Structuring information in interactive natural language generation

Alexander Koller
koller@ling.uni-potsdam.de

Nikos Engonopoulos
engonopo@uni-potsdam.de

Martín Villalba
mvillalb@uni-potsdam.de

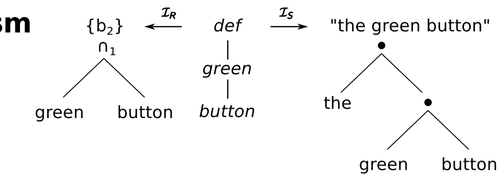
SFB 632 - D6

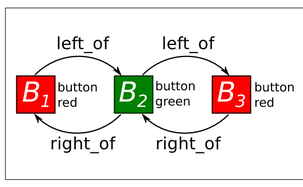
Research questions

- Given a target object, how to generate the **referring expression** that has the **highest chance** of being correctly understood?
- Given a referring expression and the user's observed behavior, how do we **predict what the user has understood** as its object?
- How do we **recognize that a user has not understood**, and generate the right instruction to **correct it** with appropriate **contrastive focus**?

Generation with semantically interpreted grammars

We use a **synchronous grammar formalism** relating **sets of objects** from a first order model of the world to **natural language strings** describing them, via an **abstract syntactic structure**.





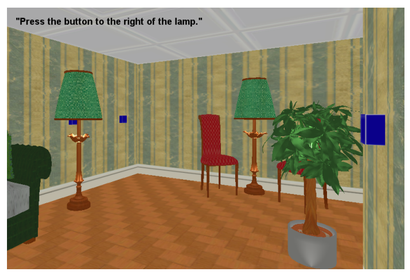
Grammar rule	String	Denotation
NP → def(N)	the • w_1	uniq(R_1) = if (R_1 is singleton) then R_1 else ∅
N → leftof(N, NP)	w_1 • to the left of • w_2	{ $a \in R_1$ | *exists* $b \in R_2$ s.t. (a,b) ∈ |left_of| }
N → green(N)	*green* • w_1	|*green*| ∩ R_1
N → red(N)	red • w_1	|*red*| ∩ R_1
N → button	button	|button|

We parse the target object into a **finite chart**, a data structure that represents the (possibly infinite) set of all valid referring expressions.

```
NP/{b1}          → def(N/{b1})
NP/{b2}          → def(N/{b2})
N/{b1}           → leftof(N/{b1, b2, b3}, NP/{b2})
N/{b1, b3}       → red(N/{b1, b2, b3})
N/{b2}           → green(N/{b1, b2, b3})
N/{b1, b2, b3}   → button
  ⋮
```

We select the **best referring expression** out of the chart, i.e. the one maximizing the probability $P_{sem}(b^*|r, s)$ of b^* being r.

The GIVE Challenge



Users have to solve a puzzle in a 3D world. They can interact with objects in the world (e.g. click on buttons) and move freely in the environment.

NLG systems guide the users by generating instructions, including referring expressions for objects in the environment.

Grounding problem: Systems have to **predict** (mis)understanding of an object and **prevent** mistakes by providing corrective **feedback**.

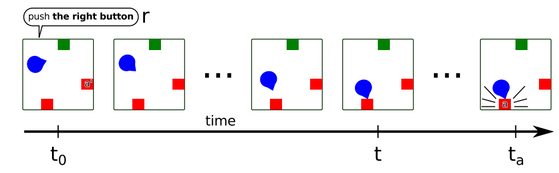
The interaction corpora comprises over 2500 games and more than 340h of recorded interactions.

Probabilistic semantic model

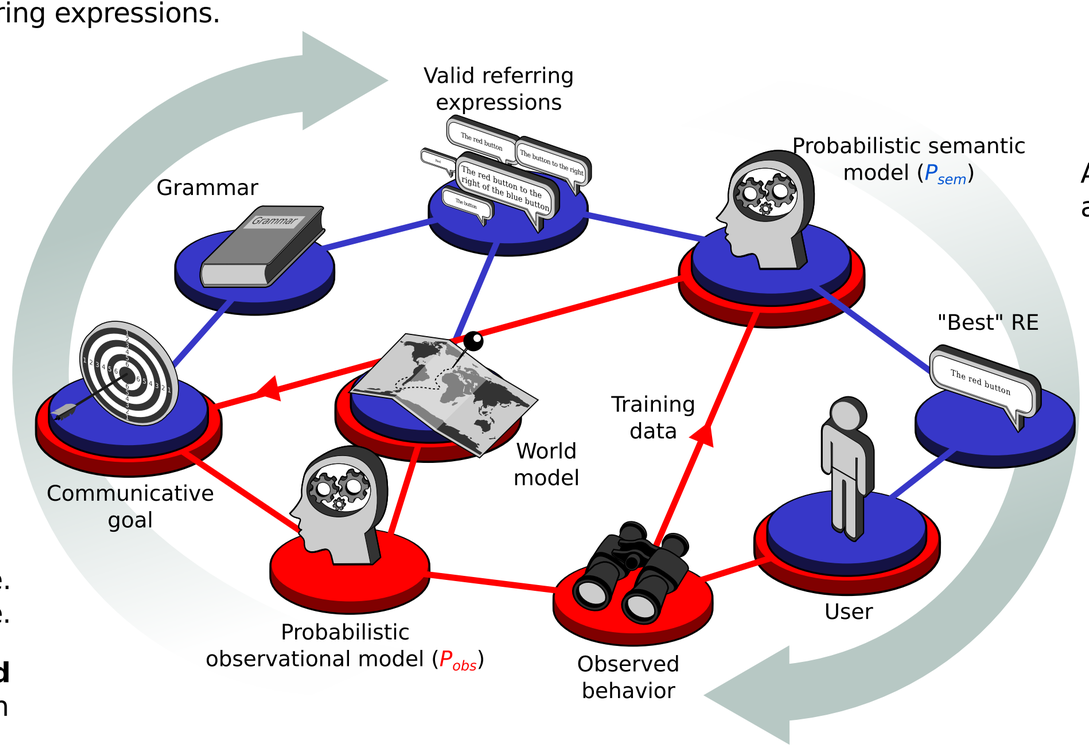
Purpose: predict the probability distribution P(*b*|*r*) of the RE *r* **being understood** as an object *b* over the set of all objects in the world.

