

Modelling Turn-Taking in Human Conversations

Nishitha Guntakandla and Rodney D. Nielsen

Hilt Lab, University of North Texas
Denton, Texas 76201

Abstract

In this work, we make a contribution to developing turn-taking mechanism in spoken dialogue systems. We focus on modelling the turn-taking behavior in human-human conversations. The proposed models are tested on the Switchboard corpus which contains conversations annotated at the utterance level. Several experiments were performed to analyze the salience of different features that are associated with the preceding utterances for the task of predicting whether there will be a change in speaker. The impact of the n-gram sequential modelling on turn-taking is studied. Machine learning techniques are also employed to perform this prediction task.

Results from the experiments suggest that a combination of the preceding dialogue sequence, previous changes in speaker information and duplicating the sequences by replacing speaker IDs plays an important role in modelling turn-taking. Utterance sequences of length 3 in N-grams resulted in higher predictability for this task. Experiments suggest that a machine learning technique with 4-grams of a combination of all these features is effective for predicting speaker changes.

Introduction

One of the practical goals of computer scientists is to implement human intelligence in computers and there by building a human-like robot which acts like a companion. For any companion robot, it is important to understand what the other speaker is trying to communicate and to respond appropriately, but it is also important to understand when to respond. Turn-taking is one of the cognitive behaviors that humans developed to make an organized conversation, where speakers in the conversation decide who is to speak next to maintain the flow of the conversation. For example, the current speaker can offer the floor to another speaker or continue.

Our aim is to model the turn-taking mechanism in human conversations based on the intentions exhibited by a speaker through each utterance. (Jurafsky, Shriberg, and Biscia 1997a) Intentions are recognized by syntactic realization and the semantic information of an utterance. This information is usually represented by labels called Dialog Acts

(DA) (which will be discussed in section 3). For example, in most cases, if the speaker's intention in the utterance is to ask a question, he will offer the floor to another speaker to answer it. Here, offering the floor indicates there will be a change in speaker

The work in this paper reports the modelling of the turn-taking mechanism as a predictive task, i.e., after this utterance will there be a change in speaker or not?. We discuss the significant features (like dialog acts and previous turn changes) that are extracted from a series of utterances to model this behavior. We also report the experiments conducted with language modelling and machine learning techniques with combinations of different features. We show that all the models and machine learning techniques with any combination that are discussed in the paper outperforms the baseline.

The rest of the paper is organized as follows. In section 2, we briefly review some of the previous work on turn-taking mechanisms. The human-human dialogue corpora, the training and the test split that we used for this task is described in section 3. In section 4, the n-gram language model and machine learning techniques are discussed along with the features that are used for this specific task. We report the results from these experiments in section 5. Finally we conclude with a discussion.

Related Work

The process of turn taking in human-human interaction has attracted attention from researchers in psychology, linguistics and engineering. Previous studies use signals, cues and rules from the conversations to model turn-taking behavior. There has been a lot of work on finding turn ending cues in conversations. Some of the cues in a conversation that indicate the end of a turn are pause duration (Schlangen 2006), speaking rate, energy and pitch levels (Ward, Fuentes, and Vega 2010), lexical cues, textual completion, IPU duration and acoustic cues, and intonation (Gravano and Hirschberg 2011) along with some other cues which are only applicable to face-to-face conversations like gaze, posture and posture shifts (Padilha and Carletta 2003), head gestures and movements, facial expressions, foot movements, shoulder movements, hand gestures and movements. Although, there has been significant amount of work on speech and visual cues, we will not discuss them in detail because we are only look-

Table 1: Sample Switchboard Conversation with utterance-level annotations

Tag	Speaker id & Turn no.	Utterance no. within the turn	Utterance
o	A.1	utt1	Okay. /
qo	A.1	utt2	{F Uh, } first, {F um, } I need to know, {F uh, } how do you feel [about, + {F uh, } about] sending, {F uh, } an elderly, {F uh, } family member to a nursing home? /
sv	B.2	utt1	{D Now, } how long does it take for your contribution to vest? /

ing at textual cues in conversations.

