response: That's true

Hedge in Prompt
Correct identification of positive examples: 72.83% Average score: 1.23
Correct identification of negative examples: 81.62% Average score: -1.16

Hedges were correctly identified as being followed by agreement or acknowledgement.

FOUND: True Positive
Prompt: Davis is kind of strange
Apparently he's kind of a wild guy
I don't really know for sure
They've had some excellent teams though obviously
Response: Yes
They have.
Suprisingly, this actually scored negatively for features including hedge, but the other structures surrounding it were very positive.

Conventional Opening in Prompt
Correct positive examples: 66.67%
Average score: 1.45
Correct negative examples: 90.00%
Average score: -1.76

The predictor (and bot) did well with conventional openings and closing. It was interesting to note that some features matching and opening to an opening were negative, but most were strongly positive. Examples with conventional openings sometimes were misidentified because the opening was one of many utterances. This may be an aspect of the corpus itself, since the topic was provided and many people just jumped right into talking without initial formalities.

FOUND: True Positive
Prompt: My name is Jean.
Response: I'm Tina

FOUND: True Negative
Prompt: I'm Tina
Response: most of the time for that

FOUND: False Positive
Prompt: Hi
I like, I love to do all kinds of craft and sewing
I find that's my one release <laughter>
Response: Yeah
Well, I'm not much of a basketball fan either
off topic (no semantics) but otherwise sounds valid

Conventional Closing in Prompt
Correct positive examples: 94.01%
Average score: 1.15
Correct negative examples: 92.72%
Average score: -2.16

Similar rules for closing. It is unusual to use a multiutterance or almost anything other than a closing after a prompt to close.

True Positive
Prompt: Bye-Bye
Response: Bye-Bye

True Negative
Prompt: Thanks for calling.
Response: [Really long utterance with +]

False Positive
Prompt: Good talking to you.
Response: All right.
Thanks a lot.
Thrown off by length features and act tags. There may be an issue with number of closings used.

Conventional Closing in Response
Correct positive examples: 86.68%
Average score: 0.98
Correct negative examples: 91.04%
Average score: -1.90

Again, similar. It was interesting to note that it was unusual to close after collaborative completion and statement of opinion.

False Negative
Prompt: Yeah, I think so.
Response: Well I think so.
I enjoyed talkin to you.
Not typical to close after statement of opinion. Seems very abrupt.

True Negative
Prompt: I haven't seen it in years.
But I used to
Response: Bye-bye

Utterance and Turn Length

These features helped overall, but it was difficult to discern a pattern that translated to obvious linguistic generalizations.

Subjects

This was a small attempt at semantic analysis. The only very strong subject feature was "I", and it scored highly as a prompt, and low as a response! Seems it may ot good to switch conversation to just be about you.

Chatting with the Bot

The chatting experience itself was interesting since the user would have to try and mimic phone conversation via a text chat, and the two forms of communication are quite different. The most significant difference is that turns in a phone conversation are longer and more rambling, while text chats tend to be terser; that was something we had to keep in mind as we chatted with our bot.

Limiting the conversation to the chosen Switchboard discourse was surprisingly effective. When beginning a chat, the bot printed out the first few turns of the discourse, so that the user could envision themselves as in the middle of a conversation before actually starting. One issue that came up again and again was the usage of acknowledgments, such as "uh-huh" or "right". Because of the frequency of these words, the bot used them a lot in responses. In many cases, they were well-timed and made sense in the conversation, but that made it more difficult to nudge the bot to contribute some original ideas. In other cases, they were simply not relevant. That being said, given the constraints of its possible response set, the bot did a decent job at sustaining semi-coherent conversations. It was not uncommon to have it respond in a proper or at least seemingly relevant way to completely unseen prompts.

(See Appendix for example conversations)

Conclusion and Further Work

The approach of modeling conversation as a set of turns, prompts and responses, and using stochastic gradient descent with various linguistic features proved a successful tool for analyzing and predicting aspects of conversational interaction. The extension of the use of dialogue acts to include turn and utterance lengths, noise analysis, multiple DAT interactions, and interruption features was effective in achieving a low predictor error rate (compared to baseline) and a respectably high choose and guess rate (considering the lack of semantic understanding). The bot outperformed the human oracle, although the human was at a disadvantage in not having access to the entire corpus.

