6-16-2017

Using Past Speaker Behavior to Better Predict Turn Transitions

Meshorer Tomer

Using Past Speaker Behavior to Better Predict Turn Transitions

Tomer M. Sagie 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

June 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. Sagie Meshorer

has been approved.

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

Steven Bedrick
Assistant Professor

Stephen Wu
Assistant 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 partially funded by the National Science Foundation under grant IIS-1321146.

Contents

Abstract

Using Past Speaker Behavior to Better Predict Turn Transitions

Tomer M. Sagie Meshorer

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

June 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. In addition, speakers with more control of the conversation floor will be less likely to yield the turn. As the conversational image of the speaker evolves as the conversation progresses, 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 seconds and words) relative to the speaker's average turn length.

Relative floor control measures the speaker's control of the conversation floor (in seconds 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 transition prediction.

Part I

Introduction

Chapter 1

Introduction

Conversations are a frequent form of everyday social interaction and are characterized by a rapid exchange of messages between the conversants. The turn exchange system in human conversation is universal in nature and crosses culture, age and language [21]. According to the seminal work by Sack et al. [35], only one conversant is usually speaking at any given time, 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. 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 dominant approaches tried to explain the mechanisms by which the conversation floor is allocated. Duncan [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 Sacks et al. [35], defines the turn allocation process in terms of a set of local rules that are applied 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 adjacency pair[1]). This work investigates whether using features that are computed over all past turns can help improve the predictability of turn transitions.

Spoken dialogue systems (SDS) are computer systems that support a conversational 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 system of turn exchanges.

Early SDS systems did not contain a turn management component and instead used a fixed

[1]Adjacency pairs are a pair of utterances which have the following attributes [37]: (1) adjacent positioning of component utterances, (2) different speakers producing each utterance, (3) the first utterance relates in a specific way to the second utterance thus forming a pair type, for example, Question-Answer.

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 on syntactic [35, 5], prosodic [10, 7, 8], and pragmatic cues [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 conversants 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 a speaker's past 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 predict 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, 13] to train random forest models [28]. First we performed data preprocessing in order to merge turn, dialog act and single word information into a single data set. Next, we trained baseline models using 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. To measure the predictive power of the different models we use standard model metrics: F1, precision, recall and area under the curve (AUC)

1.3 Contribution

Our results show that the summary features improve prediction performance in both area under the curve (AUC), 69.4% vs 66%, and F1, 69.3% vs 57.8%. In addition, the model that was trained on all of the features performs better than the local features model in both AUC, 83.7% vs 81.1%, and F1, 77.5% vs 74.8%. The results show that using summary features helps predict turn transitions.

1.4 Dissertation Structure

The rest of this thesis is as follows. Chapter 2 presents the related work in both human-human and human-computer conversations. Chapter 3 describes the theoretical model that underlies the summary features. Chapter 4 describes the data sources, data preparation and initial data analysis. Chapter 5 describes our main study in which we compute the long term features over all the turns in each conversations. The chapter compares the baseline models as well as the model that includes the summary features. Finally, in chapter 6 we present the conclusion and future work.

This thesis is based on a paper [27] that was submitted and presented at Interspeech 2016.

Chapter 2

Related Work

This chapter presents work related to turn transition prediction in both human-human conversations and human-computer conversations. Studying human-human turn allocation mechanism might inform us on the inner workings of the system that people use during conversations when regulating turn allocation. This in turn will help guide us in how to implement such mechanisms in spoken dialogue systems.

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 [19] included gaze as an additional signal for turn transitions. As the signals are local in nature (derived from the last turn), our summary features complement them.

Turn allocation was introduced in the seminal work by Sacks, Schegloff, and Jefferson [35], 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 recognize 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 conversant (with a high relative floor control score) might self select thinking that the other conversant will not take turn.

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 [35], prosodic [10] and pragmatic [10, 9]). Our work augments the local utterance features with summary features that can be used to improve projectablity.

The work of Selfridge and Heeman [39] on turn bidding suggests that each conversant measures the importance of their potential contribution when negotiating the right to the conversation floor. Their work is based on the fact that 1) the importance of speaking is the primary motivation behind turn-taking behavior, and (2) conversants use turn cue strength to bid for the turn based on this importance. Our work suggests that contestants might use summary features to signal the strength of the bid. For example conversants will take turn longer than their average if they want to signal turn importance.