However, (Sacks, Schegloff, and Jefferson 1974, Schlangen 2006) deals with speech and they consider the turn-taking modelling to be predictive like we do. Decisions on when to take or when not to take the turn in a conversation should be made prior to reaching a transition place. (Sacks, Schegloff, and Jefferson 1974) considers the turn-taking in conversations between two or more persons is based on basic set of rules governing turn allocation at every Transition-Relevance place (TRP). TRP’s usually occur at possible syntactic completion points and are conveyed by the intonation of the speaker. Whereas, (Schlangen 2006) reports word-final pitch, intensity levels and n-gram based features are the most predictive ones.

The corpus that we adopted is the same used by (Schlangen 2006) which will be discussed in the next section.

Data

In this work, we use the Switchboard corpus (Godfrey, Holliman, and McDaniel 1992), which consists of 2438 unconstrained human-human telephone conversations, averaging 6 minutes in duration and on different topics. These conversations are carried out among 543 speakers, where no pair of speakers had a conversation together more than once and no one spoke on a given topic more than once.

The Switchboard dialogue act corpus (SwDA) is a subset of the original Switchboard corpus which contains only 1155 conversations and is annotated at the utterance level with dialog acts (DA). Dialogue acts provide information about the lexical and syntactic realization of an utterance. The “SWBD-DAMSL” annotation scheme is adapted to label each utterance, and defines approximately 60 unique tags. But the utterances, at times, displayed characteristics of different tags together, which resulted in annotating them with a combination of all the appropriate tags. Although, there are 220 of these combined tags, their occurrence is infrequent across the corpus. Thus, combined tags were clustered into 42 classes based on (Stolcke et al. 2000). While clustering, (Stolcke et al. 2000) removed any utterance with the “@” tag, which indicates that the transcription is incorrect or has bad segmentation, because evaluation on in-

correctly transcribed utterances for dialogue act classification task is inappropriate. Although, there are transcription and segmentation errors with the context of the utterance, it provides information on who is the current speaker. So, for this task we also consider the utterances that are labeled @. Hence, our data has 43 class labels. The sample data of Switchboard conversations with utterance-level annotations are shown in Table 1.

We also divided the corpus into two speaker disjoint clusters, and as a result, 761 conversations are in the training set and the test set consists of 146 conversations. The remaining 248 of the 1155 total dialogues were removed from the data (i.e., were neither included with the training data nor the test data) because these conversations were between one speaker from the first cluster and the other from the second cluster. This experimental setup is to evaluate our system performance when we encounter new speakers.

Experimental Design

Similar to the work in (Schlangen 2006) we try to predict whether there will be a change in speaker or not after each utterance. For most prediction tasks, probabilistic language models are employed for better performance. Thus, we use an n-gram language model for modelling turn-taking behavior in this paper. We will also explore, the appropriate length of the dialogue history that maximizes the model accuracy.

N-Grams

Language models assign probabilities to sequences of tokens. The N-gram language model predicts the next token considering a sequence of preceding n-1 tokens, based on probability values assigned. In our task, tokens would be utterances and the information associated with them like speaker ID, dialogue act and change in speaker information which will be discussed in detail. Hence, sequences of a combination of the features from utterances are used to model the turn-taking mechanism, using the statistical properties of N-grams.

Features

- **Utterance:** An utterance is a transcribed segment of speech. It contains an uninterrupted sequence of words.

Table 2: Results of N-gram models with different combinations of features

Feature combinations	Bigrams (%)	Trigrams (%)	4-grams (%)
Dialogue acts	62.63	63.14	62.93
Dialogue acts + change in speaker	59.35	63.27	63.07
Dialogue acts + change in speaker + speaker ID	58.78	60.25	60.67
Dialogue acts + change in speaker + speaker ID + duplicating sequences by replacing speaker IDs	62.63	63.33	63.15

Occasionally, recognized words may contain errors because of the background noise or poor speech quality.

- **Dialogue acts:** Dialog acts are the labels given to utterances based on the intention of the current speaker. This feature has 43 values. Dialog acts carry semantic and syntactic information of utterances, so instead of using utterance lexical tokens as features to model, we only use dialog acts
- **Speaker information:** Switchboard corpus contains dyadic conversations. Speaker information in these dialogues are provided as speaker IDs either A or B. A is the speaker who initiates the conversation and the other participant will be B. Speaker IDs are associated with each utterance indicating the speaker of that particular utterance.
- **Change in the speaker information:** This is a binary feature (1/0) specifying whether there is a change in speaker or not. This is also the class feature to be learnt.

