Using Past Speaker Behavior to Better Predict Turn Transitions

Thesis Presentation

Tomer Meshorer

Center for Spoken Language Understanding
Oregon Health & Science University
Portland, Oregon,USA

08 June 2017

Outline

‘

Section 1

Motivation

Current Issues

1. For a natural conversation between human and machine, we want to conform to human to human turn taking system (Sacks et al, 1974)
   - One speaker at the time
   - Overlap speech is common but brief (only 17
   - Gaps between turns kept to a minimum (200 ms)
   - No fixed turn size and no fixed conversation size.
2. In Human-Human conversations conversants predict (Sacks et al, 1974) or signal (Duncan 1972) each other on coming turn transition
3. Spoken Dialogue Systems mainly use time-out for predicting turn transition.
4. However, timeouts leads to poor user interaction(Arsikere et al, 2015)
   - Not effective in noisy environment
   - Too little - machine barge in during intra turn pause.
   - Too much - user waiting for the machine.
5. Recent research tries to use utterance's local features but still do not match human performance.
   - Syntactic (Sacks et al 1978,De Ruiter et al. 2006)
   - Prosodic (Ford 1996,Stolcke 2002,Ferrer 2003)
   - Pragmatic (Ford 2001)

Thesis Statement

1. Conversant's past behavior can help predict turn transitions
2. Past behavior is represented by summary features
   - *Relative turn length*: current turn length so far (in seconds and words) relative to the speaker's average turn length
   - *Relative Floor Control*: the speaker's control of the conversation floor (in seconds and words) relative to the total conversation length
3. Computed for each dialog act.

Section 2

Related Work

Turn taking in human-human conversations

1. (Duncan 1972) Some signals and rules for taking speaking turns in conversations
   - Intonation : terminate the final clause with rising or falling pitch.
   - Drawl on the final syllable
   - Body motion,
   - Sociocentric sequence ("you know")
   - Drop in pitch or loudness
   - Syntax - completion of grammatical clause , involving subject-predicate combination.
2. (Sacks et al, 1974) A Simplest Systematics for the Organization of Turn-Taking for Conversation
   - Introduced Turn Construction Unit (TCU) (word, phrase, clause, sentence)
   - Defined Turn Allocation Unit (TCU). Local rule system is operational in TRP (Transition Relevance Place)
     - Current speaker selects the next conversant;
     - If the current speaker did not select, any of the listeners can self select; or
     - If neither of the previous two cases apply, the current speaker continue.
3. Later research focused on local features
   - Syntactic (Sacks et al 1978,De Ruiter et al. 2006)
   - Prosodic (Ford 1996,Stolcke 2002,Ferrer 2003)
   - Pragmatic (Ford 2001)

Turn taking in Spoken dialogue systems I

1. Early systems used time out based on speech signal threshold. Fixed size of 500ms (Arsikere et al, 2015)
2. A. Raux and M. Eskenazi. (2009) - A Finite-State Turn-Taking Model for Spoken Dialog System
   - Used a 6 state non deterministic state machine to model the turn taking system between the user and the system.
   - Define a set of possible actions that both user and the system can take: Grab the floor, Release the floor, Wait while not claiming the floor, and Keep the floor.
   - Created a cost matrix to assign cost to each action in the state machine.
   - Applied probabilistic decision theory principle of selecting the action with lowest expected cost
   - Improvement of 29.5% over the fixed timeout.
3. Selfridge and Heeman (2010) - A Bidding Approach to Turn-Taking
   - Utterance importance is a driving force behind turn-taking
   - Conversant measures the importance of their potential contribution when negotiating the right to the conversation floor.
   - Use Reinforcement Learning to map a given situation to the optimal utterance and bidding behavior
   - In relation to our work, conversants might use summary feature to signal their bid.

Turn taking in Spoken dialogue systems II

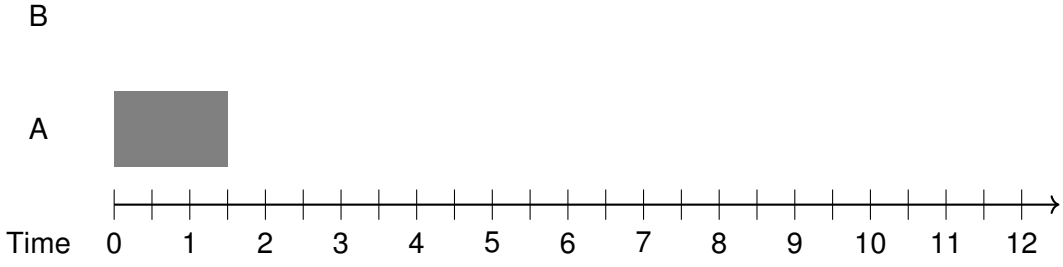
1. Gravano and Hirschberg (2011) - Turn-taking cues in task-oriented dialogue
   - Impressive study of local utterance features as cues (signals) for turn taking.
   - Formalized and verified Ducan work on signaling.
   - Defined an IPU - Inter Pausal Unit. Max Sequence of words surrounded by silence.
   - Tested over 200 features. prosodic , syntactic , IPU duration, speaking rate.
   - Found the combining features can improve turn transition prediction.
   - Our work can complement the local features.
2. N. Guntakandla and R. Nielsen (2015) - Modelling turn-taking in human conversations
   - Used dialog act as local feature.
   - Tested prediction based on Bigram, Trigram and 4-gram of dialog acts.
   - We based our base models on current and previous dialog acts.
   - We used the same corpus, but mapped the dialog act types from 148 to 9.