Wilson et al. [46] explored the role of entrainment in turn taking. They proposed there is entrainment of endogenous oscillators in the brains of the speaker and the listener on the basis of the speaker's syllabus production. In their model, the speaker and the listener are counter-phased so 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 [12]. 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 reduce speech recognition error rates, the problem of turn taking in SDS rises in importance [18]. 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 latest utterance. As different studies use a variety of features, we will focus on those that are similar to our summary features.

Gravano and Hirschberg [15] 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 milliseconds 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 also show in Section 5, the authors found that combining multiple cues leads to better accuracy.

Raux and Eskenazi [31] performed a comprehensive study of features that might affect 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.

In a more recent study, Nishitha and Rodney [16] used a model based on N grams of dialog acts to predict turn transitions. They trained decision tree models using the switchboard data. As features, they contrasted the use of the previous dialog act, the previous two dialog acts and the previous three dialog acts. They also contrasted whether to indicate if there is a speaker change between the previous dialog acts. They achieved an F1 measure of 0.67 for the model that use the previous two dialog acts. In our work, we base our baseline models on the previous dialog act and the previous two 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

Theoretical model

This chapter aims to provide a more formal description of our turn taking model. First we present a formal definition of a conversation, turn changes and local features. Next we present the formulas to compute the summary features: relative turn length and relative floor control.

3.1 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:

$$\dots s_{i-2}, d_{i-2}, s_{i-1}, d_{i-1}, s_i, d_i \dots \tag{3.1}$$

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:

$$\dots d_{i-2}, y_{i-1}, d_{i-1}, y_i, d_i, y_{i+1} \dots \tag{3.2}$$

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.

3.2 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 introduce some notation. Let 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 starting from the beginning of the first dialog act in the turn, and up to the end of the last dialog act. We measure length 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

$$avg\_t_i = \frac{\sum_{t \in S_i} length(t)}{|S_i|} \tag{3.3}$$

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).

$$rt_i = \frac{length(t_i)}{avg\_t_i} \tag{3.4}$$

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 use S_i as defined above: the set of complete turns of speaker s_i that are prior to the turn that d_i is in. We similarly define L_i to be the turns of the other conversant (the listener of d_i) that are prior to d_i. We first compute the conversation length up to d_i denoted as c_i, which excludes inter-turn pauses.

$$c_i = \sum_{t \in S_i \cup L_i} length(t) \tag{3.5}$$

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:

$$rc_i = \frac{\sum_{t \in S_i} length(t)}{c_i} \tag{3.6}$$

Note that we also 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 Chapter 5.

Chapter 4

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 [14].

The switchboard corpus is the first large collection of phone conversations and was collected in 1990-1991. The initial goal of the corpus was to facilitate speech research by providing prerecorded 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 original audio recordings were later annotated 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 types of annotations for each conversation. We used the turn annotations to assign turns to speakers and mark the turn start and end points. We used the dialog act annotations 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 used the token annotations to assigned words to dialog acts and to compute the overall number of words within each dialog act.

4.1 Data Preparation

Figure 4.1 shows our data pipeline. Data is imported from the NXT switchboard corpus [4] into a graph database [45]. Figure 4.2 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.

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

- We exclude 11 dialogue acts that were coded in Switchboard as “other.”

