Using Past Speaker Behavior to Better Predict Turn Transitions

Tomer M. Meshorer
B. A. Computer Science, Tel Aviv University, 1997

Presented to the
Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree
Master of Science
in
Computer Science & Engineering

March 2017

Center for Spoken Language Understanding
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the M. S. thesis of

Tomer M. Meshorer

has been approved.

Prof. Peter A. Heeman, Thesis Advisor
Research Associate Professor

Steven Bedrick
Research Assistant Professor

Stephen Wu
Research Associate Professor

Dedication

I would like to dedicate this work to my father, Israel Sagie ($1945-2016$), who passed away during the preparation of this work. His wisdom and support will always guide me.

Acknowledgements

I would like to express my deep appreciation to Prof. Peter Heeman for his guidance during this research. Without his valuable assistance, I could not have completed this work.

I would also like to express my appreciation to the members of the committee, Prof. Steven Bedrick and Prof. Stephen Wu, for reviewing the work and providing invaluable feedback.

I am also indebted to the staff and students of the Center of Spoken Language Understanding, and especially to the gradate program coordinator, Ms. Patricia Dickerson, for her valuable assistance during the course of my studies. This work was funded by the National Science Foundation under grant IIS-1321146.

Contents

Abstract

Using Past Speaker Behavior to Better Predict Turn Transitions

Tomer M. Meshorer

Master of Science
Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine

March 2017
Thesis Advisor: Prof. Peter A. Heeman

Conversations are at the core of everyday social interactions. The interactions between conversants are preformed within the realm of a sophisticated and self-managed turn taking system. In human conversations, the turn taking system supports minimal speaker overlap during turn transitions and minimum gaps between turns. Spoken dialogue systems are a new form of conversational user interface that permits users to use their voice to interact with the computer. As such, the turn taking capabilities of SDS should evolve from a simple timeout to a more human-like model. Recent advances in turn taking systems for SDS use different local features of the last few utterances to predict turn transition.

This thesis explores using a summary of past speaker behavior to better predict turn transitions. We believe that the summary features represent an evolving model of the other conversant. For example, speakers who typically use long turns will be likely to use long turns in the future. Moreover, speakers with more control of the conversation floor will be unlikely to yield their turn often. As the conversational image of the speaker evolves with the conversation, other speakers might adjust their turn taking behavior in response.

We computed two types of summary features that represent the current speaker's past turn-taking behavior: *relative turn length* and *relative floor control*. Relative turn length measures the current turn length so far (in time and words) relative to the speaker's average turn length.

Relative floor control measures the speaker's control of the conversation floor (in time and words) relative to the total conversation length. The features are recomputed for each dialog act based on past turns of the speaker within the current conversation. Using the switchboard corpus, we trained two models to predict turn transitions: one with just local features (e.g., current speech act, previous speech act) and one that added the summary features. Our results shows that using the summary features improves turn transitions prediction.

Part I

Introduction

Chapter 1

Introduction

Conversations are a frequent form of everyday social interaction and are characterized by rapid exchange of messages between the conversants. The turn exchange system in human conversation is universal in nature and crosses culture, age and language [20]. According to the seminal work by Sack et al. [33], a single speaker controls the conversation floor the majority of time in a conversation, conversants takes turns, and the gap and overlap between turns is kept to a minimum. These attributes apply regardless of turn length (from a single word to a full sentence) and conversation length.

In this work we are interested in the point of possible transition between speakers (the point of turn transition). To keep the gaps between turns minimal while supporting speaker change, the listener must predict a possible turn transition before the end of the speaker's utterance. Two dominating approaches tried to explain the mechanisms by which the conversation floor is allocated. [6] suggested that the speaker signals the listener about an upcoming turn transition by using a combination of one or more signals. Another approach suggested by [33] defines the turn allocation process in terms of a set of local rules that are operated at possible turn transition points. Both approaches (signaling and turn allocation rules) are based on phenomena that occur in the last few utterances (for example, the syntactic construct of the turn or the use of an adjutancy pair). This work investigates whether using features that were computed over all past turns can help to improve the predictability of turn transition.

Spoken dialogue systems (SDS) are computer systems that support a conversional speech-based user interface. Users engage in a conversation with the computer to perform a task (for example, information seeking) by using their natural apparatus, the voice. Hence, to be effective and user friendly, an SDS should also adhere to the delicate system of turn exchange.