Section 3

Theoretical Model

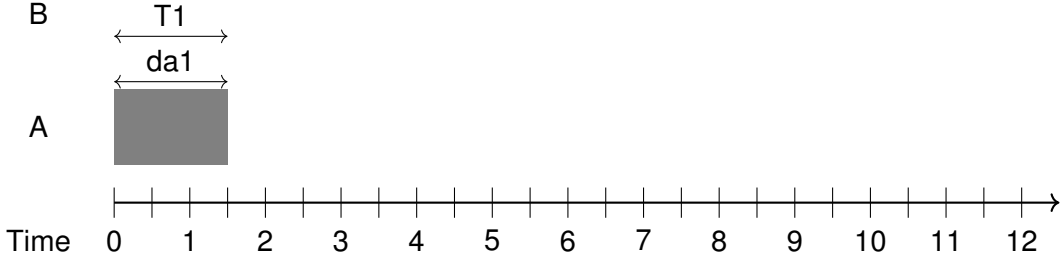
Relative Turn Length

Timing Diagram



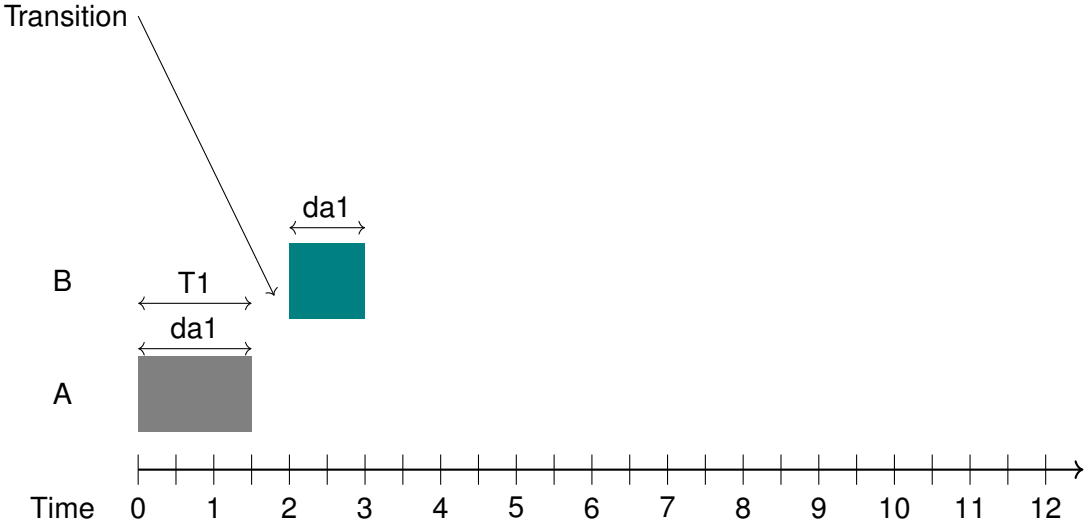
Relative Turn Length

Timing Diagram



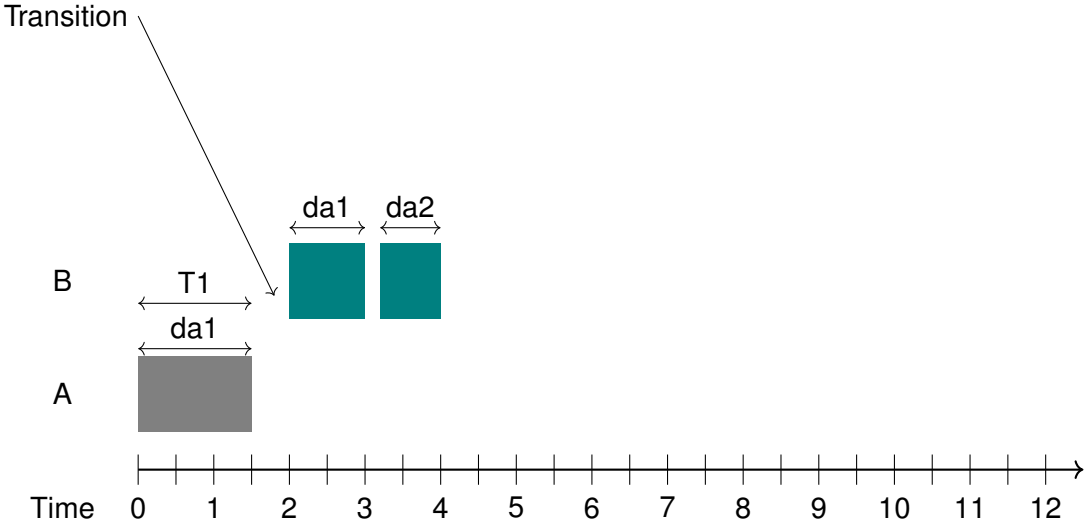
Relative Turn Length

Timing Diagram



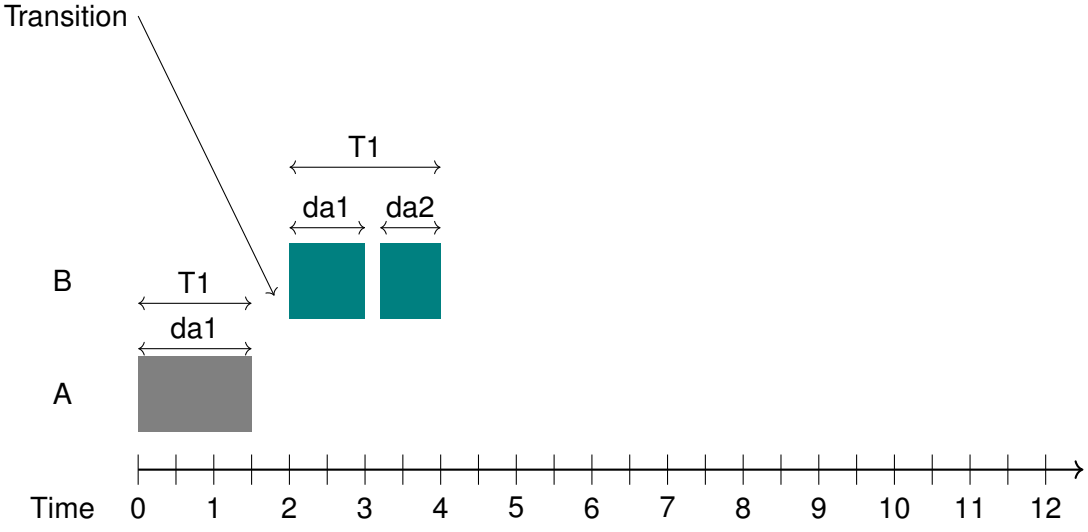
Relative Turn Length

Timing Diagram



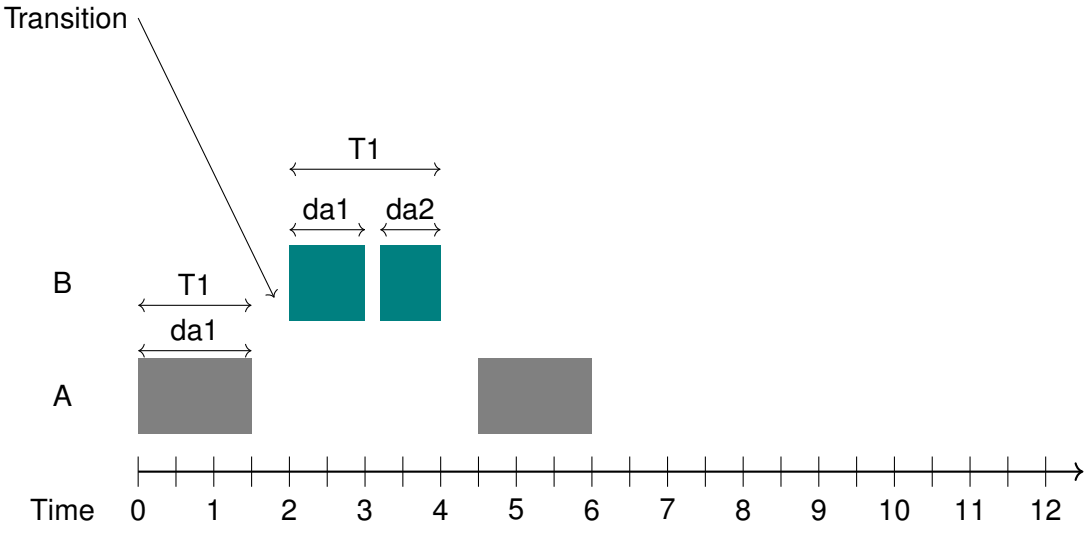
Relative Turn Length

Timing Diagram



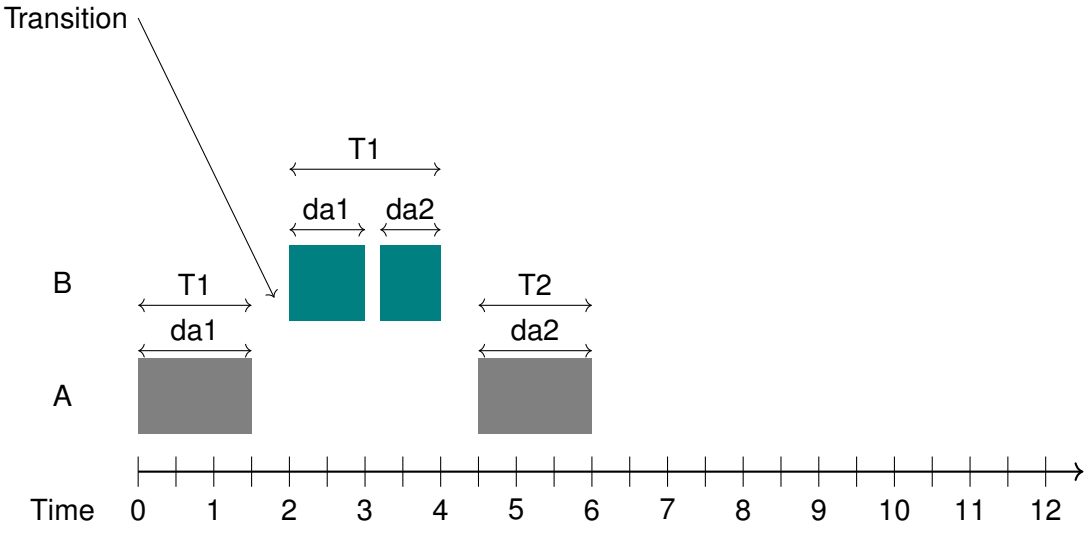
Relative Turn Length

Timing Diagram



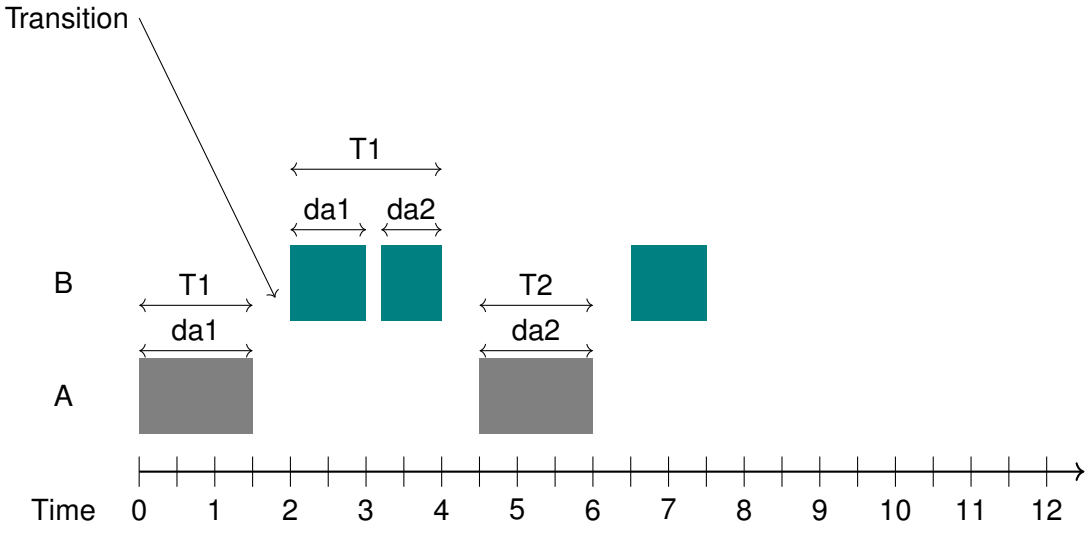
Relative Turn Length

Timing Diagram



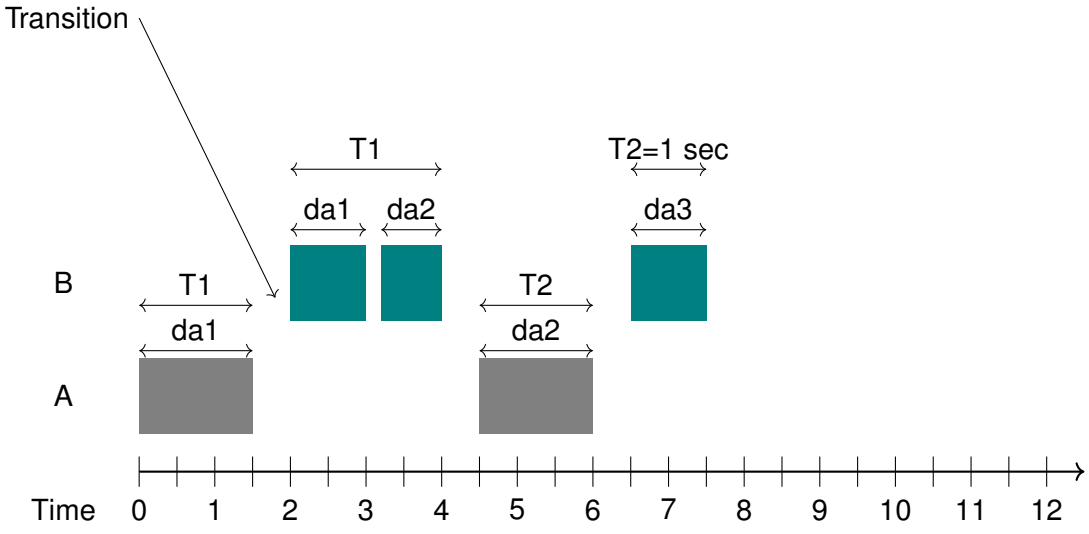
Relative Turn Length