Experimental Setup

Our training set contains 324 unique speakers. We used leave-one-out cross-validation, where a given test fold represents all of the utterances of one of the 324 speakers and the associated training data is derived from all of the utterances that did not involve any of the test speaker’s dialogue partners. So that our experimental set up is equal to our original training and test split.

Experiments

It is cumbersome to compute the probabilities of an entire dialogue history and model them. Also with the longer sequences, the performance of the model drops, because the occurrence of the exact long sequence will be rare. Thus, we try to limit the length of the history for modelling. To decide on what is the appropriate sequence length for this task, we have conducted experiments using bigram, trigram and 4-grams models.

The probabilities are computed on the features extracted from the immediate previous 1, 2 and 3 utterances of the current utterance to build bigram, trigram and 4-gram models

respectively. These probabilities are used to predict whether there will be a change in speaker or not.

We conducted experiments with different combinations of features to decide on one significant feature set for this prediction task. We computed accuracies of each model based on its ability to predict that there will be a change in speaker or not after this utterance. Experimental results of all N-gram models with different combinations of features are shown in Table 2.

Majority baseline for this task is considering there will be a change in speaker after every utterance. The accuracy computed from the training set is 55.81%

- **Dialog acts:** Only the previous dialogue acts are used to predict a change in speaker. The performance of all the models with just dialog acts as features outperforms the majority baseline class by at least 7%. This alone shows that dialogue act is an important feature to predict the turn-taking behavior.
- **Dialog acts and change in speaker:** Here, sequences of a combination of previous dialogue acts and change in speakers from previous utterances are used. For bigram model this feature set hampers the performance, but for trigram and 4-gram models there is slight increase in accuracy.
- **Dialog acts, change in speaker and speaker ID:** Speaker IDs which represent the current speaker are added to the previous feature set and sequences of those combinations are used to model turn-taking. Using this feature decreases performance for all the language models, because by adding the speaker IDs to the feature set, we are just adding the sequences that are particular to the current speaker, thus the feature lacks generality. We could overcome this by duplicating all the sequences of current speaker to other speaker by replacing their IDs (i.e., A to B and B to A). Therefore, all the sequence examples will be available to both the speakers.
- **Dialog acts, change in speaker, speaker ID and duplicating the sequences by replacing speaker IDs:** A combination of all the features and duplicating the sequence by replacing A’s with B’s and B’s with A’s as feature set

Table 3: Results of machine learning techniques in comparison with n-grams

Technique	Bigrams (%)	Trigrams (%)	4-grams (%)
N-grams	62.63	63.33	63.15
Naive Bayes	62.63	62.74	62.64
J48	62.63	63.49	63.67
Bayes Net-work	62.63	62.74	62.65

provides better results to any of the previous feature sets in all the models.

Overall, the experiments show that the trigram model with the combination of dialog act, change in speaker, speaker ID and duplicating the sequence by replacing speaker IDs performed better than any of the combinations with any model.

Machine learning techniques

We employed different probabilistic machine learning techniques to predict whether there will be a change in speaker or not by considering features from previous utterances. Experiments were conducted with the same feature combinations from the same number of previous utterances that are used in N-grams to build feature vectors for these machine learning techniques. The experimental setup is also similar to the one utilized in N-grams i.e., leave-one-speaker-out cross-validation. We accounted Naive Bayes, Bayes Network and J48 classifiers with default parameters from WEKA (Hall et al. 2009).

Predicting is similar to making decisions when some conditions were provided. Thus, decision tree classifier performed better in this task. As discussed in the N-grams model, a combination of dialogue acts, change in speaker information, speaker ID and duplicating the sequence by replacing speaker IDs stands as a very good feature set.

Accuracies obtained from leave-one-out cross-validation by N-grams, Naive Bayes, Bayes Network and J48 classifier are shown in the table 3.