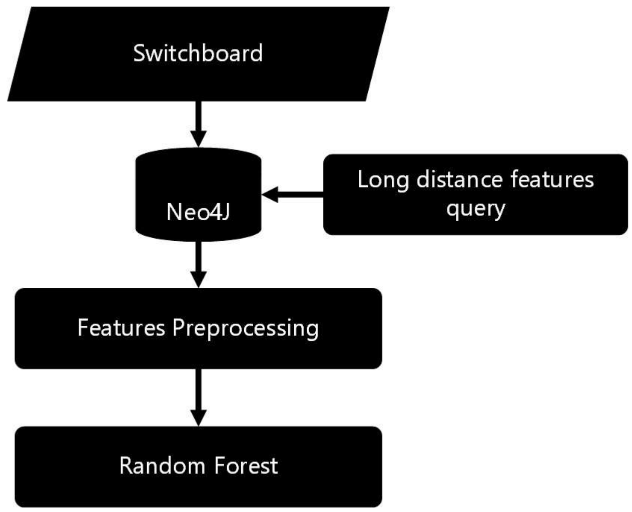

Figure 4.1: Experiment data pipeline

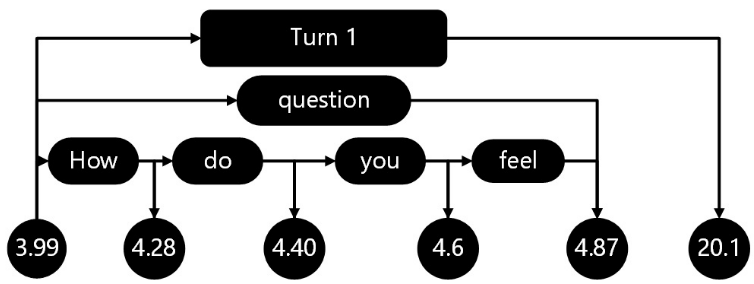

Figure 4.2: Conversation graph data model

- We filtered out all the dialogs that had data integrity issues, for example dialog acts that referred to non existent terminals. This decrease the number of conversations from 624 to 310 conversations consisting of 60595 dialog acts.

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

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

Table 4.1: Mapping from dialog act to dialog act class

4.2 Data Exploration

In this section we explore the distribution of the independent data variables and their relationships with turn taking (which is the dependent variable) We visualized the data using the python seaborn library, which is based on matplotlib. The data set for the visualization contains a row for each dialog act. For each dialog act we measure its length, and the values of the summary features (Relative turn length and Relative turn control). The independent variable denotes whether a turn change occurred after the dialog act. The variables are summarized in Table 4.2.

Variable	Description	Type
Previous Dialog Act	the dialog act before the current one	categorical
Dialog Act	the current dialog act	categorical
Length	length of the current dialog act in seconds	seconds
Relative Turn Length (RTL)	Relative turn length as defined in Section 3.2	percent
Relative Time Control (RTC)	Relative time control as defined in Section 3.2	percent
Turn Change	1 if there was a turn change after this dialog act	binary

Table 4.2: Data Fields

4.2.1 Influence of Summary Features on Turn-Taking

In this section we explore how the summary features, defined in Section 3.2, affect turn-taking. To measure the influence of relative turn length on turn taking, we first remove all the dialog acts that occurred in the first 120s of each dialog, as their relative turn lengths are not based on as much dialog history as the later ones, and so will not be as reliable. This leaves us with 42,721 dialog acts. We then divide the range of relative turn values into buckets and counted the number of dialog acts (regardless of their type) that led to a turn change and the total number of dialog acts for each bucket. We compute the probability of a turn change by dividing the former by the latter. As can be seen in Figure 4.3, the chance of turn change is higher when the speaker has the floor for smaller values of relative turn length. The really small relative turn lengths (0-15%) are probably back channels. We also observe that as the speaker has the floor for more time, the speaker tends to hold it.

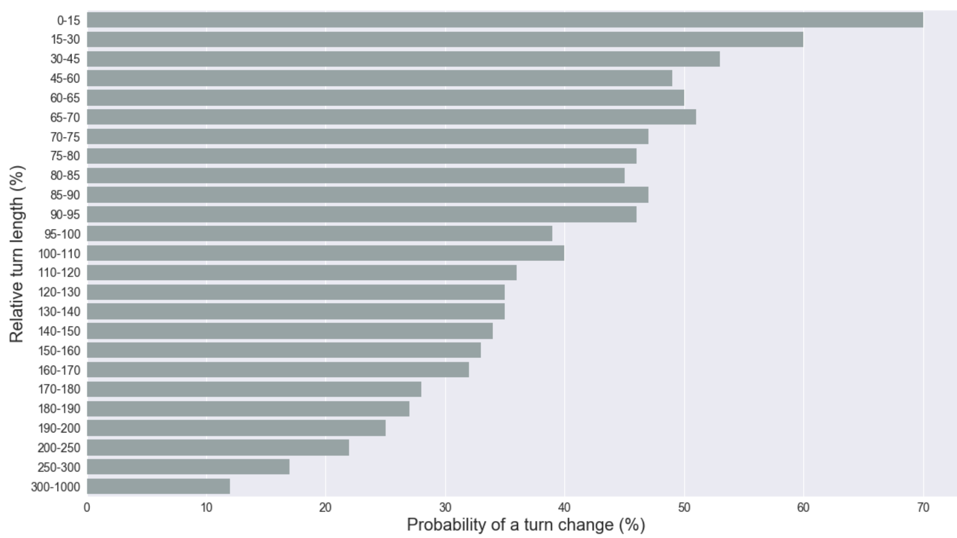

Figure 4.3: Relative turn length effect on probability of a turn change

The above finding is actually is opposite of our hypothesis that we stated in the introduction, in which we assumed that if a speaker's turn so far is less than her/her average then he/she would tend to keep speaking, and vice verse, if the speaker passed his/her average turn length he/she would tend to give up the turn. However, based on 4.3, we see the exact opposite. We will explore this difference further in Section 4.2.4. None-the-less, we do see that this feature carries information, which can be used by machine learning, even though it is opposite of what we expected.