Early SDS systems did not contain a turn management component and instead used a fixed timeout to detect the end of a user turn. Using a simple timeout led to barge-in situations, in which the system prematurely started to speak during the user's turn. Decreasing the barge-ins by

increasing the timeout caused large gaps between turns, where the user waited for the system to speak. To improve this situation, newer SDS systems are incorporating recent findings in human-human conversation in their prediction models. For example, features from the latest utterance are used to predict turn transition. Prediction in human-human conversation is based syntactic [33, 5], prosodic [10, 7, 8], pragmatic [9].

While local features of the latest utterance form an important input for prediction, this thesis postulates that speakers might also use summary features computed over many turns. The summary features form a conversational image of the speaker and contain features that represent the speaker's average behavior over many turns. For example, average turn length measures the length in both time and words of each converstants turn up to this turn. Hence, if the length of the current speaker's turn is more than its average turn length, it is more likely that a turn ending will occur.

1.1 Thesis Statement

This work investigates whether summary features based on past speaker behavior in a conversation can help improve turn transition prediction. Using the switchboard corpus, we computed for every dialog act in the conversation two summary features: *relative turn length* of the current turn, and the *relative floor control*. In addition, after each dialog act in the conversation, we marked whether a turn transition occurred. Our hypothesis is that we can make better turn transitions by using the summary features, compared to using only local features.

1.2 Approach

To test the effectiveness of the summary features, we used the NXT version of the Switchboard corpus [4, 12] to train random forest models [26]. The baseline models were trained on local features: current and previous dialog acts. We also trained a model on the summary features as well as a model that includes both the local and the summary features. The summary features are based on measurements of each speaker's behavior over the preceding turns in the dialogue.

1.2.1 Local Features

We define a conversation as a sequence of dialogue acts $d_1 \dots d_N$, where d_i is uttered by speaker s_i. We write this as the following sequence:

$$\dots s_{i-2}, d_{i-2}, s_{i-1}, d_{i-1}, s_i, d_i \dots \tag{1.1}$$

We denote whether there was a turn transition with y_i. A turn transition occurs when the speaker s_i is different from speaker s_{i-1}. Hence, (1) can be also be viewed as a sequence of dialog acts d_i followed by turn transitions y_i:

$$\dots d_{i-2}, y_{i-1}, d_{i-1}, y_i, d_i, y_{i+1} \dots \tag{1.2}$$

In our first baseline model, we try to predict the turn transition value y_{i+1} based only on the latest dialog act d_i. In our second baseline model, we try to predict turn transition y_{i+1} based on the latest two dialog acts: d_{i-1} and d_i.

1.2.2 Summary Features

As discussed in the introduction, we introduce two types of summary features in this paper: relative turn length (rt_i) and relative floor control (rc_i). These features are used in predicting whether there is a change in speaker y_{i+1} after dialogue act d_i.

To compute the summary features, at dialogue act d_i, we denote S_i to be the set of complete turns of speaker s_i that are prior to the turn that d_i is in. Let t_i represent the turn so far that d_i is in, up to d_i but no subsequent dialogue acts. Let length(t) be the length of a turn or a partial turn in seconds (or words). To compute the *relative turn length* of turn t_i we first compute the average length of all the turns in S_i

$$avg\_t_i = \frac{\sum_{t \in S_i} length(t)}{|S_i|} \tag{1.3}$$

The *relative turn length* summary feature of t_i, denoted as rt_i, measures the percent of the length of the turn t_i so far, relative to the average turn length up to t_i of the current speaker s_i (but not including t_i).

$$rt_i = \frac{length(t_i)}{avg\_t_i} \tag{1.4}$$

Note that we calculate two versions of rt_i: in seconds and in words. The purpose of rt_i is to let the listener, in predicting turn changes, take into account whether the current speaker is exceeding his or her average turn length.

The *relative floor control*, denoted as rc_i, measures the percent of time in which the current speaker controlled the conversation floor up to d_i. We again define S_i as above, and we define L_i to be the turns of the other conversant (the listener of d_i). We first compute the conversation length up to d_i denoted as c_i which excludes inter-utterance pauses.

$$c_i = \sum_{t \in S_i \cup L_i} length(t) \tag{1.5}$$

To compute relative floor control at d_i, we divide the floor time of the speaker s_i up to turn t_i by c_i:

$$rc_i = \frac{\sum_{t \in S_i} length(t)}{c_i} \tag{1.6}$$

Note that we calculate rc_i in seconds and in words. Participants can use the relative floor control as a means to determine if one speaker is controlling the conversation; a controlling speaker will probably be less inclined to give up the floor.

We use these two summary features in the *summary model* and *full model*, as described in the next section.