Experimental results show that 4-grams of combination feature set in J48 classifier performs slightly better than the trigrams language model but, there is very little difference between accuracies of the Trigram and J48 models, so we consider both for our final testing.

Results

Finally, the experiments from the above section, found that trigram model and the J48 decision tree classifier with the combination of all the features performed best on predicting whether there will be a change in speaker or not. These models are tested on the test split, which contains 146 conversations and the results of N-grams and J48 are reported in table 4.

Table 4: Majority baseline and the results of the proposed models on the test set.

Technique	Accuracy (%)
Majority class baseline	54.27
Trigram Model	61.74
J48 with 4-grams	62.70

The majority baseline that speaker changes after every utterance is 54.27%. Results depicts that the models that we proposed with combination of dialogue acts, change in speaker, speaker IDs and duplicating the sequence by replacing speaker IDs as features outperformed the majority baseline.

Switchboard corpus contains backchannels and is one of the most frequent classes among 43 classes we considered. One cannot imagine natural conversation without backchannel. Role of the backchannel in a conversation is to specify that the listener is interested in the conversation and they also act as fillers, when the current speaker pauses his speech for short duration.

Ex: Backchannels are highlighted in the sample conversation -

sd A.5 utt2: we're not being tested for drugs at all, {F uh,} /
sd A.5 utt3: our policies and procedures manual, {F uh,} the furthest it goes about drugs is in [the, +kind of the miscellaneous section, or - -
b B.6 utt1: Uh-huh. /
+ A.7 utt1: -it's reasons for immediate dismissal, /
sdq A.7 utt2: it says, use of narcotics on company premises. /
b B.8 utt1: {F Um. } /
sv A.9 utt1: {C So } that's pretty general, /

Introducing backchannels in a conversation involves quick change in speakers. There are varying opinions among researchers about considering the speaker change while backchannel as a turn.

All the experiments reported earlier in this work considered speaker change during backchannel as a turn. Some of the previous work (Schlagen 2006, Koiso et al. 1998) does not consider the backchannels as instantiating turn-transitions. As mentioned earlier, (Schlagen 2006) worked with the Switchboard corpus to predict turn-taking after each utterance. (Schlagen 2006) trains a set of machine learning classifiers by extracting features like f0 curve, intensity, pausal durations from speech and claims that Bayesian Network among all the machine learning techniques performed best. This work considers, take (a different speaker takes the floor) and wait (same speaker continues) as class labels.

(Schlagen 2006) evaluated the proposed model on Section 2 dialogues of switchboard which contains 20 dialogues.

Table 5: F1-Measure for change in speaker and no change in speaker classes when backchannels are not considered as turn initiators.

Technique	Change in speaker/take (F1-Measure)	No change in speaker/wait (F1-Measure)
Bayes Network (Schlangen 2006)	0.46	0.74
Trigrams	0.48	0.80
J48 with 4-grams	0.51	0.79

Table 6: F1-Measure for change in speaker and no change in speaker classes when backchannels are considered as turn initiators.

Technique	Change in speaker/take (F1-Measure)	No change in speaker/wait (F1-Measure)
Trigrams	0.67	0.55
J48 with 4-grams	0.67	0.57

Whereas, our work is tested on 146 dialogues which has no speaker in common with the training set. To compare our results with (Schlangen 2006), we did not consider backchannels as turn initiators. F1-Measure for each class is computed on the test set and the corresponding results are shown in the table 5.

The trigram language model and J48 with 4-grams modeled on sequences of a combination of dialogue acts, change in speaker, speaker IDs and duplicating the sequence by replacing speaker IDs that we proposed slightly outperforms the Bayes Network machine learning that is proposed with audio features by (Schlangen 2006). The area under the curve (AUC) value for J48 with 4-gram feature combinations when backchannels are not considered as turn initiators is 0.686 for both the classes.

F1-Measure for change in speaker class and no change in speaker class when backchannels are considered as turn initiators are reported in the table 6.

The F1-Measure for change in speaker class when considering backchannels as transition is more than the F1-measure of the same class when backchannels are not the transition points. Backchannels constitute 19% of Switchboard corpus. When backchannel is considered as transition point there is a significant raise in F1-measure for change in speaker. This shows that the models we proposed learns well from the sequences of combination of features of utterances.