The current set of features performs well, but it makes little use of key words and parts of speech. Analysis of use of full sentences versus phrases, age, dialect, and gender may also be useful.

While this algorithm provides a reasonable predictor and a chat bot that appears to converse with some understanding of conversational flow, this project is really meant to supplement related work in chat bots and conversation. Integration with the dialogue act tagging algorithm, a parser, and a semantic analysis and logic parser could provide a chat bot that is not just intelligent in an assistive role, but a conversive role as well.

Errata

Division of work

Dustin primarily implemented the predictor and the actual chat bot itself as well as the linguistic analysis. Rupa primarily did the initial corpus research and related scripting, the feature design and analysis, and the weights analysis. Both also spent a significant amount of time chatting with the bot and collecting interesting conversations.

Relation to 221

Dustin is also enrolled in CS 221, and is using the same project for that class. The class focuses primarily on the model and method, as opposed to linguistic feature design and analysis.

For CS 221, he set up a special baseline and oracle test, tried different SGD hyper-parameters and loss functions. He also spent time trying other models altogether, including a reinforcement learning model and hidden markov model.

For CS 224N, he also worked on creating various features with DATs, set up the chat bot and extended it to deal with noise and interruption, and did linguistic analysis on the training examples.

Citations

Stolcke et al. *Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech*. "http://web.stanford.edu/~jurafsky/ws97/CL-dialog.pdf"

Christopher Potts, Stanford Linguistics. *The Switchboard Dialogue Act Corpus - Data and Code*. "http://compprag.christopherpotts.net/swda.html"

Open Subtitles Corpus: http://opus.lingfil.uu.se/OpenSubtitles.php

Stanford CoreNLP Tools: http://nlp.stanford.edu/software/corenlp.shtml

APPENDIX

Conversation 1

---- (discourse context)
YOU (qw): {D So, } who's your favorite team? /
ME (sd): My favorite team is the Pittsburgh Steelers. /
YOU (br^m): Pittsburgh? /
ME (na): # You bet. # /
YOU (sd): # I used # to be a big Pittsburgh fan when I was little. /
ME (%): {D Well, } I, {F uh, } - /
ME (sd): when ((John)) ((Stallworth)) played he was from my hometown in Alabama, /
----- **(chatting begins)**
YOU: That must have been fun.
Act Tag: sd
YOU: Did you ever get to meet him?
Act Tag: qy
YOU:
ME (na): {D Well, } because I'm right here in Dallas [I, + I] kind of pull for Dallas. /
SCORE
1.75
YOU: Excuse me?
Act Tag: br
YOU:
ME (b): Right. /
SCORE
1.7
YOU: I've never been to Dallas.
Act Tag: sd
YOU: But my mom grew up there, actually.
Act Tag: sd
YOU: She didn't like it very much.
Act Tag: sv
YOU: Too many cowboys, am I right?
Act Tag: ^g
YOU:
ME (+): # I'll almost always # root against them. /
SCORE
1.75
YOU: Dallas has great teams, though
Act Tag: sv
YOU: I don't remember the name
Act Tag: h
YOU: Cowboys maybe?
Act Tag: ^g
YOU:
ME (%): {C So } that was, - /
ME (sd): [I, + I] watched that game /
ME (%): {C and, } {F uh, } that was, {F uh, } -/
SCORE
1.0
YOU: The Steelers versus the Cowboys?
Act Tag: ^g
YOU: Yeah, that was awesome.
Act Tag: sv
YOU:
ME (ba): {F Boy, } /
ME (sv): [they are, + they are] just a fun team to watch. /
SCORE
1.8

Conversation 2

---- (discourse context)
YOU (qy): {D So. } Have you heard about Saturns? /
ME (^h): <Swallowing> <hum> <breathing> <<he seems to be out of breath>> I've heard some about Saturns. /
ME (sd): I don't know <inhaling> a lot about them. /
ME (sd): I haven't been over, {F uh, } to a dealer to look at them although I did see something in tonight's paper that said that their, {F uh } reliability is rated to be equal to that of [the, + I guess, the] Japanese cars. /
YOU (ba): {F Ooh, } that's great. /
YOU (sd): {D Actually } that's kind of the type of car that [I, + I'm] thinking about we might get. /
YOU (sd): Saturns, you can't go to a dealer unless you go to [a, + a] Saturn dealer.