1.3 Dissertation Structure

Chapter 1 contains the introduction, the definition of the summary features and the thesis statement. Chapter 2 presents the related work in both human-human and human-computer conversations. Chapter 3 describes the study which compute the long term features over all the turns in each conversations. The chapter describes the experiment and includes the description of the corpus, the data preparation pipeline, and the result, which compares the baseline models as well as the model that includes the summary features. Finally, in chapter 4 we present the conclusion and future work.

1.4 Contribution

Our results show that the summary features improve prediction performance in both area under the curve (AUC), 69% vs 65.5%, and F1, 65.34% vs 62.24%. In addition, the model that was trained on all of the features performs better than the local features model in both AUC, 83.51% vs 81.08%, and F1, 75.45% vs 74.82%. The first study results show that using the summary features can help predict turn transitions.

In the second study, our results show that as you shorten the summary features the prediction performance increase..

Chapter 2

Related Work

This section presents work related to turn transition prediction in both human-human conversations and human-computer conversations.

2.1 Human-Human Conversations

Duncan [6] argued that speakers signal when they want the listener to take the turn and presented six signals used by the speaker to accomplish this: intonation, drawl on the final syllable, body motion, sociocentric sequence, drop in pitch or loudness, and syntax. Kendon [18] added gaze as a signal to turn transition. Our summary features complement the set of signals as suggested by [6].

Turn allocation was introduced in the seminal work by Sacks, Schegloff, and Jefferson [33], who observed that conversations are "one speaker at a time" and gaps between turns as well as speaker overlaps are kept to a minimum. To satisfy these constraints, Sacks et al. suggested an ordered set of rules for turn allocation: (a) current speaker selects the next conversant; (b) if the current speaker did not select, any of the listeners can self select; or (c) if neither of the previous two cases apply, the current speaker continues. For the first rule, Sacks et al. suggested that the current speaker uses adjacency pairs as the main apparatus for selecting the next speaker. Hence, we recognized the importance of dialog acts in turn allocation and chose them as the atomic turn components. In addition, our work might impact the second rule, in which the conversant self selects. While Sacks et al. suggested that the first starter is the next speaker, we suggest that a conversant might use the conversational image of the speaker and of themselves when self selecting. For example, a controlling speaker (with a high relative floor control score) has a better chance to gain control of the conversation floor when self selecting. The work on turn bidding is also related [36], which suggested that each conversant measures the importance of their utterance when negotiating the right to the conversation floor.

In addition to the turn allocation system, Sacks et al. also suggested that turn construction units (TCU) should support projection of turn ends by the participants. The projectability attribute was later extended to other features of the speaker's utterance: (syntactic [33], prosodic [10] and pragmatic [10, 9]). Our work augments the local utterance features with summary features that can be used to improve projectablity.

Entrainment was presented in [43], which suggested entrainment of endogenous oscillators in the brains of the speaker and the listener on the basis of the speaker syllabus production. In their study, the speaker and the listener are counter-phased such that speech overlaps and gaps are minimized. Although our work does not imply cyclic synchronization between speaker and listener, we do suggest that each conversant creates a conversation image of the other conversant and uses it during turn transition.

The importance of using dialog acts was emphasized by a very recent study of Garrod and Pickering [11]. The study suggested that turn production is a multi-stage process in which the listener performs simultaneous comprehension of the existing turn as well as production of the new turn content. They suggested that the first step in the process is dialog act recognition, which is done as soon as possible and acts as the basis for the listener's turn articulation and production. In our study we use dialog act as the main turn component.

2.2 Human-Computer Conversations

As recent advances in machine learning [17] reduce speech recognition error rates, the problem of turn taking in SDS rises in importance. Traditional SDS systems use a simple silence timeout approach to trigger turn transitions. This creates three issues [1] first, the model might not be robust enough in a noisy environment (for example when driving); second, if the timeout is too short, the system might detect intra-turn pauses (for example, the user pausing to think) as a turn transition and will cut into the user's turn; and third, if the timeout is too long the system will wait too long to take the turn, resulting in large gaps between turns.

Recent studies tried to improve over the simple threshold model by using machine learning to train models based on features derived from the last utterance. As different studies use a variety of features, we will outline those that used counting features that are close to the summary features.

Arsikere et al. [2] focused on utterance segmentation in the context of an incremental dialog system. Using the switchboard corpus, they used a decision tree algorithm to decide if a word is the final utterance by using various features and in particular the number of words in the turn so far. The usage of count features improves precision by 10% but has very low recall (7%), which

might have occurred, according to the author, from turns with only one word.

