Take or Wait? Learning Turn-Taking from Multiparty Data

Iolanda Leite, Hannaneh Hajishirzi, Sean Andrist, Jill F. Lehman
Disney Research, Pittsburgh
4720 Forbes Avenue, Pittsburgh, PA
{iolanda.leite, hannaneh.hajishirzi, sean.andrist, jill.lehman}@disneyresearch.com

Abstract

We build turn-taking models for autonomous characters in language-based interactions with small groups of children. Two models explore the use of support vector machines given the same multimodal features, but different methods for collecting turn-taking labels.

Introduction

Turn-taking plays an important role in the dynamics of human social interactions (Sacks, 1974). It may feel effortless, but turn-taking relies on a complex mix of contextual, verbal, and gestural cues that unfold over time, especially in multiparty settings. Such complexity favors machine learning over a static, rule-driven solution.

Although there is a large body of prior research on multimodal (Thomaz and Chao, 2011) and multiparty turn-taking in adults (Tur et al. 2010, Bohus and Horvitz 2011), most prior computational work with children has focused on turn-taking between an agent and a single child. Thus, we explore the use of Support Vector Machines to build turn-taking models for autonomous characters interacting with small groups of children.

We investigate two different turn-taking models with multimodal features. The first model was trained on the turn-taking decisions made by the human wizard who controlled the virtual character that interacted with our participants. To overcome the effects of his variable reaction times, a second model was created based on the judgments of annotators who were asked to make *take-or-wait* decisions at a theory-driven subset of moments in the video records of the children's games. We present results for these two models and some directions for future work.

Scenario, Participants and Setup

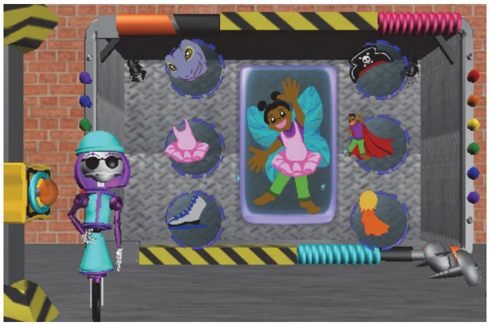

Figure 1. Robo Fashion World.

Our testbed for this work is Robo Fashion World, an interactive game designed to facilitate the collection of audio-visual language data from young children in groups of up to four, with or without adults. The game is hosted by Edith, an animated character who is responsible for mediating the interaction (Figure 1). During the game, children can request a silly change to the model by naming one of the fashion items on the board, or can request a picture of the model as it is to be taken home later.

A total of 65 children (34 females and 31 males), ages four to ten (M/SD = 6.8/1.9 years), played in 29 groups of ~3.2 members (with adults). Games lasted about nine minutes and were recorded with frontal and lateral cameras, as well as close-talk and linear array microphones.

A human wizard performed speech processing for Edith. The wizard's interface allowed signaling of a small number of linguistic events, e.g., reference to a game item, request for a picture, long silence, or multiple voices talking at once. The wizard was not given an explicit turn-taking paradigm, simply told that the children should have fun. As more than one interface option might be applicable at any given time, the wizard's decision about whether and what to signal implicitly defined the character's turn-taking behavior. Log files containing the timing and content of wizard actions, the behaviors employed by the character as a result, and the changing state of the game board were generated automatically. Our goal is to use this data to create a model that allows Edith to make autonomous and socially appropriate turn-taking decisions in future games.

Feature Extraction

Using the audio, video, and interaction logs, annotators extracted a set of behavioral and contextual features for each child. Video was used for *head orientation* (toward or away from Edith) and *gesture* (head shakes, pointing and emphasis). The close-talk microphone recordings were used to define six additional features for each utterance. *Pitch*, *power,* and *prompt* (a dialog context feature) were derived automatically. *Addressee* (whether an utterance was to Edith or not), *yes/no words,* and *valid asset words* (references to a fashion item visually available on the game board) were labeled by hand.