YOU (+): [This, + this] offshoot of G M <thumping>. /
ME (b): <Laughter> Yeah. /
YOU (sv): {C And, } {F uh, } supposedly, they've got great customer <sniffing> satisfaction from what things I've heard. /
ME (aa): {F Um. } Yeah, /
ME (sd): I've heard a lot of people like them. /
ME (sd): {F Uh, } yet I don't know [wh-, + whether] I would buy Saturn or not at this point. /
ME (%): I'm kind of, - /
ME (sv): my card and, {F uh, } desires tend to be pretty picky /
ME (%): {C and so, } {F uh. } -/
---- (chatting begins)
YOU: Yeah, they are rather expensive.
Act Tag: sv
YOU: But I'm OK with paying a little extra for that quality and customer service.
Act Tag: sv
YOU: Assuming the car doesn't break on me though.
Act Tag:
YOU:
ME (b): Uh-huh. /
ME (b): Sure. /
ME (%): # Those are # -- -/
SCORE
0.875
YOU: My brother had a Jaguar, and he spent thousands in repair bills.
Act Tag: sd
YOU: Seriously, I asked him why he didn't just get rid of it.
Act Tag: sd
YOU: But he just said it was his "baby"
Act Tag: sd
YOU: Can you believe that?
Act Tag: qh
YOU:
ME (aa): {F Um. } Yeah, /
ME (sd): I've heard a lot of people like them. /
ME (sd): {F Uh, } yet I don't know [wh-, + whether] I would buy Saturn or not at this point. /
ME (%): I'm kind of, - /
ME (sv): my card and, {F uh, } desires tend to be pretty picky /
ME (%): {C and so, } {F uh. } -/
SCORE
0.692307692308
YOU: What are you looking for?
Act Tag: qo
YOU:
ME (sv): {C And } I find that for [a normal, + {D you know, } everyday things.] It's really very easy to work on /
ME (sd): {C and so, } {F uh, } {D you know, } I'm kind of spoiled. /
ME (sd): I used to have a (( )) Chevy. /
SCORE
0.0663507109005
YOU: Ah, so robust and a smooth drive.
Act Tag: bf
YOU: My dad used to have a Mercedes when we lived in Germany.
Act Tag: sd
YOU: And he said it was the smoothest ride of his life.
Act Tag: sd
YOU:
ME (b): Yeah. /
SCORE
0.0672963400236
YOU: I think smoothness is overrated though.
Act Tag: sv
YOU: I just don't drive enough to care that much.
Act Tag: sd
YOU: And what can I say
Act Tag: h
YOU: I just hate driving.
Act Tag: sv
YOU:
ME (sd): # With a # three fifty in it. /
SCORE
0.3125

Conversation 3

---- (discourse context)
YOU (o): Okay. /
ME (qo): {D So } how serious is the,

YOU (x): <Laughter>.
ME (+): subject of crime in your area? /
YOU (sv): {D Well, } needless to say, here in Washington, D C <laughter> this is the war zone. /
YOU (sd): {F Uh, } D C around here stands for drug capital or death capital. /
ME (x): <Laughter>.
---- (chatting begins)
YOU: Seriously, it's horrendous.
Act Tag: sv
YOU: Every hour, someone dies from a drug related crime or gang violence.
Act Tag: sd
YOU:
ME (b): Yeah, /
ME (sv): there aren't that many,
SCORE
0.0707317073171
YOU: Wow, I can't imagine what that must be like.
Act Tag: ba
YOU: I had it lucky, but some of this kids just don't know anything else.
Act Tag: sd
YOU: Can you imagine growing up like that?
Act Tag: qh
YOU:
ME (b): Yeah. /
SCORE
0.0971867007673
YOU: What's the crime in your area like?
Act Tag: qo
YOU:
ME (%): {E I mean, } [there's, + there's] been a lot of, {F uh, } - /
ME (sd): {C and } I used to think that this whole argument was completely bogus, /
ME (sv): {C but } then, the --
SCORE
0.4
YOU: If there even is any.
Act Tag: sd
YOU:
ME (x): <Laughter>.
SCORE
0.0609065155807
YOU: I don't know if I'm making any sense or not.
Act Tag: h
YOU:
ME (sd): It's illegal, /
ME (sv): {C but } it's not wrong because all their friends do it. /
SCORE
0.131578947368
YOU: Oh, okay.
Act Tag: bk
YOU: Kids these days.
Act Tag: t1
YOU: The government has to step in.
Act Tag: sv
YOU: There's no other alternative.
Act Tag: sv
YOU:
ME (x): <Laughter>.

