Large Language Models Know What To Say But Not When To Speak

Muhammad Umair, Vasanth Sarathy, J.P de Ruiter
Department of Computer Science
Tufts University

Released Dataset

Background

Turn-Taking is a fundamental aspect of human communication, enabling smooth, fluid conversations. In everyday conversations, speakers alternate between speaker and listener roles without predetermined cues, relying on **Transition Relevance Places (TRPs) –** opportunities within a speaker's utterance where the listener may, but is not required to, take over the turn.

Large Language Models (LLMs) have shown promise in improving the turn-taking abilities of Spoken Dialogue Systems (SDS), particularly by identifying turn-final TRPs.

However, these models often struggle to predict more subtle, **within-turn TRPs**, where listeners could, but do not always, respond.

Challenges

Current LLM-based approaches to predicting *opportunities* for speech in natural, unscripted, interaction face two major challenges:

1. **Lack of ground-truth data for TRPs:** While TRPs between-turns can be easily identified due to speaker switches, TRPs within-turns are more difficult to label, particularly because there are few observable cues.
2. **Written vs. Spoken Language:** Most LLMs are trained on written language, which differs significantly in structure and usage from spoken language.

TRPs: The Data Problem

Common Methods for Identifying TRPs:

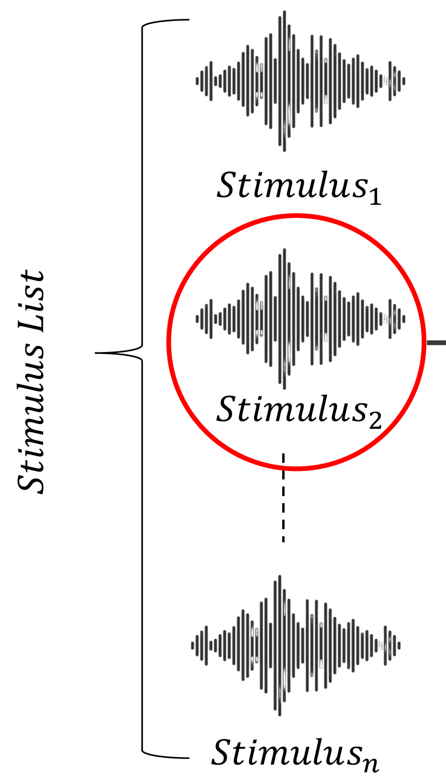
1. Detect points in conversation where a speaker switch occurs, as this most often occurs at a turn-final TRP.
2. Use expert annotators to identify speaking opportunities based on conversational cues.

Limitations of Current TRP Identification Methods:

1. **Limited Scope:** Speaker changes capture a small subset of all TRPs, as listeners can choose not to speak at a TRP, resulting in no visible transition.
2. **Subjectivity of Expert Annotations:** Expert labeling is subjective and does not reflect the same anticipatory process that interlocutors engage in.

Contributions

1. **Novel Dataset for TRP Detection:** We develop a highly ecologically valid participant-labeled dataset with annotations of **within-turn TRPs** in natural conversations.
2. **Simple TRP Prediction Task Formulation:** We provide a simple binary decision task for models to predict TRPs based on preceding linguistic information – in line with human mechanisms of TRP anticipation.
3. **Evaluation of LLMs:** We establish baseline performance by testing state-of-the-art LLMs on their ability to predict within-turn TRPs, offering insights into their limitations in natural dialogue.