Next we measure the relation between values of relative floor control and probability of turn change. As with relative turn length, we only include dialog acts after 120s into the conversation. We also divide the range of relative floor control into buckets such that each bucket contains roughly the same number of dialogs acts. We then count the number of dialog acts (regardless of the type) that led to a turn change and the total dialog acts for each bucket. Last, we compute the probability of a turn change by dividing the former by the latter. The result is shown in Figure 4.4. As we hypothesized in the introduction, we observe that high values of floor control correlate with the willingness of the current speaker to give up the turn. When the speaker has relative floor control scores above 50% the speaker tends to keep the floor and avoid a turn change.

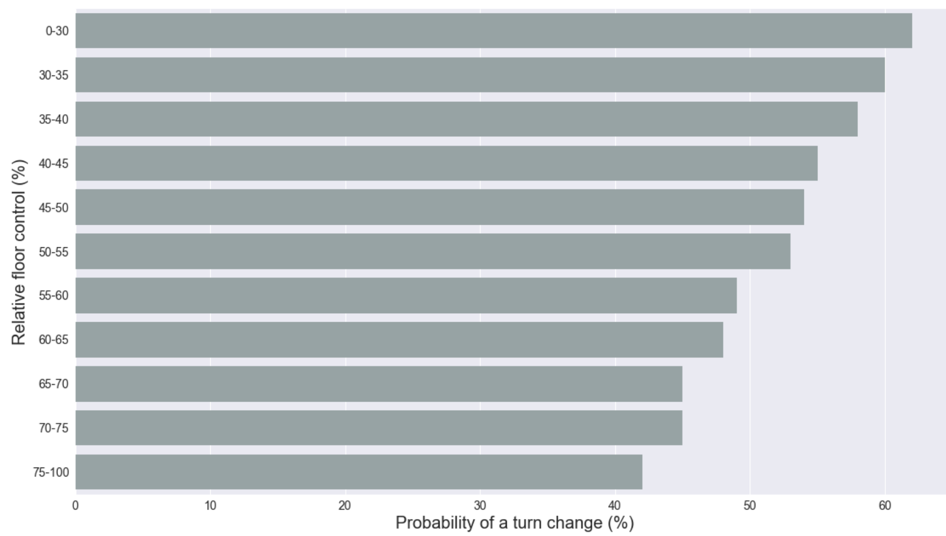

Figure 4.4: Relative floor control probability of turn change

4.2.2 Influence of Dialogue Acts Types

Next we explore the influence of dialog acts types (which are the local features in the theoretical model) on turn changes. First we compute the distribution of dialog acts type over all the dialog acts in corpus. Figure 4.5 plots the distribution of each dialog act in the corpus. Each bar is a count of the dialog act type, divided by the total number of dialog acts. We observe that the majority of dialog acts are statements, backchannels and opinions. This might be due to the nature of the switchboard corpus, which consists mainly of casual conversations.

Next we measure the correlation between dialog act type and the probability of a turn change. In Figure 4.6, each bar measures the number of times a turn change occurred after a dialog act type divided by the total number of dialog acts of this type: the probability that a dialog act of the said type will lead to a turn change. We observe that the majority of back channels and questions (which are usually the first utterance in an question-answer adjacency pair) will lead to a turn change.

