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

June 2016

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.

Peter A Heeman, Thesis Advisor
Research Associate Professor

Alison Presmanes Hill
Research Assistant Professor

Xubo Song
Research Associate Professor

Chad C. Hagen
Assistant Professor

Jan van Santen
Professor

Dedication

Acknowledgements

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

June 2016
Thesis Advisor: Peter A Heeman

Conversations forms the bulk of everyday spoken interaction. Talk in conversation is organize in turns such that only one speaker speaks at a time, and the gaps between turns is minimized. Research in human-human conversations shows that turn transition projection is at the core of fluent and efficient conversation. Hence, correct projection of turn transition in spoken dialogue system (SDS) is important in order to improve user engagement. Recent SDS systems are using machine learning models based on local features of the speaker utterance in order to predict turn transition. For example, features based on utterance syntactic structure, pragmatics (dialog acts), prosodic. This thesis investigates if features that are a summary of past speaker behavior can improve turn transition predication. The suggested features are computed over multiple past turns of the current speaker. The features measure the *relative turn length* of the current turn, and the *relative floor control* of the current speaker. we believe that the summary features represent an evolving model of the other conversant. For example, speakers who, on average, use long turns, will 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.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 improve turn transitions prediction.

Part I

Introduction

Chapter 1

Introduction

Conversations are the main form for social engagement in which participant exchange utterances in order to reach a specific goal. Beside the basic level of utterances, a conversations contain different organizational systems that operate at different levels. For example, the speech act layer map each utterance as an action performed by the speaker and contain construct like adjutancy pair etc. Turn taking and assignment are another talk organization system. This system concern itself with turn construction and turn allocation.

Seminal work done by [**?**, **?**] which analyzed core characteristic of turn taking in conversations found that: Overwhelmingly, on party talks at a time; transition from one turn to the next with slight or no gap and slight or no overlap make the vast majority of transition; turn order is not fixed; turn size is not fixed.

More recent research measured the actual gaps between turns and the time it takes to produce a single word or a clause . The timing suggest that in order to keep the conversation fluent, listener must predict turn transition before the current speaker ended its current turn. To anticipate transitions, conversant uses mainly features that are derived from the last few utterances of the speaker: syntactic [**?**, **?**], prosodic [**?**, **?**, **?**], pragmatic [**?**]. To naturally engage in conversations with humans, spoken dialogue system (SDS) should incorporate a turn taking component. The component should enable it to predict turn ending and speaker transitions. However, most SDS use simple thresholds on the silent time. More recent advances use machine learning models [**?**] on local features (syntactic and prosodic). of the last one or two turns.

1.1 Thesis Statement

Conversation participants takes turns when speaking. Turn taking is a social organization system which assure the only one speaker speaks at a time (minimum overlaps), and that the gaps between turns is minimize. To minimize gaps between turns, speakers project turn transition using

various features of the speaker utterance (syntactic structure, pragmatic, prosodic). To participate in conversation with humans, spoken dialogue systems (SDS) should participate in turn taking. However, up until recently, turn taking in SDS systems consisted only of fixed time out mechanism. More recent research create machine learning models based on features extracted from the speaker utterance (referred to local features) and try to predict turn transition. This thesis investigates if features that a summary of past speaker behavior can help. The suggested features are computed over multiple past turns of the current speaker. The features measure the *relative turn length* of the current turn, and the *relative floor control* of the current speaker. we believe that the summary features represent an evolving model of the other conversant. For example, speakers who, on average, use long turns, will 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.

1.2 Approach

To test the summary features effectiveness, we used the NXT version of the Switchboard corpus [**?**, **?**] to train random forest models [**?**]. 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.

1.3 Dissertation Structure

The paper is organized as follows: section 2 presents related work. Section 3 introduces the local and the summary features. Section 4 describes the experiment. Section 5 shows the results obtained by training random forest models with and without the summary features. Finally, in section 6 we present our conclusion.

1.4 Contribution

Our results show that the summary features improve prediction performance in both area under the curve (AUC), 0.65 vs 0.63, and F1, 66.42% vs 54.97%. In addition, the model that was trained on all the features performs better than the local features model in both AUC, 0.82 vs 0.79, and F1, 74.87% vs 74.08%. The results show that using the summary features can help predict turn transitions.

Chapter 2

Related Work

This section covers the related work in both human-human conversations (conversation analysis and psycholinguistics) as well as human-machine conversations (spoken dialogue systems).