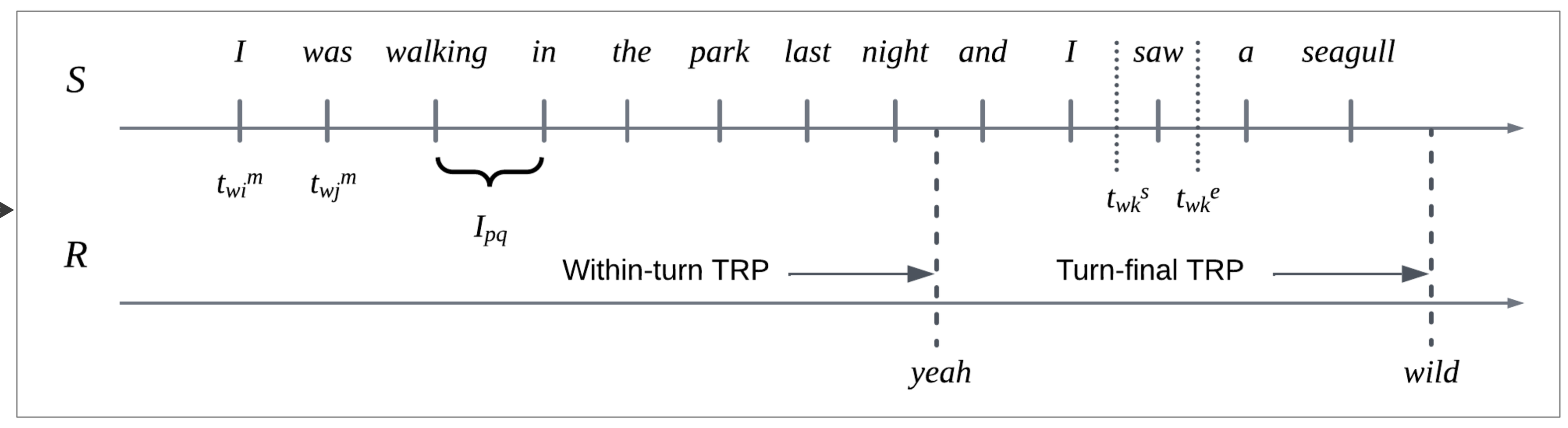


Figure: Participants listened to a stimulus (S) and produced auditory responses (R) to indicate their perception of TRPs. Each word in the stimulus and the response has a start and end time. Intervals are between adjacent words.

Identifying TRP Locations

Participant Task: Each participant listened to conversational **stimulus turns** – short segments of natural audio – and gave auditory feedback when they perceived a point where it was appropriate to speak i.e., a TRP.

Stimulus Data: Was drawn from the *In Conversation Corpus* (ICC), a collection of 93 unscripted, 25-minute, conversations between pairs of undergraduate students. From this corpus, 55 turns were selected (**28.33 minutes of talk)**, focusing specifically on segments that contained multiple opportunities for turn-taking.

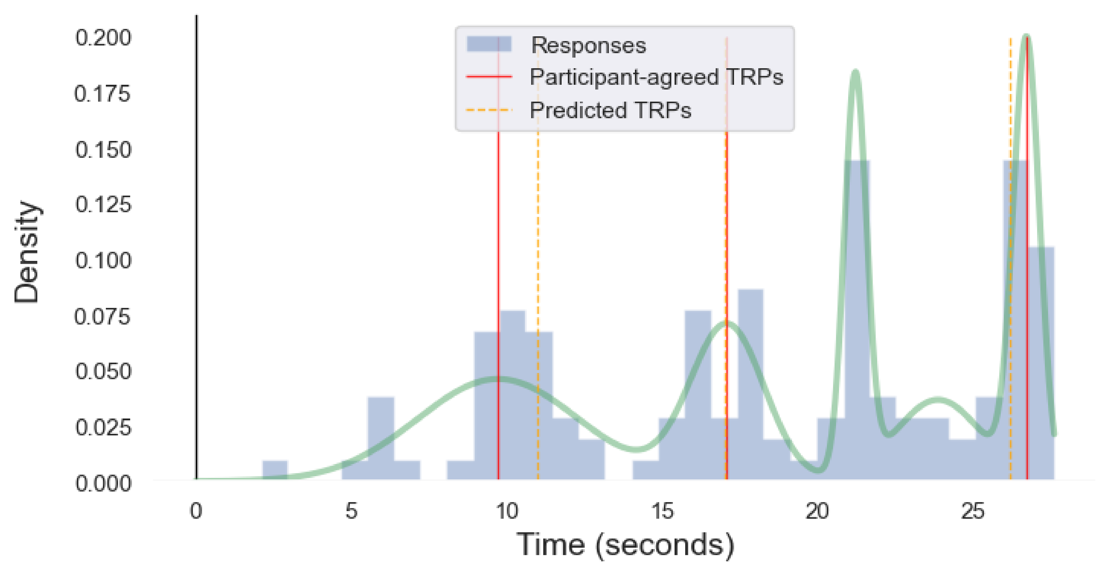


Figure: Distribution of participant responses, the times at which participants agreed a TRP occurred, and model predictions of TRPs for a single stimulus S.