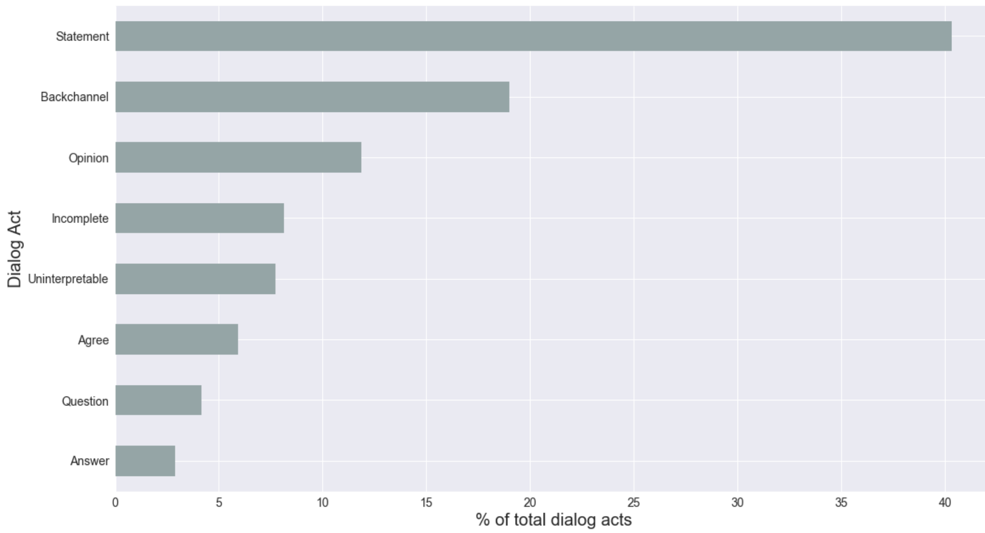

Figure 4.5: Dialog act relative count

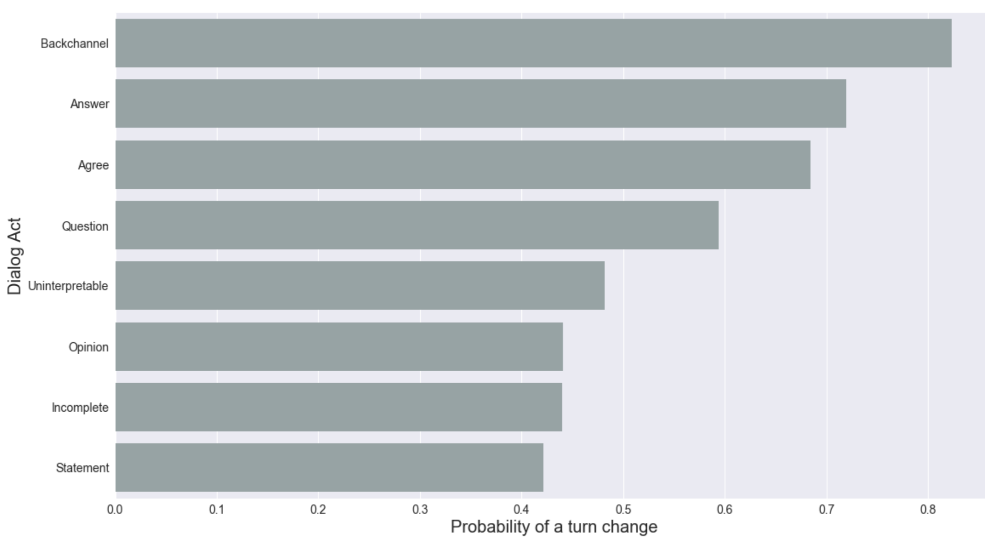

Figure 4.6: Dialog act probability of turn change

4.2.3 Dialogue Act Types and Summary Features

In this section we explore how the summary features are distributed for each dialog act. Moreover, we look at the score distribution for dialog acts that are followed by a turn change and ones that do not. In Figure 4.7, we see that, for most dialog act types, the median relative turn length that led to a turn change is smaller than when it does not. This suggests that speakers who are using the floor more than their average turn length will tend to hold the floor even longer. This is consistent with our finding from Section that longer relative turn lengths result in a lower rate of turn-changes, but shows that it holds across speech act types as well.