Gravano and Hirschberg [14] used the Columbia games corpus to study the effectiveness of different turn transition cues. The authors define inter-pausal units (IPU) as a maximum sequence of words surrounded by silence of more than 50 ms. A turn is the longest sequence of IPUs by the same speaker. One of the features studied is IPU duration in ms as well as number of words. As in our findings, the authors found that long IPUs are a good indication of upcoming turn changes (long IPUs might correlate with a speaker passing its average turn length). Moreover, as we show in Section 5, the authors found that combining multiple cues leads to better accuracy.

Raux and Eskenazi [29] performed a comprehensive study of features that inform turn changes. The study found that timing features, such as turn duration and number of pauses, have relatively strong predictive power. While Raux and Eskenazi use features of the current turn, in our study we use the timing features for the turns that have occurred so far in the current conversation.

Ina more recent study, Nishitha and Rodney [15] used a model based on N grams of dialog acts to predict turn transitions. They trained a decision tree model using the switchboard data and tested bigram, trigram and 4 grams models of dialog acts with and without speaker id. They achieved an F1 measure of 0.67 for the trigram model. In this paper, we based our baseline models on bigrams and trigrams of dialog acts. We also mapped the switchboard dialog acts from 148 dialog acts down to 9 to reduce data dimensionality. The prediction performance of our baseline model is comparable to their results.

Chapter 3

Study

3.1 Data

To evaluate the importance of the summary features in predicting turn transitions we used the 2010 version of the switchboard corpus [4] which is based on the original release [13].

The switchboard corpus is the first large collection of phone conversations and was collected in 1990-1991. The initial goal of the corpus is to facilitate speech research by providing pre recorded audio files of day to day conversations. The original corpus was composed of 2483 phone calls involving 520 speakers. Each call ranged from 1.5 minutes to 10 minutes, with an average length of 6 minutes. Conversations involved a randomly chosen topic between two randomly selected speakers. The corpus was later improved and was released as part of the Penn Treebank 3 corpus, which included 650 annotated conversations. The current release used for this research includes 642 conversations and just over 830000 words.

The current corpus contains multiple annotations for each conversation. We used the turns annotation to assign turns to speakers and mark the turn start and end points. We used the dialog act annotation to assign dialog acts to turns such that each turn interval contain all the dialog acts that occured between turn start and turn end. Next, we use the token annotation to assigned words to dialog acts and to compute the overall number of words within each dialog act.

3.2 Data Preparation

Figure 3.1 shows the experiment data pipeline. Data is imported from the NXT switchboard corpus [4] into a graph database [42]. Figure 3.2 shows the data structure as it is represented inside the graph database. For each conversation, the conversation entities (words, dialog acts and turns) are represented as edges between time points, which are represented as vertices. The structure leads to a direct computation of the summary features using the graph query language.

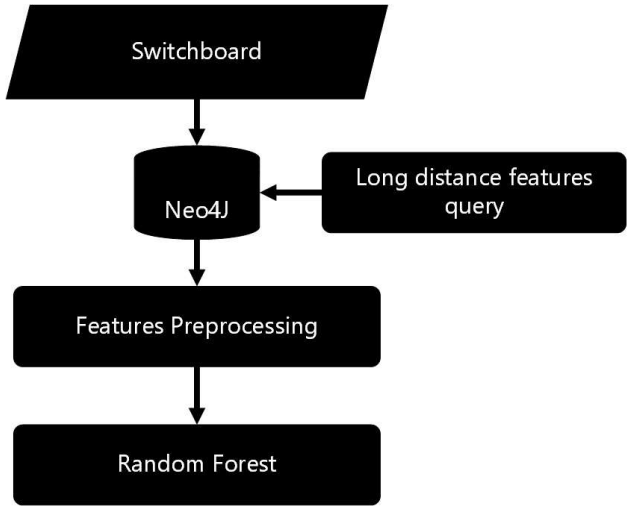


Figure 3.1: Experiment data pipeline

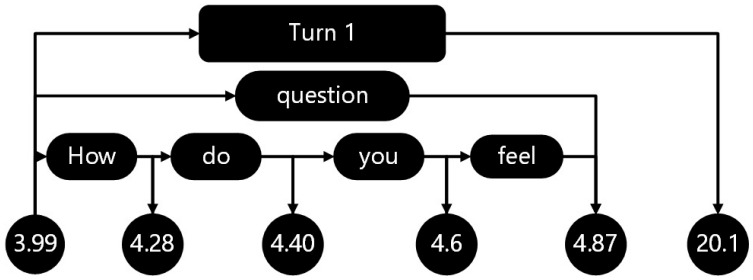


Figure 3.2: Conversation graph data model

After computing the summary features, we perform the following data transformation:

- We exclude 11 dialogue acts that were coded in Switchboard as "other."
- Since we believe that it takes a certain amount of time to build a stable conversational image, in evaluating our model, we removed all turns that occurred in the first part of each conversation. For this paper, we used an estimate of 120 seconds. This reduced the number of dialog acts from 50,633 to 37,508.
- To reduce data sparsity, we grouped switchboard dialog acts into dialog act classes. This reduced the number of dialog acts from 148 to 9 dialog act classes. See Table 1 for examples of the mapping.
- We added a binary y_{i+1} feature to each dialog act. As explained in Section 3, the variable is 1

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

Table 3.1: Mapping from dialog act to dialog act class

if there is a turn change from dialogue act d_i to d_{i+1}.

3.3 Data Exploration

Before performing the actual learning and prediction tasks, this chapter explores the data in order to understand the statistics of the different features.

3.3.1 Features

After pre processing the data the data set to be explorted is describe in the following table:

Field	Description	Type
Previous Dialog Act	dialog act before the current one	categorical
Dialog Act	current dialog act	categorical
Length	length of the current dialog act in seconds	numeric
Relative Turn Length (RTL)	The relative of the dialog act	numeric
Relative Time Control (RTC)	The relative control of the current turn in words	numeric
Turn Change	1 if there was a turn change after this dialog act	binary

Table 3.2: Data Fields

3.3.2 Dialog Acts

Figure 3.3 shows the count for each dialog act. We can observe that the majority of dialog acts are statements, backchannels and opinions. This is true to the nature of the switchboard corpus which consists mainly of casual conversations.

Figure 3.4 shows an histogram of the length in seconds for each dialog act. It is evident that the distribution is left tilted.

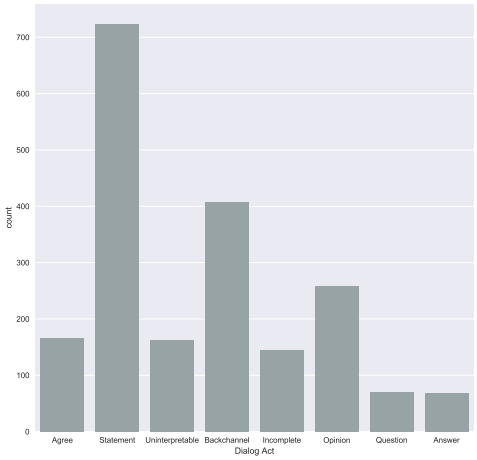


Figure 3.3: Dialog act size

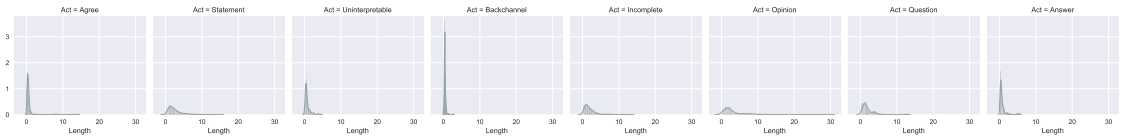


Figure 3.4: Dialog act length in words

Figure 3.5 shows the and histogram for each dialog act length conditioned on the occuruce of a turn change. We can observe that the length distribution is the same, however when a turn change occur, some dialog acts tends to be longer than regular.

Figure 3.5: Dialog act length in words conditioned on turn change

In figure 3.6 , we measure the probability that a dialog act will lead to turn change. We can observe that the majority of back channels and question will lead to a turn change.

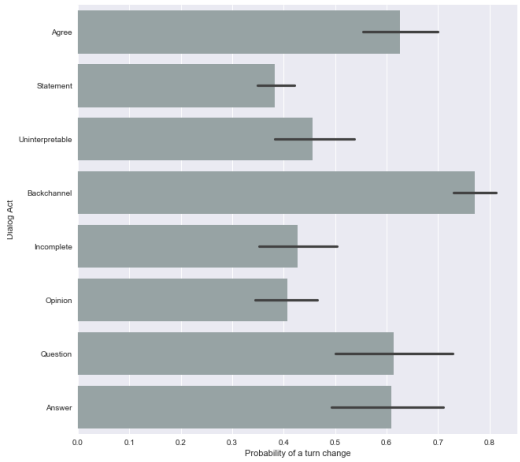


Figure 3.6: Dialog act probability of turn change

3.3.3 Relative Turn Length

Figure 3.7 shows the distribution of the relative turn length for each dialog act. We can observe that