Timing Diagram



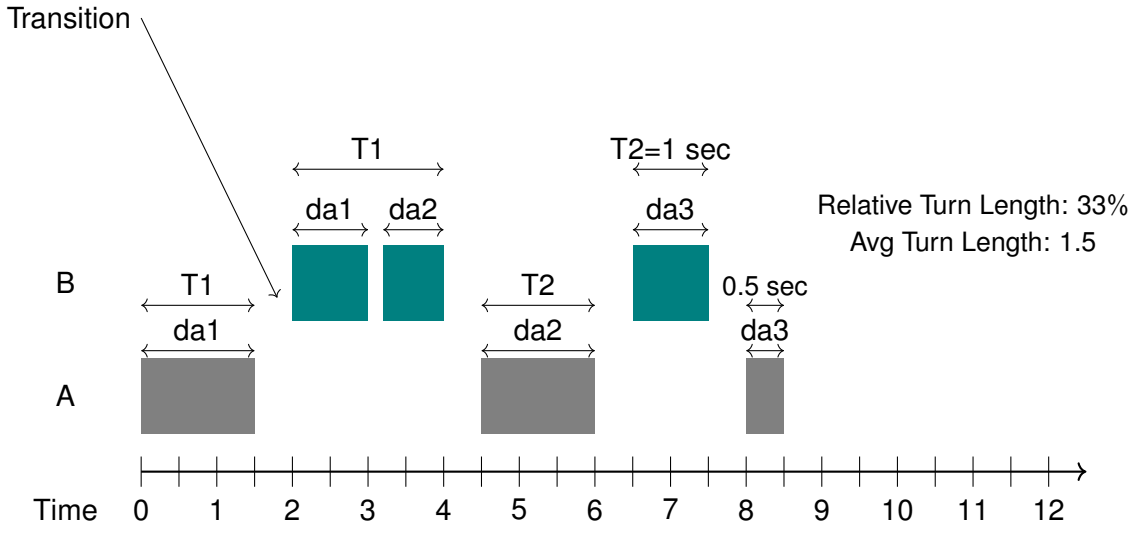
Relative Turn Length

Timing Diagram



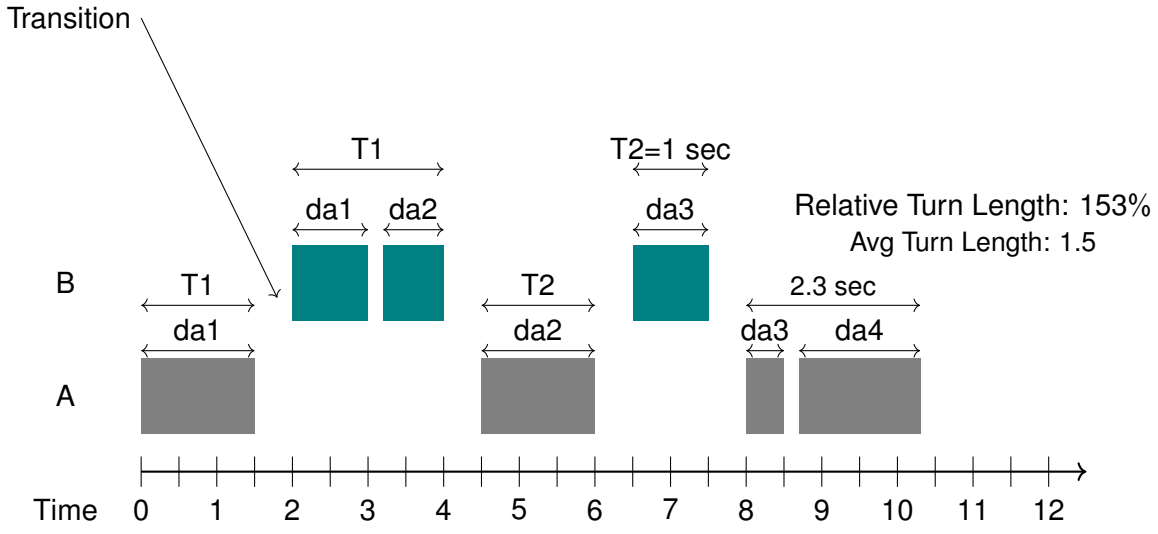
Relative Turn Length

Timing Diagram



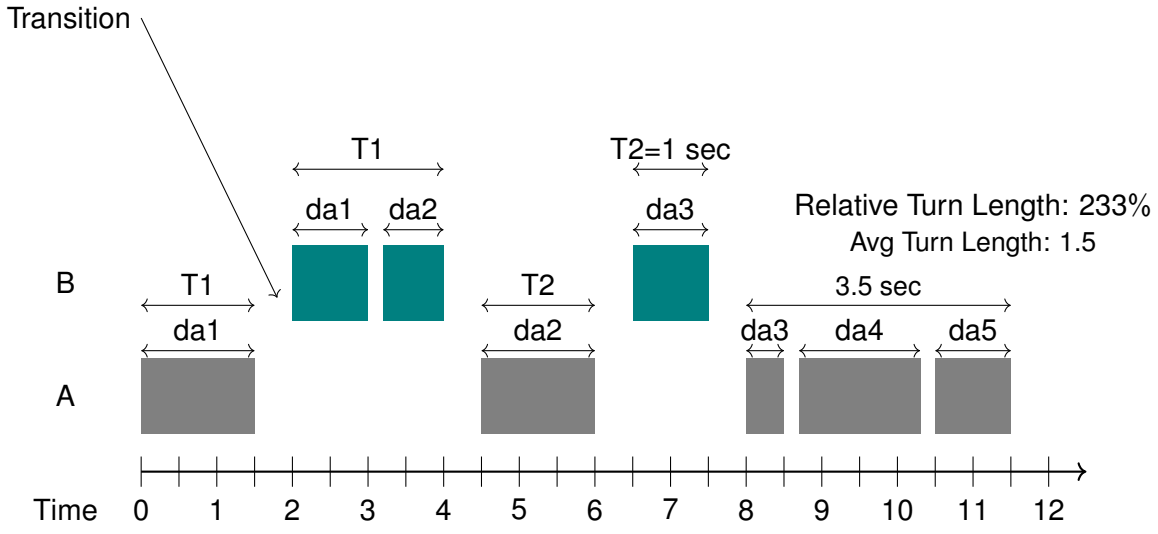
Relative Turn Length

Timing Diagram



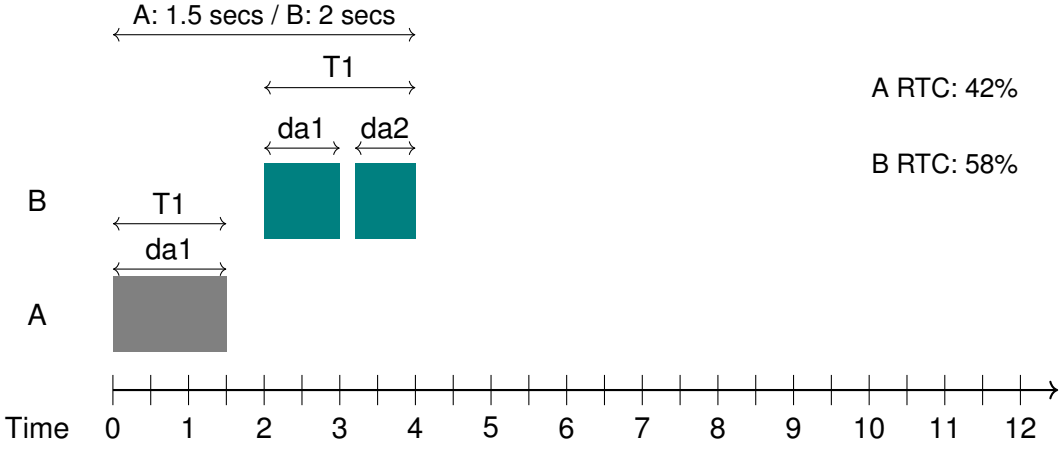
Relative Turn Length

Timing Diagram



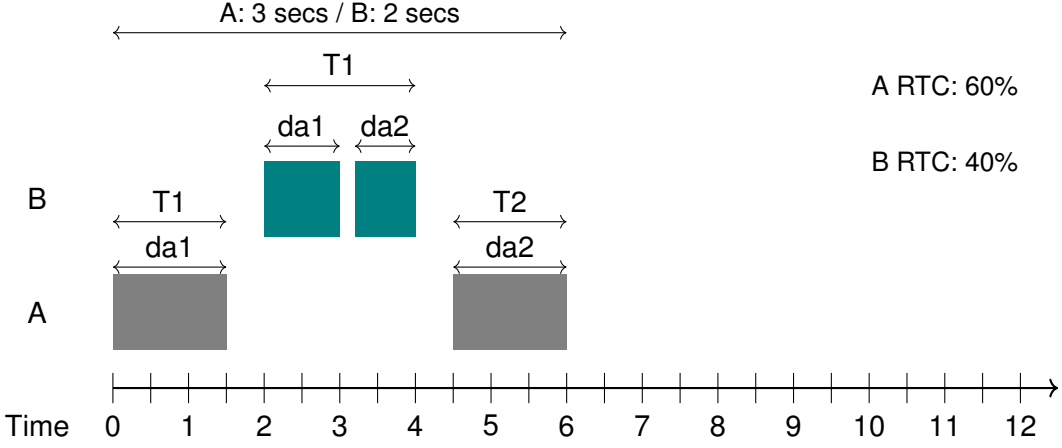
Relative Floor Control

Timing Diagram



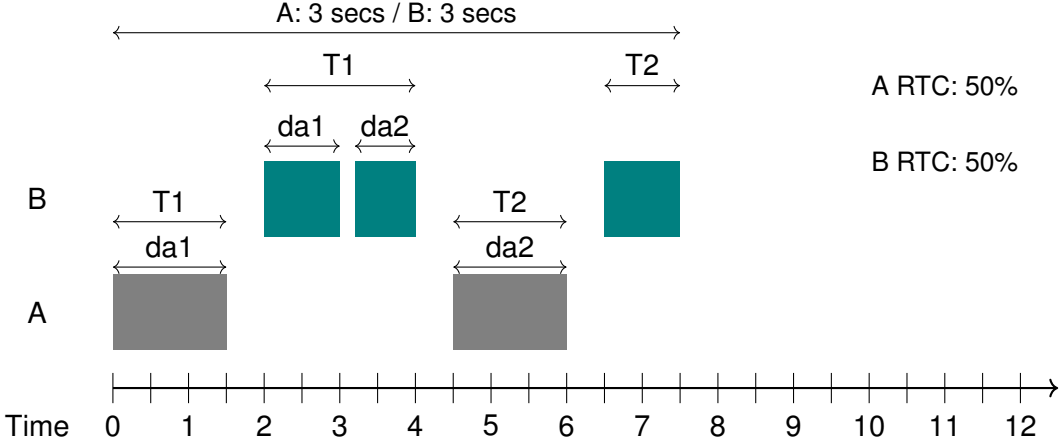
Relative Floor Control

Timing Diagram



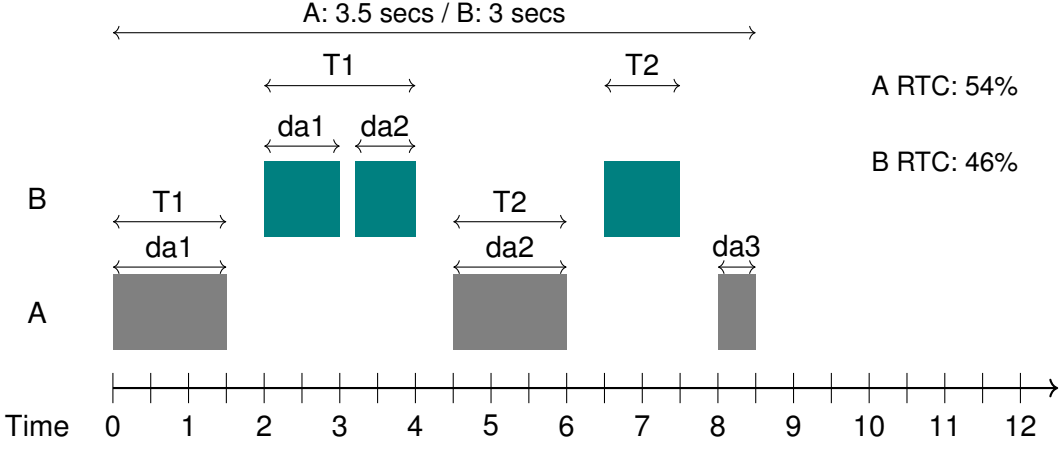
Relative Floor Control

Timing Diagram



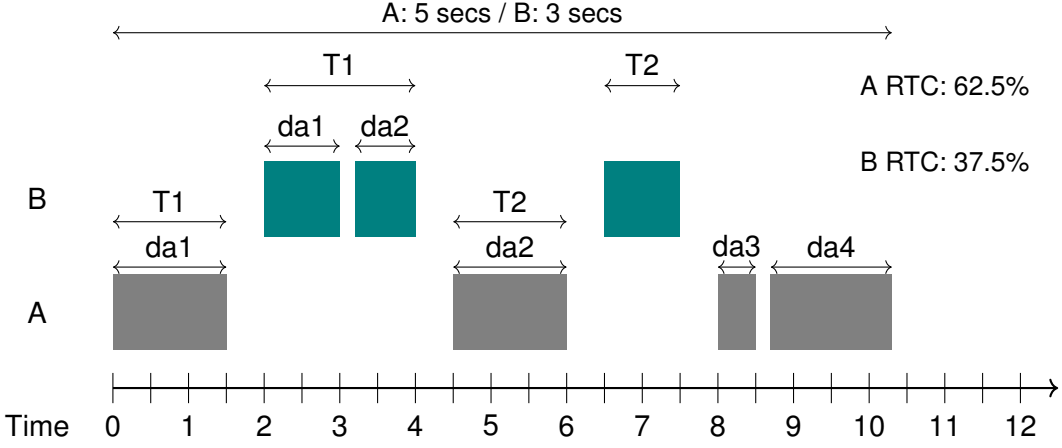
Relative Floor Control

Timing Diagram