Binary TRP Prediction Task

Formally, we can define a stimulus S as having N words, each with a start and end time. Participants produce M responses for S, also with a start and end time.

$$S = \langle (w_1, t^s_{w_1}, t^e_{w_1}), \ldots, (w_n, t^s_{w_n}, t^e_{w_n}) \rangle$$

$$R = \langle (\tilde{w}_1, t^s_{\tilde{w}_1}, t^e_{\tilde{w}_1}), \ldots, (\tilde{w}_M, t^s_{\tilde{w}_M}, t^e_{\tilde{w}_M}) \rangle$$

We further define a Prefix *P* as a sequence of words in *S* from the first up to the i^th word, and P_S as the set of all prefixes in S.

$$P_i \quad = \langle w_1, \ldots, w_I \rangle ; \forall w_i \in P_i, w_i \in S$$

$$|P_S| = N$$

We also define $T_i \in \{0, 1\}$ as a binary random variable for intervals $I_{i,j}, 1 \le i, j \le N, j = i + 1$ between subsequent words, and $T_{R,S}$ as the set of predictions after each word.

$$T_{R,S} = \langle T_1, \ldots, T_N \rangle$$

Task Definition: *Given a stimulus S, and the set of all prefixes P_S, where each T_i in $T^{Predicted}_{R,S}$ occurs after each of the prefixes P_i in P_S.*

Evaluation

Precision and Recall: Measure the accuracy of the predicted TRPs and the model's ability to identify all relevant TRPs.

F1 Score: The harmonic mean of precision and recall, highlighting the balance between them.

Free-Marginal Multi-rater Kappa: A measure of agreement over change between model predictions and participant-labeled TRPs.

$$k_{free} = \frac{[\frac{1}{Nn(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{K} n_{ij}^2 - Nn] - \frac{1}{k}}{1 - \frac{1}{k}}$$

Temporal Metrics: Quantify how closely model predictions align with participant-identified TRPs, measuring the error in timing predictions.

$$d^S_{i,j} \quad = min(|i - j|) \forall j \in T_j^{Participants} = 1$$

$$D_S = \langle d^S_{I,j}, \ldots, d^S_{p,q} \rangle$$

$$NMAE \quad = \sum_{i=1}^{|D|} d^S_{i,j} \qquad NMSE \quad = \sum_{i=1}^{|D|} (d^S_{i,j})^2$$

Takeaways

1. **LLM Performance:** Despite their success with written-language, state-of-the-art LLMs **perform poorly in predicting within-turn TRPs** in natural, unscripted conversation.
2. **Timing and Alignment:** Models showed low precision, recall, and agreement with human-labeled TRPs, with substantial timing errors.
3. **Ecological Validity:** LLMs should be exposed to ecologically valid, natural spoken-first language during training to attempt to mimic human-like turn-taking behavior.

Limitations

1. **Linguistic Focus:** LLMs were asked to predict TRPs using linguistic information only, whereas human participants had access to both prosodic and linguistic information.
2. **Dataset Specificity:** The stimuli used were obtained from the ICC, which contains information, unscripted dialogues. Replicating this work across existing datasets is necessary to confirm broader applicability.
3. **ICL Limitations:** Our task is highly sensitive to prompt design, and it may be the case that we need to further explore task adaptation strategies.

This research was supported in part by Other Transaction award HR00112490378 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program. We would also like to thank **Grace Hustace** for her contributions in data collection and **Dr. Julia Mertens** for early-stage discussions.

EMNLP 2024