**Introduction**

Natural communication between humans is a highly interactive process.
- Speakers choose an utterance which they believe has high chance of achieving their communicative goal.
- They will then monitor the listener’s behavior to see whether this goal is actually being achieved and give feedback when necessary.

**Goal:** improve interactive NLP systems by adding monitoring and feedback capability in real time.

**The GIVE domain**

- Users have to solve a puzzle in a 3D environment.
- They can interact with objects in the world (e.g. click on buttons) and move freely in space.
- NLG systems guide users by generating instructions, including referring expressions (REs) for objects in the environment.
- Grounding problem: Systems have to predict (mis)understanding of a referent and prevent mistakes by providing corrective feedback.

**Our research question**

Given a referring expression, how do we predict what the user has understood as its referent?

---

**Model of RE resolution**

When receiving an instruction containing a referring expression $r$ at a given world state $s$, the user resolves $r$ to an object $a$. The user then moves towards $a$, exhibiting behavior $\sigma$. A probabilistic model over possible referents $p(a|r,s,\sigma) \propto p_{\text{sem}}(a|r,s) \cdot p_{\text{obs}}(a|\sigma)$

- $p_{\text{sem}}$ and $p_{\text{obs}}$ are separately trained, log-linear models.
- Both can generalize to unseen worlds.
- Features:
  - $p_{\text{sem}}$: semantic properties, potential sources of confusion, and visual salience
  - $p_{\text{obs}}$: distance, angle, visual salience and their evolution in time

**Data**

- Interaction corpora from the GIVE challenges, consisting of
  - automatically generated instructions
  - recorded user movements and actions
- Test data: 5028 episodes from the GIVE-2 challenge
- Training data: 3414 episodes from the GIVE-2.5 challenge
- Different virtual worlds, users & NLG systems between training/test data

**Episode**

An episode consists of the events between an instruction and the user’s action.

---

**Results**

**Prediction accuracy**

Given $r$, $s$, and $\sigma$ until some $t > t_0$, predict the referent understood by the user: $\arg \max_\alpha p(a|r,s,\sigma)$

**Feedback decision**

Given $r$, $s$, and $\sigma$ until some $t > t_0$, decide to give feedback if $p(a) - p(a') > \theta$ for some object $a \neq a'$ (here $\theta = 0.1$).

Feedback should be provided if the user was going to make a mistake, i.e. $a \neq a'$.

---

**Conclusions and Future Work**

- Our model predicts how a user is resolving the REs generated by an interactive system.
- The model updates initial estimate continuously based on observations.
- Next steps:
  - more time-aware model for $p_{\text{obs}}$
  - evaluate model in an end-to-end situated NLG system
  - explore use in other domains, e.g. navigation systems or less situated environments.