The Wizard Model (M_W)

With LibSVM (Chang and Lin 2011), the extracted features were used to train a binary classifier that predicts whether Edith should *take* the turn or *wait* at every 500 msec boundary (the time slice to be used in the autonomous character to balance component synchrony against perceivable delay). The training labels for M_W were based on the actions taken by the wizard during data collection. A *take* occurred at the end of a time slice in which the wizard performed an interface action. Slices while Edith spoke were excluded from the model because at present her speech cannot be interrupted (we anticipate building a complementary *hold/release* model when she can be). The remaining slices were assigned *waits.* M_W has three key characteristics: it encodes decision making at every moment by a human *in situ*; it reflects the inherent reaction time lag between a *take* decision and an action at the interface during which feature values may have changed; under the extant decisions, children had fun.

Several versions of M_W were built and tested with 29-fold cross validation against the wizard's behavior. Versions differed with respect to the number of prior time slices included, ranging from no prior history to feature vectors using eight previous slices (four seconds). Accuracy improved significantly up to four slices (two seconds) of history, and then flattened.

As Table 1 shows, M_W is a poor match to the wizard's actual decision making. This might be because the model's features do not capture regularities in the wizard's decisions, but we believe it is more likely due to the variable delays between when the wizard made his decision and when he signaled that decision at the interface. Indeed, given the large amount of overlapping speech in our games, the conditions M_W appeared to be reacting to could be quite different from the conditions at the time of intent, particularly because our models encode ongoing behavior over a two-second window.

The Annotator Model (M_A)

To overcome the reaction time lag in M_W, we asked two annotators to make *take-or-wait* decisions at the end of a subset of video segments of the children's games. The video segments began immediately after Edith changed the board and stopped at moments where, according to theory, it is appropriate to take the turn (e.g., after a long silence or at the end of a child's utterance). Every end-of-utterance in all 29 sessions was included in the set as well as an evenly distributed number of silences that did not occur at the end of utterances. Inter-rater agreement for the 4949 segments was high ($k = .739$, $p < .001$).

Using the same features described above, we then built M_A from the ~4000 segments where the annotators agreed. As with M_W, we explored multiple versions of M_A in a 29-fold cross validation and found that accuracy to its own ground truth peaked at two seconds of history. Table 1 shows that when *take/wait* labels are paired with the features that exist at the moment the decision is made, the feature set is able to capture much of the regularity in the annotators' decision making. We conjecture that M_A might, in fact, be a better representation of the wizard's intent than M_W.

Table 1. Performance with history = 4 time slices/2 seconds.

Model	Max F1	AUC	TPR	TNR
M_W	0.40	0.51	0.68	0.75
M_A	0.82	0.67	0.83	0.81

Of course we cannot tell from these results whether M_A is a good turn-taking model or one under which children would have fun. It is also important to note that M_A is of potential value only because its data were reactions to the same real behavior as M_W. We cannot stop the world to capture the wizard's decisions at the moment of intent, so it is the combination of the data generated by the wizard and the annotators that is critical to M_A's potential success.

Conclusions and Future Work

This work is a first step towards building turn-taking models that enable virtual characters to decide when they should take the turn in open-ended multi-child interactions. We presented the results of two models built using the same multimodal features but different approaches for collecting turn-taking labels. To further understand the differences between these models, we are planning to conduct a study to evaluate the impact of the different models' predictions across the full data set by asking subjects what they would do in situations where the models' actions differ. We also intend to compare the SVM-based models with a baseline model where the turn-taking decisions are purely rule-based.

References

Bohus, D., and Horvitz, E. 2011. Decisions about turns in multiparty conversation: from perception to action. In *Proceedings of the 13th international conference on multimodal interfaces*, ICMI ’11, 153–160. New York, NY, USA: ACM.

Chang, C.-C., and Lin, C.-J. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):27.

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

Thomaz, A. L., and Chao, C. 2011. Turn-taking based on information flow for fluent human-robot interaction. *AI Magazine* 32(4):53–63.

Tur, G., Stolcke, A., Voss, L., Peters, S., Hakkani-Tur, D., Dowding, J., Favre, B., Fernandez, R., Frampton, M., Frandsen, M., Frederickson, C., Graciarena, M., Kintzing, D., Leveque, K., Mason, S., Niekrasz, J., Purver, M., Riedhammer, K., Shriberg, E., Tien, J., Vergyri, D., and Yang, F. 2010. The calo meeting assistant system. *IEEE Transactions on Audio, Speech, and Language Processing,* 18(6):1601–1611.