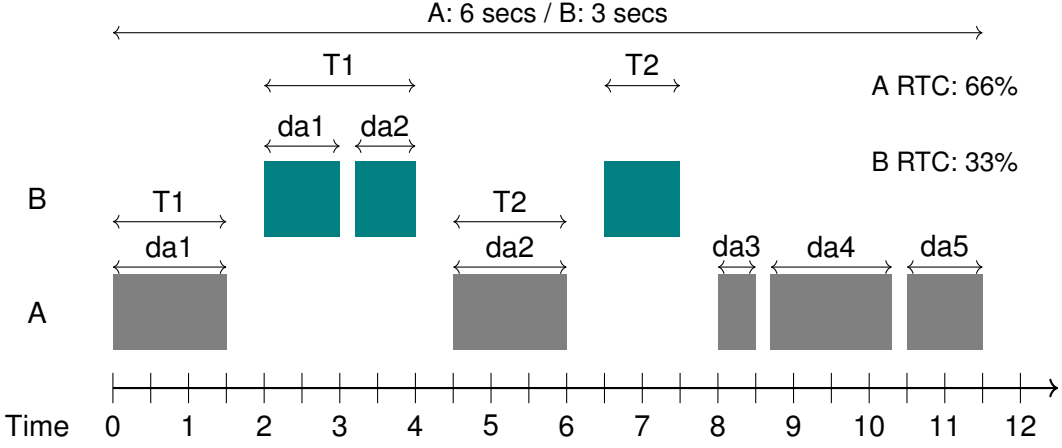
Relative Floor Control

Timing Diagram



Relative Floor Control

Timing Diagram



Section 4

Data Preparation

Corpus

1. Switchboard corpus . Originally recorded in 1991.

Corpus

1. Switchboard corpus . Originally recorded in 1991.
2. Audio recording of casual conversations between randomly chosen speakers.

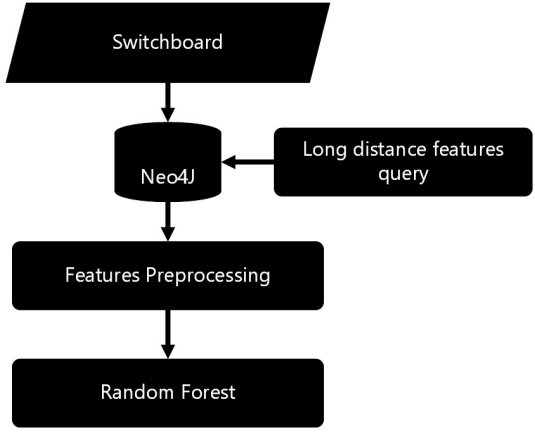
Corpus

1. Switchboard corpus . Originally recorded in 1991.
2. Audio recording of casual conversations between randomly chosen speakers.
3. 2483 conversation, involving 520 speakers

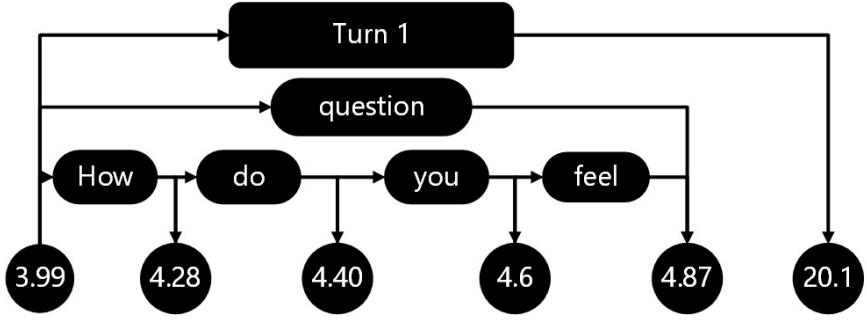
Corpus

1. Switchboard corpus . Originally recorded in 1991.