2.1 Human - Human Conversations

Early research suggested that turn transitions are either controlled by the speaker, which cue the listener using prosody [**?**] or gaze [**?**], or are projected by the listener based on local features (syntactic [**?**], prosodic [**?**] and pragmatic [**?**, **?**]). Measurements by Levinson et al. [**?**] of turn gaps and overlaps suggested that prediction plays a significant role in turn taking, while speaker cues mainly contribute as a final signal to begin talk [**?**]. Turn transition projection was verified in experiments by Riest et al. [**?**], which used conversation fragments with a controlled lexicostatistics information, intonation and amplitude. Prediction of turn transition was done by button press. The study concluded that the lexicostatistics information is critical to turn prediction while intonation less so. The study results prove the inherit capabilities of humans to predict turn transition based on features derived from the latest utterance. The study also helped identify what features can help SDS system designers when training a turn transition model.

Ford and Thompson [**?**] suggested that three types of cues converge at turn transition relevance point - syntactic cues, intonational features and current action completion. Of those three, we chose to use action completion as the local feature noting that those points often represent the end of syntactic units.

In a very recent study, Garrod and Pickering [**?**] suggest that turn production is a multi-stage process, which involves simultaneous comprehension and production. The listener tries to decode the speaker's intention as soon as possible, and decides on its own dialog act. Levinson et al. [**?**] suggested that based on the selected dialog act, the listener performs utterance conceptualization of, lemma retrieval, phonological retrieval and phonetic encoding. This background production

occur in parallel with closed monitoring of the speaker. Moreover, Garrod and Pickering [**?**] suggest that the listener tries to simulate the speaker's based on the speaker intention and rate of speech (also see an early theory of conversation oscillator in [**?**]). In this paper we try to extend these findings, and test if the listener monitors the speaker's past behavior in addition to the most recent turn.

2.2 Human - Computer Conversations

Traditional turn transition in spoken dialogue system were based on silence threshold. The system has predefined silence time after which it is assume that turn change occurred. The fact that continuous speech includes many gaps which are not transitions, and that the surrounding environment is noisy, led to many false positives and thus overall low usability of the system.

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 prosodic and word level. Schlangen [**?**] used 29 prosodic features, and 8 syntactic features based on ngram model of both the words and POS tags. In addition, similar to our model, they used measurement of the utterance length and turn length up to the last word. Arsikere et al. [**?**] describes a system that try to segment speech to utterance using prosodic as well as syntactic features. The prosodic features include pitch and energy, while the syntactic feature include words and POS ngram. Gravano and Hirschberg [**?**] used more complete set of local feature to predict speaker transition. The study used intonation, speaking rate (syllables per second), intensity and pitch level, utterance duration and voice quality. The model was trained and tested on the Columbia games corpus using various machine learning methods. The best result was 80.0% accuracy and was achieved by using linear SVM.

Nishitha and Rodney [**?**] used a model based on N gram 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. The result shows F1 measure of 0.67 for the trigram model. In this paper we based our baseline models on unigram and bigram of dialog acts. In addition, we filtered out all the non complete dialog acts which appear in the switchboard corpus.We also mapped the switchboard dialog acts from 148 dialog acts down to 9 in order to reduce data dimensionality. The prediction performance of our baseline model is comparable to the results in [**?**].

Chapter 3

Experiment

3.1 Data

This is the data (corpus itself)

3.2 Data Preparation

Figure **??** shows the experiment data pipeline. Data is imported from the NXT switchboard corpus [**?**] into a graph database [**?**], Figure **??**, 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.

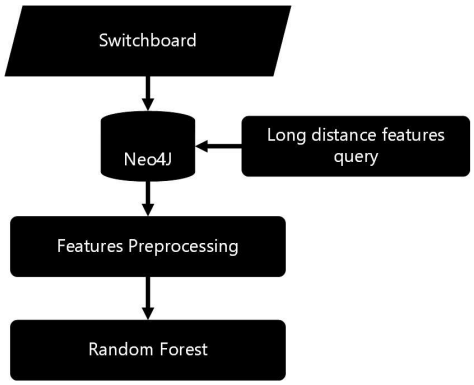

Figure 3.1: The experiment data pipeline

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

- Since we believe that it takes certain amount of time to build a stable conversational image,

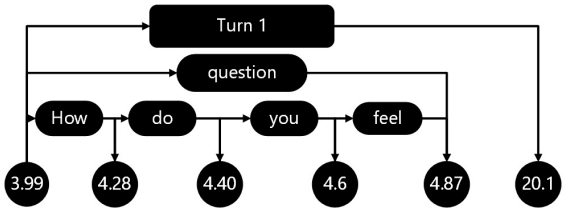

Figure 3.2: conversation graph data model

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

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 50633 to 37508.