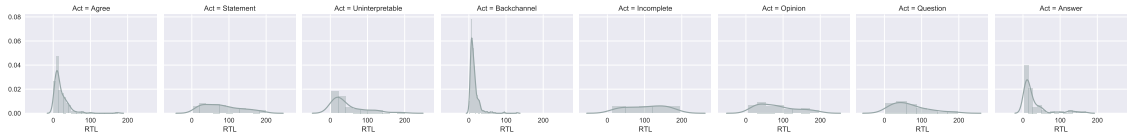


Figure 3.7: Relative turn length for each dialog act

3.3.4 Relative Floor Control

Figure 3.8 shows distribution of floor control or each dialog act conditioned on a turn change. We can observe that regardless of the dialog act, most distributions are normal with mean of 50%

3.4 Machine learning

3.4.1 Classification Models

To test the contribution of the summary features, we used a binary classifier with y_i as the outcome variable. We trained four models, which used the following sets of features:

Figure 3.8: Relative flow control conditioned on turn change

baseline 1: Predict turn transition based only on the current dialog act label.

baseline 2: Predict turn transition based on the labels of the current and previous dialog acts.

summary model: Predict turn transition using just the summary features.

full model: Predict turn transition using the summary features and the current and previous dialog acts.

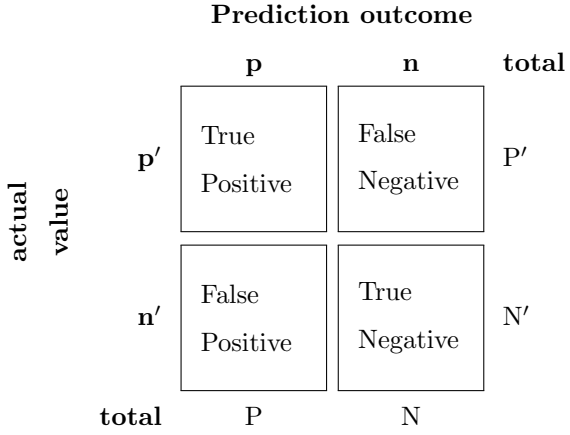
We used random forests to build the binary classifiers ($N = 200$) [3]. Random forests build an ensemble of decision trees during training, and during testing, each decision tree votes on the outcome. Like decision trees, they can account for interactions between variables, such as making greater use of the summary features when the current speech act is not a question. Random forests though are not as sensitive to overfitting and data fragmentation.

To find the optimal hyper parameters, we ran a grid search over the *max_features* and *max_depth* hyper parameters for each model. The hyper parameters search was done over $\{sqrt, log2, 10\}$ for *max_features* and $\{5, 7, 9\}$ for *max_depth*. When training the model, we used the optimal hyper parameters for each feature set.

We performed 10 fold-labeled cross validations. We made sure that each conversation was entirely in a single fold. This way, each dialogue was entirely used for training or testing, but never for both at the same time.

3.4.2 Metrics

To evaluate our hypothesis, we use the trained model to perform prediction of the test data. We than compare the results against the truth values. The results are recorded in a confusion matrix. Each row in the matrix represent the actual class and each column represent the predicted class. The cell in the matrix are compute as follow :



1. True Positive - Labels which were predicted as positive and are positive
2. False Positive- Labels which were predicted as positive but are in fact negative
3. True Negative - Labels which were predicted as negative and are negative
4. False Negative - Labels which were predicted as negative but are in fact positive

Based on the confusion matrix, we compute the following metrics for each model:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

The accuracy measure how many instances were classified correctly out of the total instances.

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

The precision metric measure how accurate is the classifier for the positive instances. I.e. how many instances that were classified as positive, were in fact positive.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

The recall metric measure how many positive instances were detected by the classifier.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3.4}$$

In order to have one measurement which encompass both recall and precision, we compute F1 which is the harmonic mean of recall and precision.

Note that there is a trade off between recall and precision. I.e. if we want to increase recall we will reduce precision. To measure this tradeoff, we will use ROC (Receiver Operating Characteristic) curve. The ROC curve measure the true positive rate (TPR) or recall, against the false

positive rate (FPR). When comparing different ROC curves, we measure the area under the curve (AUC).

3.5 Results

We first analyzed the results in terms of accuracy: how often the models correctly predicted whether a turn transition occurred; in other words, how often the model predicts the correct value of y_{i+1}. Table 2 shows the results of training a random forest for each model. We see that using the summary features provides better accuracy than baseline 1, which use only the current dialog act (66.14% vs 60.26%). In addition, using the full model yields an improvement of over 1.58% in the result.