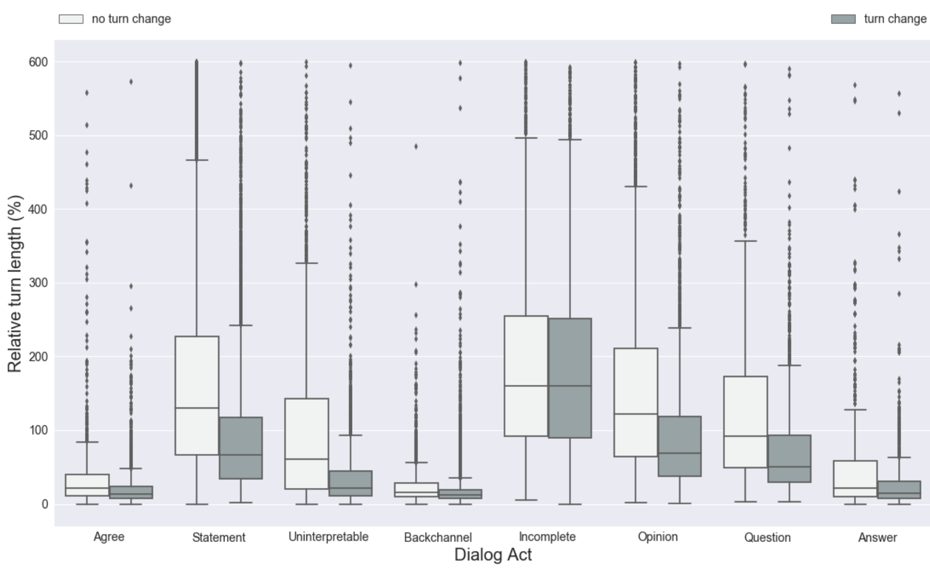

Figure 4.7: Relative turn length for dialog act type

Next we measure the relative floor control score for each type of dialog act. In Figure 4.8, for each dialog act we show the distribution of relative floor control for acts that lead to turn changes and acts that do not. We see that for most of the dialog acts, the median score is about equal and close to 50%. We also observe that 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. Again, this is consistent with our earlier result that higher values for relative floor control tend to lead to less turn-changes, but here we see that it also applies to each dialog act.

4.2.4 Explanation of the correlation between Relative turn length and turn changes

Our initial hypothesis 1 was that low relative turn change would lead to a speaker holding the floor, and high relative turn change would lead to a turn change. However, in section 4.2.3, we reported that the relative turn length feature was giving us the opposite effect. We found that the

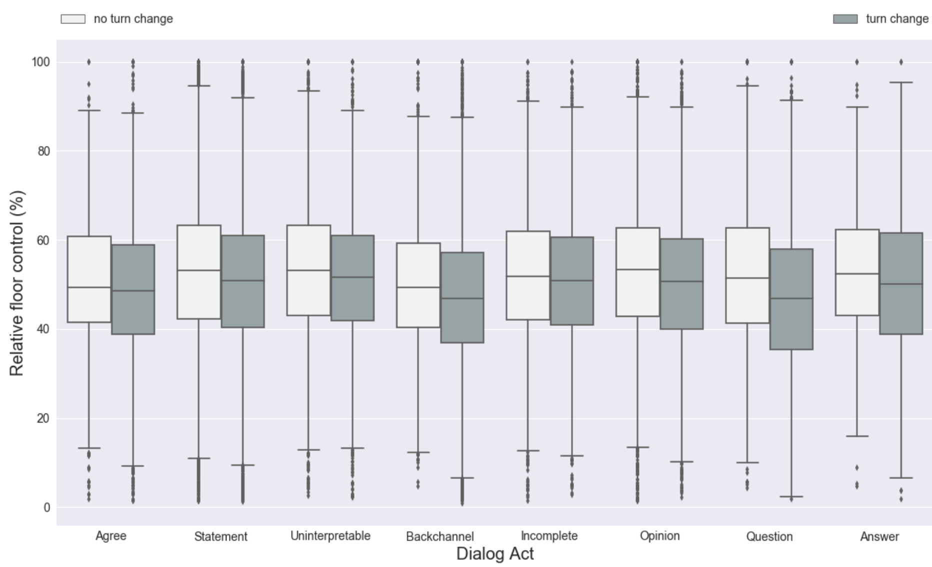

Figure 4.8: Relative floor control by dialog act

chance of turn change is higher when the speaker has the floor for shorter than its average turn. Also, the speaker will likely keep the floor when speaking more than the average turn length.

We suspect that the reason that our initial hypothesis proved to be incorrect is due to the nature of the corpus. Figure 4.9 shows the actual distribution of turn lengths (in seconds).

First we explain why dialog acts with small relative turn length will more likely lead to a turn change. We suspect that the set of short turns is composed of turns with a single dialog act. Moreover, this dialog act is likely to be back channel or an answer, both of which have low relative turn length, see figure 4.7. If we look at figure 4.6, we can see that the chance of a back channel or a answer causing a turn change is 60-80%. From both observation we can conclude that a short dialog act with low relative turn length will likely to lead to a turn change.

On the other end, we want to explain why high relative turn length would not lead to a turn change. This can be attributed to the structure of the distribution curve of figure 4.9, which has a long flat tail. Hence if the speaker spoke for high relative turn length, 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.