2. Audio recording of casual conversations between randomly chosen speakers.
3. 2483 conversation, involving 520 speakers
4. In our research we used the NXT version (S. Calhoun, 2010) of the corpus which contain 642 annotated conversations (XML)

Preprocessing pipeline



Conversation representation



Preprocessing

- Removed 11 dialogue acts that were coded as other in switchboard.
- Reduce data sparsity by collapsing 65 dialog acts into 9.
- Performed using python-pandas.

Switchboard dialog acts	Dialog act classes
sd,h,bf	statement
sv,ad,sv@	statement - opinion
aa,aaˆ	agree accept
%.%-,%@	abandon
b,bh	backchannel
qy,qo,qh	question
no,ny,ng,arp	answer
+	+
o@,+@	NA

Table: Mapping from dialog act to dialog act class

Section 5

Data Exploration

Overview

1. Want to understand distribution of the input variables.

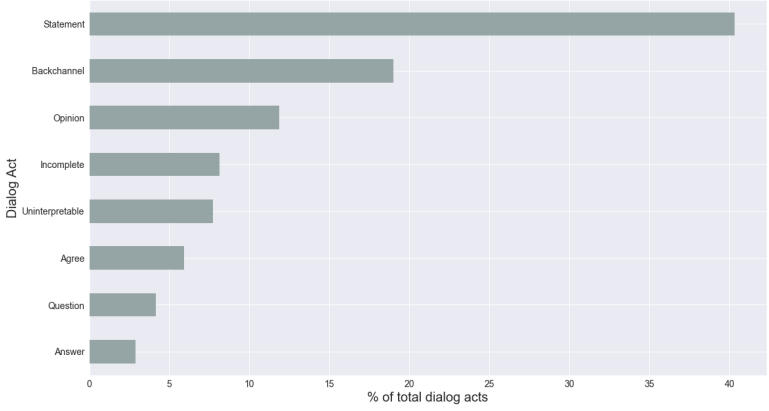
Overview

1. Want to understand distribution of the input variables.
2. Want to understand correlations between the input features and outcome (turn transition)

Overview

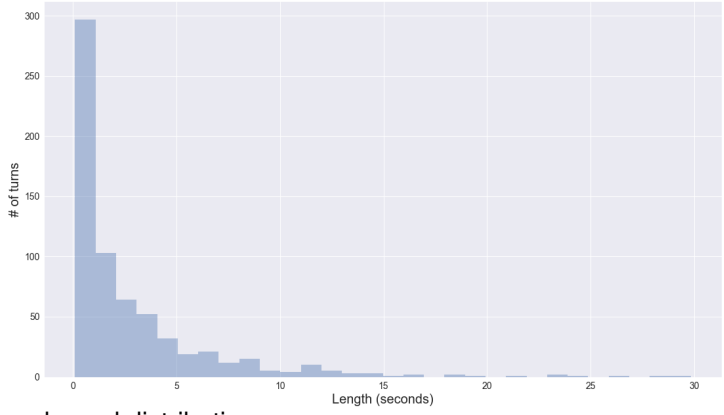
1. Want to understand distribution of the input variables.
2. Want to understand correlations between the input features and outcome (turn transition)
3. Done using python pandas for data preparation and python seaborn for data visualization