Model	Accuracy	AUC	hyper parameters
Baseline 1	60.26%	0.63	max_features=sqrt, max_depth=7
Baseline 2	74.43%	0.79	max_features=log2, max_depth=9
Summary	66.14%	0.65	max_features=sqrt, max_depth=5
Full	76.05%	0.82	max_features=10, max_depth=9

Table 3.3: Accuracy and AUC results

The effect can also be seen in Figure 3, which shows the ROC curves and the AUC for each model. We notice that the AUC of the summary model is better than baseline model (0.65 vs 0.63), and when adding the summary features to the local features in the full model, we see the AUC improves (0.82 vs 0.79). This suggests that while the discrimination facility of the summary features is lacking (AUC < 0.7), adding them to a classifier that uses local features (full model) yields better results than the baseline.

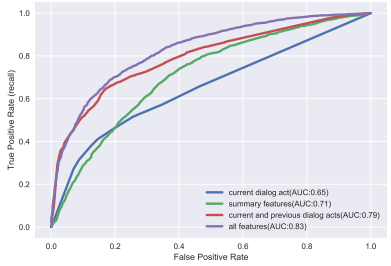


Figure 3.9: ROC curves and AUC of the different models

In addition to analyzing the results in terms of accuracy, we also analyze the results of the four models in terms of how well we predict that there is a change in speaker (i.e., y_{i+1} indicates that there was a turn switch). Table 3 shows the results in terms of recall, precision, and F1, which combines the two scores. Although baseline 1 has high precision, it has very low recall. Using

only the summary model improves recall and decreases precision by less, leading to a higher F1 score and overall better performance. Using the full model improves precision, which means that dialog acts that were considered to lead to turn transitions are classified correctly. If we use the full model, we lose precision (over baseline 2 model), but gain recall, leading to the highest F1 score and the best performance.

	Accuracy	F1	Precision	Recall	AUC
baseline 1	0.624320	0.577152	0.746507	0.470676	0.655748
summary	0.653455	0.692469	0.670798	0.714937	0.690383
baseline 2	0.748215	0.747843	0.823304	0.685486	0.810832
all	0.754587	0.775238	0.773176	0.777437	0.835127

Table 3.4: Precision, recall and F1 results

Chapter 4

Conclusions

This paper explores the use of features that capture speakers' past turn-taking behavior in predicting whether there will be a turn transition. These summary features include (a) relative turn length: how the current turn under construction compares to the current speaker's average turn length; and (b) relative floor control: the percentage of time that the current speaker has held the floor. We include two versions of each, one based on time, and one based on number of words. Relative turn length should capture whether one or both of the speakers tends to hold the turn over multiple utterances, while relative floor control captures whether one speaker is dominating the conversation. Both of these factors should influence who will speak next.

In evaluating our model on data from the Switchboard corpus, we find that our summary features on their own do better than just using the previous speech act (accuracy of 66.14% vs 60.26%). We also find that when we add these features to a model that uses the last two speech acts, we also see an improvement (76.05% vs 74.43%). These results show the potential of modeling speakers' past turn-taking behavior in predicting upcoming turn-transitions. Better modeling of turn-taking should lead to more natural and efficient spoken dialogue systems.

4.1 Future Direction

In this work, the local features that we considered in our baseline model were just the last two speech acts. Other work on turn-taking prediction use a richer set of local features, such as syntactic [6, 33, 10, 5, 24, 2], prosodic [6, 10, 38, 8, 5, 30, 29, 16, 2], pragmatic [10, 11, 29], semantic [29] and non-verbal [18]. In future work, it would be good to evaluate the contribution of our summary features with a richer set of local features.

In our work, we evaluated our model on the Switchboard corpus. In future work, it would also be good to evaluate our summary features on other corpora, especially task-based dialogues. Tasks in which there is a difference in the role of the user and speaker, such as in Trains [**?**], should

benefit from modeling the past turn-taking behavior of each speaker.

In our work, we computed the relative turn length and relative speaker control using the turn length avrage as computed over all the previous turns. In future work , it would be interesting to use simple moving average (measured over multiple window width) as well as exponential moving average.

More generally, the summary features introduced in this work represent just one aspect of the conversational image of the user. Future work should try to "summarize" other local features by creating the average value of a local feature over past turns. For example, we can compute relative speech rate, or relative use of stereotyped expressions.

Bibliography

[1] H. Arsikere, E. Shriberg, and U. Ozertem. Enhanced end-of-turn detection for speech to a personal assistant. In *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*, 2015.