Conversation 4

---- (discourse context)
YOU:
I've, {F uh, } <clicking> <<telephone>> - /
YOU:
are you there? /
ME:
Yes, sir, /
ME:
I'm right here. /
YOU:
All right <laughter>. /
YOU:
Okay, {F um, } /
YOU:
I'm in an area where, {F um, } we can, {F uh, } enjoy it year round, /
YOU:
{C so, } I've actually [been, + {F uh, } {F uh, } been] out within the past couple of weeks. /
YOU:
I'm mainly a freshwater fisherman. /
YOU:
How about yourself? /
ME:

{D Well, } {D actually } I've done both kinds -- /
YOU:
Uh-huh. /
ME:
-- {F uh, } I'm originally from the State of Virginia --
---- (chatting begins)
YOU: Me, too!
Act Tag: sd
YOU: Is that right?
Act Tag: bh
YOU: What part?
Act Tag: qw
YOU:
ME: {D Well, } [[I, + I,] + the disease I] caught was simply because, {F uh, } [I, + I] was in the Norfolk Portsmouth area, /
ME: {C and } --
SCORE
0.4
YOU: Oh no
Act Tag: %
YOU: I am so sorry.
Act Tag: fa
YOU:
ME: I, - /
ME: wherever the area, - /
ME: whatever [the +] fish is the specialist in that area, [I, + {F uh, } I] enjoy fishing for it, /
ME: {C and } I try to, - /
ME: I haven't fished in several years --
SCORE
1.4
YOU: I can imagine, what with the disease and all.
Act Tag: ba
YOU: Sounds awful.
Act Tag: sv
YOU:
ME: {D You know. } -/
SCORE
1.2
YOU: Well, no
Act Tag: ar
YOU: Not really.
Act Tag: ar
YOU: I am not diseased.
Act Tag: sd
YOU:
ME: {F Uh } -- /
SCORE
1.4
YOU: Sorry, that was rude of me.
Act Tag: fa
YOU:
ME: I've, {F uh, } <clicking> <<telephone>> - /
ME: are you there? /
SCORE
0.35
YOU: Yeah, I'm hre
Act Tag: ny
YOU:
ME: Yeah. /
SCORE
1.5

Conversation 5

---- (discourse context)
YOU:
Do you know anyone that, {F uh, } [is, + is] in a nursing home or has ever been in one? /
ME:
No. /
ME:
{C But } [I, + my grandparents] were looking into it before /
ME:
{C so } I know what they've said. /
YOU:
Uh-huh. /
ME:
Uh-huh. /
YOU:
{D Well, } I'm trying to think. /
YOU:
My, {F uh, } {F uh, } wife's grandmother had Alzheimer's /
YOU:
{C and } they were going to put her into [a, + a] nursing home /
YOU:

{C and, } {F uh, } [they, + when they] put her in, she had all kinds of trouble / 1.4
YOU:
{C and } the nursing home made them come <noise> and take her back because she <laughter> was being [[[a, + a,] + {D you know, } a,] + a] nuisance. # Or worse than a nuisance, /
YOU:
{E I mean } # --
ME:
{F ((Oh)) } they thought it was too much of a bother. # /
---- (chatting begins)
YOU: Exactly
Act Tag: aa
YOU: I thought it was very distasteful of them
Act Tag: sv
YOU: I mean, it was terrible.
Act Tag: sv
YOU: Right?
Act Tag: ^g
YOU:
ME: {C So } it's another problem. /
SCORE
0.85
YOU: Yep
Act Tag: ny
YOU:
ME: Uh-huh. /
SCORE
1.5
YOU: I am not sure what I am supposed to do.
Act Tag: h
YOU:
ME: # That, {F uh, } - /
ME: # you can't always get in when you want too /
ME: {C and } of course, you can't just sit around and wait. /
SCORE
1.6
YOU: Well what can I do at this point?
Act Tag: qw
YOU:
ME: Right. /
SCORE