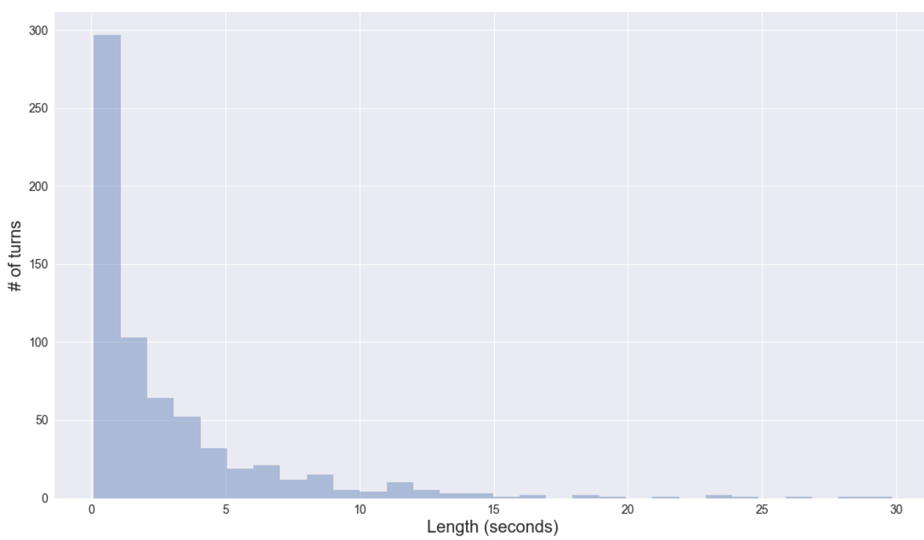

Figure 4.9: Distribution of the turn length

Chapter 5

Study

5.1 Classification Models

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

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.

For the first classifier, 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.

For the second classifier we used gradient boosting [11]. The gradient boosting classifier uses a combination of a loss function and weak classifiers to create an ensemble of weak classifiers, which perform a majority vote. At each step of the classifier boosting, the weight of hard to classify examples is boosted, such that new weak classifiers are trained on the hard to classify data.

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.

5.2 Metrics

To evaluate our hypothesis, we use the trained models to predict the turn taking behavior in our data. We than compare the predictions against the truth values. The results are shown in the confusion matrix shown in Table 5.1. Each row in the matrix represent the true class and each column represent the predicted class. The cells in the matrix are computed as follows:

		Predicted to cause turn change	
		yes	**no**
Caused turn change	**yes**	True Positive	False Negative
	no	False Positive	True Negative

Table 5.1: Confusion Matrix

1. True Positive: Count of the dialog acts that were predicted to cause a turn change and that did lead to turn change.

2. False Positive: Count of the dialog acts that were predicted to cause a turn change but in fact did not cause one.

3. True Negative: Count of dialog acts that were predicted to not cause a turn change and did not cause turn change.

4. False Negative: Count of dialog acts that were predicted to not cause a turn change but in fact caused one.

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

Accuracy measures how many dialog acts are classified correctly out of the total dialog acts.

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

Precision measures how accurate the classifier is for the dialog acts that caused turn change; i.e., how many dialog acts that were classified as causing a turn change, did in fact led to a turn change.

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

Recall measures how many dialog acts that predicted turn change were detected by the classifier.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5.4}$$

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: if we want to increase recall we will reduce precision. To measure this tradeoff, we will use ROC (Receiver Operating Characteristic) curves. The ROC curve measures the true positive rate or recall, against the false positive rate (FPR). When comparing different ROC curves, we measure the area under the curve (AUC).

5.3 Results

We first analyze 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 5.2 shows the results of training a random forest of 200 trees for each model using 10 folds cross validation. We see that using the summary features provides better accuracy than baseline 1, which use only the current dialog act (65.54% vs 62.79%). In addition, using the full model yields an absolute improvement of over 1.08% in the accuracy. In addition, baseline 1 has high precision, but very low recall. Using only the summary model improves recall and decreases precision, but 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	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 5.2: Precision, recall and F1 results using Random Forests

The effect can also be seen in phwas: figureFigure 5.1, which shows the ROC curves and the AUC for each model. We notice that the AUC of the summary model is better than baseline 1 model (0.69 vs 0.66), and when adding the summary features to the local features (the full model), we see the AUC improves (0.84 vs 0.81). 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 baselines.