[2] M. Atterer, T. Baumann, and D. Schlangen. Towards incremental end-of-utterance detection in dialogue systems. In *COLING (Posters)*, pages 11–14, 2008.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver. The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419, 2010.

[5] J. P. De Ruiter, H. Mitterer, and N. J. Enfield. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, pages 515–535, 2006.

[6] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292, 1972.

[7] L. Ferrer, E. Shriberg, and A. Stolcke. Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody. In *INTERSPEECH*, 2002.

[8] L. Ferrer, E. Shriberg, and A. Stolcke. A prosody-based approach to end-of-utterance detection that does not require speech recognition. In *ICASSP*, pages 608–611, 2003.

[9] C. E. Ford. At the intersection of turn and sequence. *Studies in Interactional Linguistics*, 10:51, 2001.

[10] C. E. Ford and S. A. Thompson. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. *Studies in Interactional Sociolinguistics*, 13:134–184, 1996.

[11] S. Garrod and M. J. Pickering. The use of content and timing to predict turn transitions. *Frontiers in Psychology*, 6, 2015.

[12] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP*, pages 517–520, 1992.

[13] J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520 vol.1, Mar 1992.

[14] A. Gravano and J. Hirschberg. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634, 2011.

[15] N. Guntakandla and R. Nielsen. Modelling turn-taking in human conversations. In *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*, Stanford CA, 2015.

[16] R. Hariharan, J. Hakkinen, and K. Laurila. Robust end-of-utterance detection for real-time speech recognition applications. In *ICASSP*, pages 249–252, 2001.

[17] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[18] A. Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[19] W. J. M. Levelt. *Speaking : from intention to articulation / Willem J.M. Levelt.* MIT Press Cambridge, Mass, 1989.

[20] S. C. Levinson. Turn-taking in human communication–origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1):6–14, 2016.

[21] S. C. Levinson and F. Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers In Psychology*, 6, 2015.

[22] R. Levitan, S. Benus, A. Gravano, and J. Hirschberg. Entrainment and turn-taking in human-human dialogue. In *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction, Stanford, CA*, 2015.

[23] L. Magyari, M. C. Bastiaansen, J. P. de Ruiter, and S. C. Levinson. Early anticipation lies behind the speed of response in conversation. *Journal of Cognitive Neuroscience*, 2014.

[24] L. Magyari and J. P. De Ruiter. Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology*, 3:376, 2012.

[25] W. McKinney. pandas: a foundational python library for data analysis and statistics.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[27] A. Raux. *Flexible turn-taking for spoken dialog systems.* PhD thesis, US National Science Foundation, 2008.

[28] A. Raux and M. Eskenazi. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 629–637, 2009.

[29] A. Raux and M. Eskenazi. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing (TSLP)*, 9(1):1, 2012.

[30] B. S. Reed. Units of interaction: intonation phrases or turn constructional phrases. *Actes/Proceedings from IDP (Interface Discours & Prosodie)*, pages 351–363, 2009.

[31] C. Riest, A. B. Jorschick, and J. P. De Ruiter. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in Psychology*, 6, 2015.

[32] S. G. Roberts, F. Torreira, and S. C. Levinson. The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in Psychology*, 6, 2015.

[33] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *language*, pages 696–735, 1974.

[34] R. Sato, R. Higashinaka, M. Tamoto, M. Nakano, and K. Aikawa. Learning decision trees to determine turntaking by spoken dialogue systems. In *ICSLP*, pages 861–864, 2002.

[35] D. Schlangen. From reaction to prediction: Experiments with computational models of turn-taking. *Proceedings of Interspeech 2006, Panel on Prosody of Dialogue Acts and Turn-Taking*, 2006.

[36] E. O. Selfridge and P. A. Heeman. Importance-driven turn-bidding for spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 177–185, Uppsala Sweden, 2010.

[37] M. Selting. The construction of units in conversational talk. *Language in Society*, 29(04):477–517, 2000.

[38] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1):127–154, 2000.

[39] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

[40] M. Tice and T. Henetz. Turn-boundary projection: Looking ahead. In *The 33rd Annual Meeting of the Cognitive Science Society [CogSci 2011]*, pages 838–843. Cognitive Science Society, 2011.

[41] F. Torreira, S. Bögels, and S. C. Levinson. Breathing for answering: the time course of response planning in conversation. *Frontiers in Psychology*, 6, 2015.

[42] J. Webber. A programmatic introduction to neo4j. In *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, SPLASH '12, pages 217–218, New York, NY, USA, 2012.

[43] M. Wilson and T. P. Wilson. An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, 12(6):957–968, 2005.