Dialog act relative count



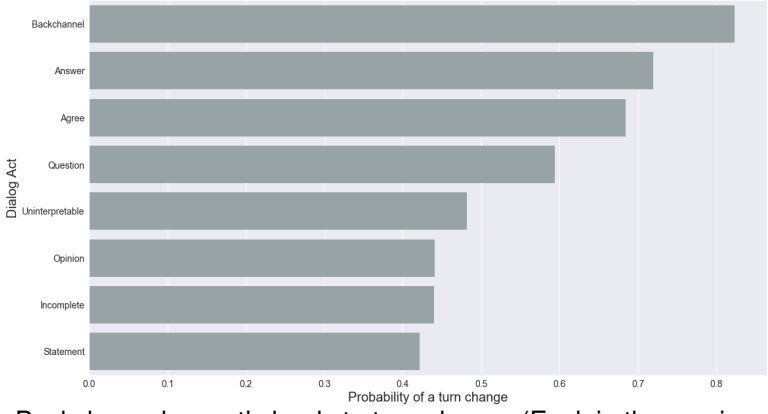
- Mainly Statements and Backchannels.
- Representative of casual conversations.

Turn Length Distribution



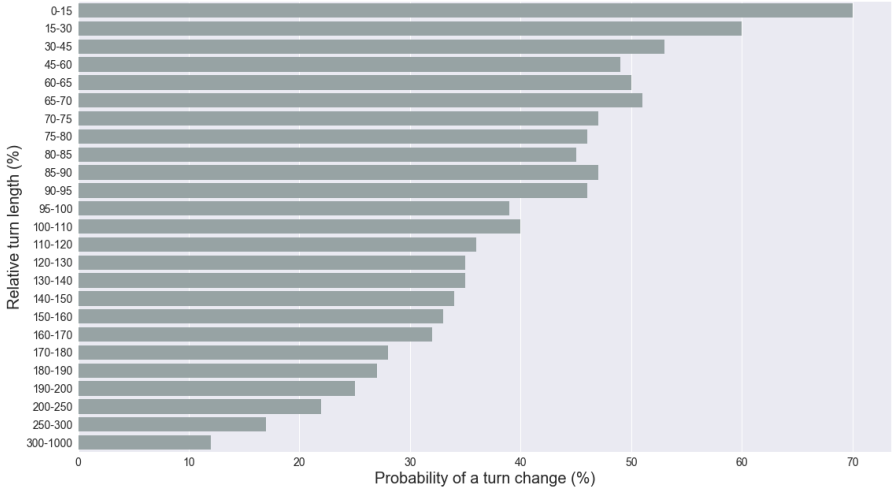
- Very skewed distribution
- Long flat tail

Dialog act probability of turn change



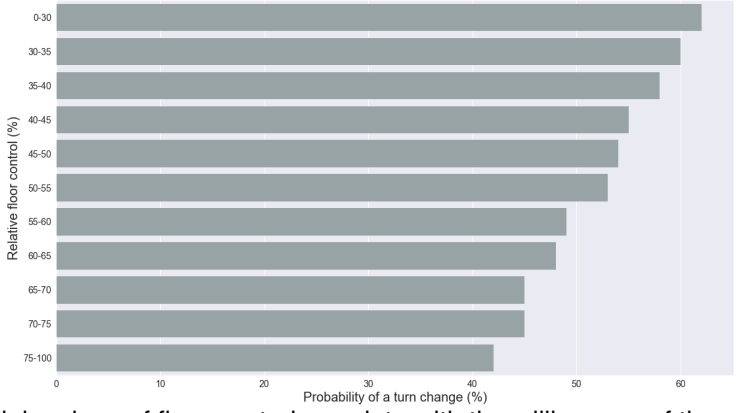
- Backchannels mostly leads to turn change (Explain the previous slide)

Relative Turn Length effect on probability of turn change



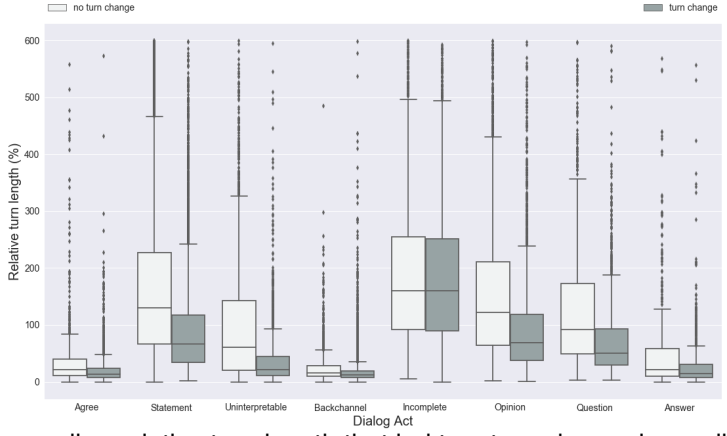
- Dialog act with small relative length lead to turn change.
- As the speaker has the floor for more time, the speaker tends to hold it.

Relative Turn Control effect on probability of a turn change



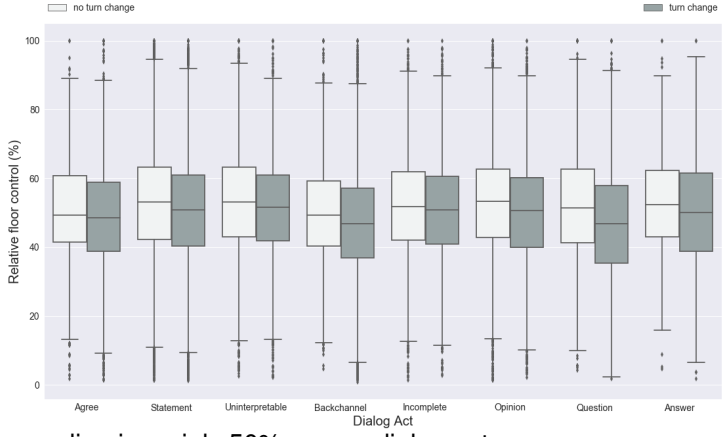
- High values of floor control correlate with the willingness of the current speaker to give up the floor.