- To reduce the dimensionality of the data, we grouped together switchboard dialog acts into dialog act classes. This reduced the number of dialog act from 148 to 9 dialog acts 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 was a turn change from dialogue act d_i to d_{i+1}.

To test the contribution of the summary features, we used a binary classifier with y_i as the outcome variable. We trained 4 models, 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 features: 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 classifiers ($N = 200$) [**?**] as the binary classifier. In order to find the optimal hyper parameters, we run 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 validation. We made sure that each conversation was entirely in a single fold. This way, each dialogue is entirely used for training or testing, but never for both at the same time.To perform the labeled cross validation, we labeled all the dialog acts in each conversation using a hash on the conversation name, such that all the dialog acts in a given conversation are either in the test or the train data. We report the accuracy of each model as well as the ROC curves and the area under the curve (AUC).

3.3 Results

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 3.2: Accuracy and AUC results

Table 2 shows the results of training a random forest for each model. We see that using the summary features provides better accuracy than using 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.

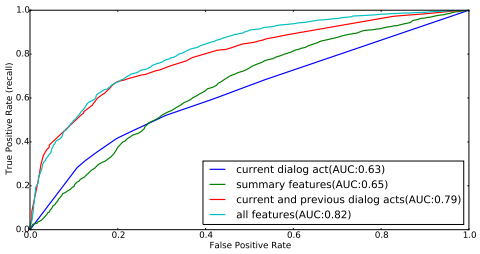

Figure 3.3: ROC curves and AUC of the different models

The effect can also be seen in Figure 3 which shows the ROC curves and the AUC for each set of features. We notice that the AUC of the summary feature is better than the base line (0.65 vs 0.63) and when adding the summary features to the local features we see the AUC improve (0.82

vs 0.79). This suggests that while the separation power of the summary features by themselves is lacking (AUC < 0.7), adding them to a classifier that uses local features yields better results.

Model	Precision	Recall	F1
Baseline 1	69.49%	45.52%	54.97%
Baseline 2	80.38%	68.80%	74.08%
Summary	64.55%	68.88%	66.42%
Full	76.17%	77.25%	74.87%

Table 3.3: Precision, recall and F1 results

Table 3 shows the precision, recall and F1 scores. Although our base line has high precision for the dialog act that leads to turn transition, it has very low recall. Using only the summary features improves recall and decreases precision by less leading to higher F1 score and overall better performance. Using all the local features improves precision which means that from dialog acts that were considered to lead to turn transition are classified correctly. If we use all the features, we lose precision (over just local features), but gain recall leading to the highest F1 score and the best performance.

Chapter 4

Conclusions

This paper suggests that adding summary features of past speaker behaviour can aid in anticipating turn transition when used with traditional local features.

The results show that the summary features, for the switchboard corpus, when combined with current and previous dialog acts does improve the prediction precision and recall. We thus conclude that using summary features should help a turn taking sub-system which is based only on local features.

Additional analyses should include testing on more local features (like syntax and prosody) as well as training and testing on other conversational corpora, such as task-based dialogues.

Also, we should consider introducing more summary features. For example, by creating a distribution over the syntactic structure types used by a speaker, we can measure the probability that the user will use a sentence or a clause in an average turn. The combination of these and other summary features might help future SDS systems to become more natural and user friendly.

4.1 Summary

This is the summary

4.2 Future Direction

Additional analyses should include testing on more local features (like syntax and prosody) as well as training and testing on other conversational corpora, such as task-based dialogues.

Also, we should consider introducing more summary features. For example, by creating a distribution over the syntactic structure types used by a speaker, we can measure the probability that the user will use a sentence or a clause in an average turn. The combination of these and other summary features might help future SDS systems to become more natural and user friendly.

Bibliography

[1] H. Arsikere, E. Shriberg, and U. Ozertem. Enhanced end-of-turn detection for speech to a personal assistant. In *2015 AAAI Spring Symposium Series*, 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 and E. Schapire. Random forests. In *Machine Learning*, pages 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*, 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 *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–608. IEEE, 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. pages 517–520, 1992.

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

[14] N. Guntakandla and R. Nielsen. Modelling turn-taking in human conversations., 2015.

[15] R. Hariharan, J. Hakkinen, and K. Laurila. Robust end-of-utterance detection for real-time speech recognition applications. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, pages 249–252. IEEE, 2001.

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

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

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

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

[20] 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.

[21] 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.

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

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

[24] 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.

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

[26] 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. Association for Computational Linguistics, 2009.

[27] 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.

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

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

[30] 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.

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

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

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

[38] 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. ACM.

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

Biographical Note

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.