The **probabilistic semantic model** (P_{sem}) uses information available at *instruction giving time*: syntactic and semantic features of the utterance, potential sources of confusion, visual salience, spatial features (distance, angle, etc.).

Model parameters are automatically **learnt from recorded interactions** using machine learning methods.



An **episode** consists of the recorded events between an instruction and the user's action.



Probabilistic observational model

Purpose: predict the probability distribution *P(b|σ)* of resolving the RE to the object *b* **based on the behavior *σ***.

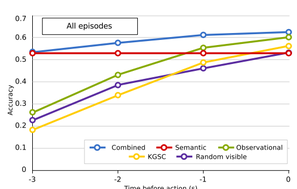
The **probabilistic observational model** (P_{obs}) collects new information continuously: eye-tracking, spatial features, visual salience, and the evolution of this information over time. The model **does not rely** on the text of the given instruction.

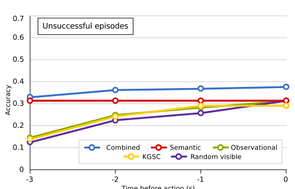
Like the P_{sem} model, parameters are automatically learnt from recorded interactions.

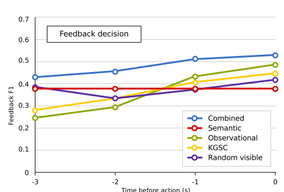
Evaluation

Training data: 3414 episodes for P_{sem} and 6478 tuples <σ,b> for P_{obs} from the GIVE-2.5 Challenge.
Test data: 5028 episodes from the GIVE-2 Challenge.

The model combining P_{obs} and P_{sem} **outperformed its components and two baselines** on prediction and feedback accuracy.







Prediction accuracy
Predict the object understood by the user: given r, s and σ until some $t > t_0$, predict the object understood by the user: $\text{argmax}_a\ P(a|r, s, \sigma)$

Feedback accuracy
Provide feedback if the system predicts the user is about to make a mistake: given r, s and σ until some $t > t_0$, give feedback if $p(a')-p(a^*) > \theta$ for some object $a' \neq a^*$ ($\theta = 0.1$)

Eye-tracking information has shown to improve accuracy for the P_{obs} model in **harder** episodes. Evaluation of the success and perceived quality of the generation system output with live users using **crowdsoucing** is underway.

Combined probabilistic model

When receiving an instruction containing a **referring expression *r*** at a given **world state** *s*, the user resolves *r* to an object *b*. The user then moves towards *b*, exhibiting **behavior** σ.

$$P(b \mid r, s, \sigma) \propto P_{sem}(b|r,s) \bullet P_{obs}(b|\sigma)$$

P_{sem} and P_{obs} are log-linear models, trained separately on recorded interactions and then combined. Both can generalize to unseen worlds.

Conclusions and ongoing work

- We have developed a **probabilistic semantic model** $P_{sem}(b \mid r, s)$ that predicts to which object *a* will the user resolve the RE *r* in the scene *s*, and a **probabilistic observational model** $P_{obs}(b \mid \sigma)$ that updates in real time, based on the user's behavior σ.
- We also presented an algorithm that computes the RE *r* that has the highest probability of being understood as the target object *a*.
- Our evaluation on recorded data shows that the combined model predicts a more accurately than the semantic model alone. We are currently training and testing our systems on live interactions with real users, using a **crowdsourcing** platform.
- Development of contrastive focus is also underway: as P_{obs} determines which object has captured the user's attention, we can identify the most distinguishing properties between the target object and the main distractor, and generate a contrastive RE.

Further reading:

Nikos Engonopoulos, Martín Villalba, Ivan Titov and Alexander Koller: *Predicting the resolution of referring expressions from user behavior.* Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP). Seattle, WA, USA. October 2013.

Nikos Engonopoulos and Alexander Koller: *Generating effective referring expressions using charts.* Proceedings of the 8th International Natural Language Generation Conference (INLG). Philadephia, PA, USA. June 2014.

Alexander Koller, Kristina Striegnitz, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore and Jon Oberlander: *The First Challenge on Generating Instructions in Virtual Environments.* In E. Krahmer and M. Theune (eds.), Empirical Methods in Natural Language Generation, volume 5790 of LNAI. Springer, 2010.

Nikolina Koleva, Martín Villalba, Maria Staudte and Alexander Koller: *The impact of listener gaze on predicting reference resolution.* Currently under review.