Relative turn length for dialog act type



- The median relative turn length that led to a turn change is smaller than when it does not.
- High RTL do not lead to turn change. Holds across dialog acts

Relative floor control by dialog act



- The median is mainly 50% across dialog acts
- The median for relative floor control is slightly higher for each dialogue act when it not followed by a turn change, than when it is

Exploration Summary

1. The chance of turn change is higher when the speaker has the floor for shorter than its average turn

Exploration Summary

1. The chance of turn change is higher when the speaker has the floor for shorter than its average turn
2. Contradict our initial assumption (that high RTL will lead to turn change)

Exploration Summary

1. The chance of turn change is higher when the speaker has the floor for shorter than its average turn
2. Contradict our initial assumption (that high RTL will lead to turn change)
3. Possible explanation

Exploration Summary

1. The chance of turn change is higher when the speaker has the floor for shorter than its average turn
2. Contradict our initial assumption (that high RTL will lead to turn change)
3. Possible explanation
   - for small relative turn length, this is due to short turn with single dialog act which is likely to be back channel or an answer, both of which have low relative

Exploration Summary

1. The chance of turn change is higher when the speaker has the floor for shorter than its average turn
2. Contradict our initial assumption (that high RTL will lead to turn change)
3. Possible explanation
   - for small relative turn length, this is due to short turn with single dialog act which is likely to be back channel or an answer, both of which have low relative
   - for high relative turn length, we attribute to the flat tail of turn length distribution - the chance that the current dialog act will lead to a turn change are smaller and smaller and hence the speaker will likely keep the floor

Section 6

Machine Learning Models

Classifiers

- Used random forests (N=200) / Gradient Boosting to train and test the following models
  - baseline 1: current dialog act label.
  - baseline 2: current and previous dialog acts.
  - summary model: just the summary features.
  - full model: summary features and the current and previous dialog acts.
- Used pandas for data pre processing and scikit-learn for model training and evaluation.
- Evaluation was done using 10 fold cross validation.
- Run grid search to find the optimal hyper parameters.

Result for Random Forest Classifier

	Accuracy	F1	Precision	Recall	AUC
baseline 1	62.79%	57.81%	74.98%	47.04%	65.99%
baseline 2	74.89%	74.87%	81.84%	69.00%	81.11%
summary	65.54%	69.32%	67.22%	71.36%	69.46%
full	75.75%	77.59%	77.50%	77.83%	83.78%

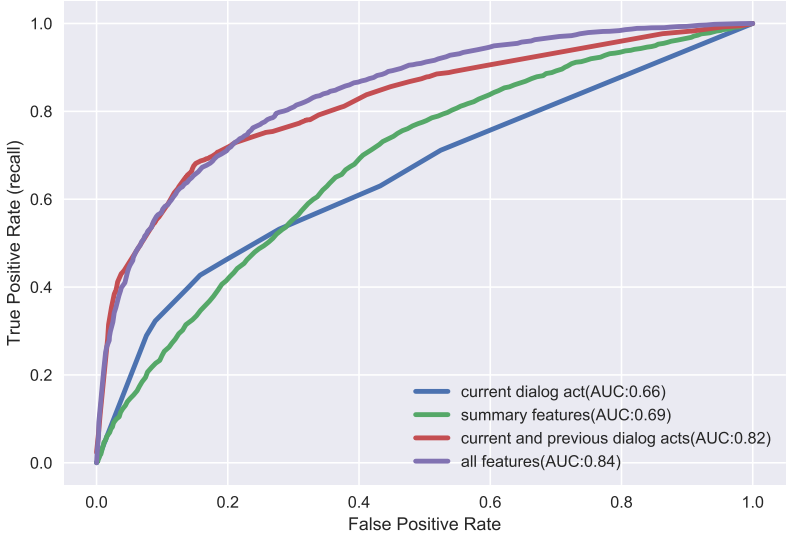
Table: Precision, recall and F1 results using Random Forests

Result for Gradient Boosting

	Accuracy	F1	Precision	Recall	AUC
baseline 1	62.79%	57.81%	74.98%	47.04%	65.99%
baseline 2	74.88%	74.82%	81.92%	68.86%	81.10%
summary	67.91%	71.30%	69.20%	73.55%	72.64%
all	76.57%	78.74%	77.44%	80.11%	84.84%

Table: Precision, recall and F1 results using Gradient boost classifier

ROC curves and AUC of different models



Sensitivity to Measurement Start Time

	0s	15s	30s	45s	60s	**120s**	180s
baseline 1	65.99%	66.10%	66.12%	66.09%	66.02%	65.98%	66.05%
baseline 2	81.11%	81.21%	81.24%	81.20%	81.15%	80.92%	80.68%
summary	69.46%	69.51%	69.43%	69.49%	69.57%	69.10%	69.21%
full	83.78%	83.87%	83.85%	83.80%	83.61%	83.19%	82.80%

Table: AUC Score in relation to the start of the dialog

Section 7

Summary

Conclusion

1. Summary features do provide improvement over local features.

Conclusion

1. Summary features do provide improvement over local features.
2. However, the affect for our data is the opposite of our initial assumption

Conclusion

1. Summary features do provide improvement over local features.
2. However, the affect for our data is the opposite of our initial assumption
   - Short turn (Low RTL) leads to turn change

Conclusion

1. Summary features do provide improvement over local features.
2. However, the affect for our data is the opposite of our initial assumption
   - Short turn (Low RTL) leads to turn change
   - In long turn the speaker will actually hold the floor.

Future work

1. Combine the summary features with other local features (semantic/prosadic)

Future work

1. Combine the summary features with other local features (semantic/prosadic)
2. Test the hypothesis on another type of corpus (for example task based corpus)

Future work

1. Combine the summary features with other local features (semantic/prosadic)
2. Test the hypothesis on another type of corpus (for example task based corpus)
3. Instead of measuring the affect from the start of the conversation, use moving averages with different window length.

Future work

1. Combine the summary features with other local features (semantic/prosadic)
2. Test the hypothesis on another type of corpus (for example task based corpus)
3. Instead of measuring the affect from the start of the conversation, use moving averages with different window length.
4. Perform the experiments where back channels are not considered as turn change.

Future work

1. Combine the summary features with other local features (semantic/prosadic)
2. Test the hypothesis on another type of corpus (for example task based corpus)
3. Instead of measuring the affect from the start of the conversation, use moving averages with different window length.
4. Perform the experiments where back channels are not considered as turn change.
5. In general, any local features can be turned into a summary feature by taking the avarage over past turn. Hence this area of research can be expanded to other local features.

Acknowledgement

1. This work was partially funded by the National Science Foundation under grant IIS-1321146.

Acknowledgement

1. This work was partially funded by the National Science Foundation under grant IIS-1321146.
2. This thesis is based on a paper that was submitted and presented at Interspeech 2016.

Acknowledgement

1. This work was partially funded by the National Science Foundation under grant IIS-1321146.
2. This thesis is based on a paper that was submitted and presented at Interspeech 2016.
3. Thesis advisor and collaborator : Prof. Peter Heeman.