The area under the curve (AUC) value for J48 with 4-gram feature combinations when backchannels are considered as turn initiators is 0.677 for both the classes.

Discussion

Although the proposed models outperformed the baseline, there is still room for improvement with the features discussed in this work. We believe that from our results in experimental section, dialogue acts alone play a major role in predicting the turn-taking behavior. Unfortunately, the Switchboard corpus contains some dialogue acts, which do not provide the intention of the speaker, instead they indicate a type of error like *bad segmentation*, *transcription errors* etc..

Some of the dialogue acts which do not help in predicting whether there can be a change in speaker or not are:

- @-which indicates that the transcription is incorrect and/or has bad segmentation. It constitutes 1% of the Switchboard utterances.
 - **Ex:** o@ B.106 utt1: [It, + *[[this "it" needs separate slash mark because B.108 is an "aa" with A.107, not necessarily the original intended continuation of this utt]]
- %- Uninterpretable indicates that the utterance is cut off in such a way that it cannot be annotated with any dialog act. Switchboard contains 7% uninterpretable utterances.
 - **Ex:** % A.17 utt2: {C But } then down here I li, - /
- item x- Non-verbal behavior especially background noise, child crying, coughs etc? Occured 2 % of the times in Switchboard.
 - **Ex:** x A.83 utt1: .

If all the utterances were classified with correct dialog acts, we could build a better model to predict turn-taking which is crucial for the correct behavior of the dialogue system.

Also, current spoken dialogue systems wait for the speaker to provide a long pause to segment the speech. This reduces the responsiveness of spoken dialogue system, as further processing can start after the current speakers turn and some silence. Segmenting (Atterer, Baumann, and Schalangen 2008) and classifying the utterances in ongoing speech recognition stream will help build a good model for a turn-taking mechanism.

Conclusion

In this work we considered modelling turn-taking behavior as a predictive task. We conducted several experiments to observe the impacts of n-gram sequential modelling on turn-taking. It is shown by the fact that prediction is better until it reached trigrams and after that, despite the addition of more context, the model has worse results. Thus, the history considered is restricted to length three. Dialogue acts, change in speaker information, speaker IDs and duplicating the examples of one speaker to another were shown useful for deciding whether there should be a change in speaker at the

end of an utterance. A decision tree machine learning technique performed better with 4-grams of a combination of all the features. The proposed model with the significant feature set outperformed the baselines

We believe that the textual features of a dialog that were discussed in this work can be combined with speech and visual cues for building a good model to predict turn changes.

References

- Atterer, M., Baumann, T., Schlangen, D., 2008. Towards incremental end-of-utterance detection in dialogue systems. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, UK.
- Godfrey, J., Holliman, E., McDaniel, J., 1992. Switchboard: telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* 517 - 520. San Francisco, Calif.: IEEE
- Gravano, A., and Hirschberg J. 2011. Turn-Taking cues in task-oriented dialogue. *Computer Speech and Language*, 25(33):601-634.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, Ian H. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11:10-18.
- Jurafsky, D., Shriberg E., and Biasca, D. 1997a. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13, University of Colorado, Boulder.
- Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A. and Den, Y. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map-task dialogs. *Language and Speech* 41(3-4),295-321.
- Padilha, E., and Carletta, J. 2003. Nonverbal behaviours improving a simulation of small group discussion. In *Proceedings of 1st Nordic Symposium on Multimodal Communication*
- Sacks, H., Schegloff, E. A., and Jefferson, G. A. 1974. A simplest systematic for the organization of turn-taking in conversation. *Language* 50:735-996.
- Schlangen, D., 2006. From reaction to prediction: Experiments with computational models of turn-taking. In *Proceedings of Interspeech*, Panel on Prosody of Dialogue Acts and Turn-Taking, Pittsburgh, USA.
- Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Ess-Dykema, C. V., Ries, K., Shriberg, E., Jurafsky, D., Martin R., and Meteor, M. 2000. Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics* 26(3):339-371.
- Ward, N. G., Fuentes O., and Vega, R. 2010. Dialog Prediction for a General Model of Turn-Taking. In *Proceedings of Interspeech*.