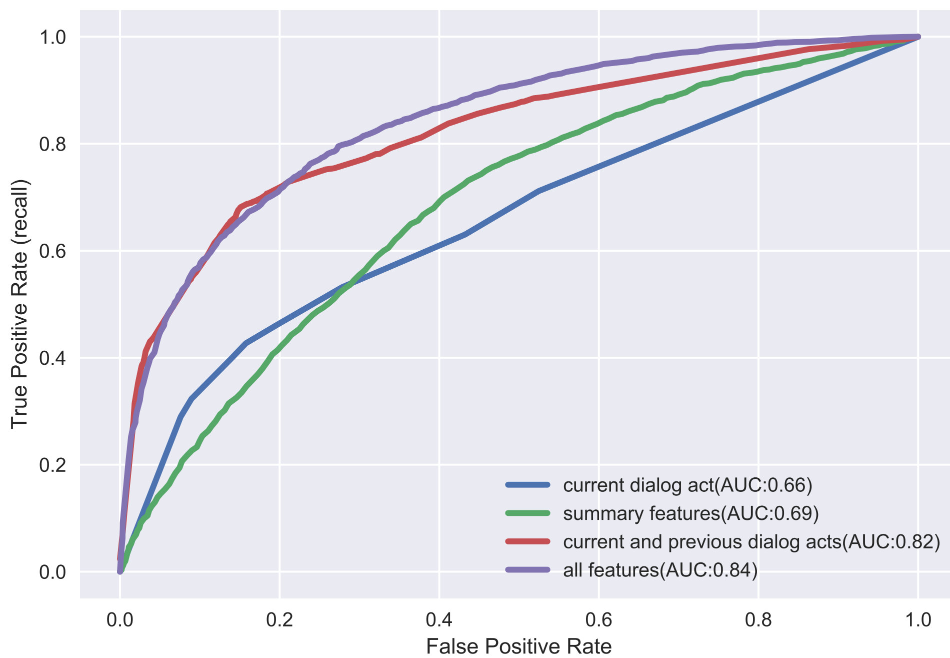

Figure 5.1: ROC curves and AUC of the different models

Table 5.3 shows the results of training gradient boosting classifer for each model using 10 folds cross validation. In general, the random boosting classifier perform better than random forest. Moreover, we can observe that the summary model is better than the beseline mode and the full model is better than the baseline.

	accuracy	f1	precision	recall	roc_auc
summary	67.91%	71.30%	69.20%	73.55%	72.64%
baseline 1	62.79%	57.81%	74.98%	47.04%	65.99%
baseline 2	74.88%	74.82%	81.92%	68.86%	81.10%
all	76.57%	78.74%	77.44%	80.11%	84.84%

Table 5.3: Precision, recall and F1 results using Gradient boost classifier

5.3.1 Sensitivity to Mesurement Start Time

In Meshorer and Heeman [27], we assumed that it take 120 seconds for the conversational image to form. So we only test our results on dialogue acts that happen more than 120 seconds into each conversation. We now evaluate this assumption by building models based on different starting times, ranging from 0s to 180s. For the smaller start times, we also have more data, as we start analyzing the performance of our models earlier in the dialogues. This is why the results for baseline 1 and 2 also differ with the start times.

The results are shown in Table 5.4 in terms of the AUC scores. As expected, we do not see much difference in the AUC values for baseline 1 and 2, so there is not much effect due to the increase in the amount of data that we are able to use. We also do not see much difference in AUC values for the summary and full model. This shows that the summary features predictive strength is not affected by the 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 5.4: AUC Score in relation to the start of the dialog

Chapter 6

Conclusions

This thesis explored 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 included 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.

6.1 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, 35, 10, 5, 25, 2], prosodic [6, 10, 41, 8, 5, 32, 31, 17, 2], pragmatic [10, 12, 31], semantic [31] and non-verbal [19]. 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 system, 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 average as computed over all the previous turns. In future work, it would be interesting to use a simple moving average (measured over multiple window widths) as well as an exponential moving average.

Our work treated back-channels as counting toward turn-taking events: if conversant A makes a statement, then conversant B back channel, and then A makes another statement, we count B's back channel as a full turn. However, there is not a clear consensus in the research community that back channels should be treated as independent turns. Our treatment of back channels as turns is partially responsible for the large number of short turns that we observed in Figure 4.9. This probably contributed to our finding that smaller values for relative turn length result in a greater chance for a turn change, which is opposite from what we expected. Future work should investigate the correlation between turn length and turn changes when back channels are not counted as independent turns.

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] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

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

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

[14] 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.

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

[16] 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.

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

[18] 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.

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

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

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

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

[23] 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.

[24] 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.

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

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

[27] T. Meshorer and P. A. Heeman. Using past speaker behavior to better predict turn transitions. In *INTERSPEECH*, 2016.

[28] 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.

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

[30] 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.

[31] 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.

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

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

[34] 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.

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

[36] 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.

[37] E. A. Schegloff and H. Sacks. Opening up closings. *Semiotica*, 8(4), 1973.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

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