Predicting grounding state for adaptive explanation generation in analogical problem-solving

Lina Mavrina and **Stefan Kopp**
Social Cognitive Systems Group,
Bielefeld University, Bielefeld, Germany
{lvaronina,skopp}@techfak.uni-bielefeld.de

Abstract

This paper's main contribution is a Bayesian *hierarchical grounding state prediction model* implemented in an adaptive explainer agent assisting users with analogical problem-solving. This model lets the agent adapt dialogue moves regarding previously unmentioned domain entities that are similar to the ones already explained when they are instances of the same generalised schema in different domains. Learning such schemata facilitates knowledge transfer between domains and plays an important role in analogical reasoning. An explainer agent should be able to predict to what extent the explainee has learned to induce a schema in order to build up on this in the explanation process and make it more cooperative. This paper describes the approach of hierarchical grounding state prediction, introduces the analogy-based explanation generation process and the agent architecture implemented for this approach, as well as provides some example interactions as the first developers' evaluation of the system in preparation for upcoming empirical studies.

1 Introduction

Explanations are complex social processes that are actively shaped by both explainer and explainee throughout the course of their interaction (Miller, 2019, Rohlfing et al., 2021). Dynamic changes in the mental states of the explainee pertaining to their understanding of the *explanandum* (i.e., the object of the explanation) should be monitored and predicted by the explainer based on observable evidence, such as conversational feedback or clarification requests posed. These predictions should be then used to continuously re-conceptualise the *explanans* (i.e., the way in which the explanandum is presented by the explainer during explanation) (Rohlfing et al., 2021). Similar principles can be applied to human communication in general: active cooperation of the interlocutors and their stepwise co-construction of the interaction and the *common ground*, i.e., "their mutual, common, or joint knowledge, beliefs, and suppositions" (Clark, 1996, p. 93), as well as mentalising over relevant mental states of each other based on observable evidence (Kopp and Krämer, 2021). However, these principles of cooperative communicative behaviour are rarely applied in modern dialogue systems.

While explanations in a narrower sense serve as answers to *why?*-questions and explain causes of events, they can also serve other functions such as providing process narratives or instructions (Miller, 2019). In assistive scenarios, instruction and guidance during problem-solving are important functions of explanations. Analogy-based explanations specifically can help people transfer knowledge from one domain to another, for instance, via the process of schema induction. During this process, a generalised schema, i.e., "an abstract category that the individual analogs instantiate in different ways" (Gick and Holyoak, 1983, p. 8), can be induced from a range of specific examples and then applied to a new target domain.

This paper introduces an architecture for an assistive agent that guides the user through the process of problem-solving via adaptive explanations. The agent presents analogous stories from other domains hinting at the desired solution of the target problem, and helps the user understand similarities and differences between these stories, as well as induce and apply generalised schemata instantiated in the stories. In order to find good analogies, the agent uses graph-based knowledge representation to compare the examples and the target problem according to *structure-mapping theory* (Gentner, 1983). In order to be adaptive, the agent bases its explanation generation on predictions of grounding state of domain entities (DEs). These predictions are continuously updated via Bayesian inference.

The main contribution of the current research is the *hierarchical grounding state prediction model*.

This approach allows the agent to adapt dialogue moves regarding previously unmentioned domain entities if they are related to the ones already explained via a common schema. The model and the architecture facilitating this kind of inference will be described in more detail in section 3, and their limitations will be discussed in a special section after the conclusion.

So far, the system has only been tested by the authors using different types of feedback and observing the behaviour of the agent. Some example dialogues showcasing the adaptivity of the system will be presented in section 4. The empirical evaluation of the system requires a series of laboratory studies in order to gain a comprehensive understanding of the impact of various factors present during adaptive spoken interaction. These studies are currently being planned and prepared for by the authors.

2 Background and related work

2.1 Adaptive explanation generation

With the rise of machine learning and specifically deep learning, the focus of research on explanations in human-machine interaction has been primarily on explanations of artificial systems and their decisions (Mueller et al., 2019). However, often these explanations are conceptualised and presented in a one-off and static way that may not be sufficient for diverse stakeholders interested in them (Suresh et al., 2021; Lakkaraju et al., 2022). An increasing amount of research is currently calling for incorporation of findings from social and cognitive sciences into explanation generation to make it interactive and adaptable towards specific goals, needs, expertise and changing levels of understanding of the explainee (Miller, 2019; El-Assady et al., 2019; Shvo et al., 2020; Sokol and Flach, 2020; Dazeley et al., 2021; Rohlfing et al., 2021; Lakkaraju et al., 2022).

The process of explanation generation can be divided into two parts: the *cognitive process* responsible for the generation of causes, and the *social process* responsible for construction and presentation of the explanans, as well as interpretation of the signals of explainee's understanding (Dazeley et al., 2021). The social process can also be seen as an interaction pattern consisting of joint actions that are facilitated by the processes of co-construction and scaffolding, during which the explainer should strive to build explanations from the knowledge the explainee already possesses, yet enrich it with additional relevant information (Rohlfing et al., 2021). This work focuses on the social process of explanation generation that can be studied and applied across a multitude of domains, not just in the field of explainable artificial intelligence.

An explainer agent incorporating the complexity of the explanation generation process requires (1) a rich and dynamic explainee model, describing relevant mental states and the level of understanding of the explanandum with appropriate granularity, (2) representations of domain knowledge, dialogue state and history, as well as (3) capabilities to continuously reason over these representations to select explanation strategies, dialogue moves and content under uncertainty inherent to communication. Concepts such as *Theory of Mind*, i.e., the ability to attribute mental states such as beliefs, goals and intentions to self and others (Premack and Woodruff, 1978), *mentalising*, i.e., the ability to predict the actions of others based on their desires, knowledge and beliefs (Frith and Frith, 2006), and *common ground* play an important role here (Miller, 2019; Shvo et al., 2020; Kopp and Krämer, 2021; Rohlfing et al., 2021). A major challenge for this research is the lack of high-quality training data for explanation dialogues, which means that the parameters of the models are hard to pre-train in advance and the system has to be able to adapt online relying only on the data observed during interaction.

Hereby, approaches used in older expert and tutoring systems can be revisited and adapted. One example is the EDGE explanation system described in Cawsey (1993). Here, inference rules are used to update the level of knowledge of the explainee stored in the user model. There are direct inference rules that concern entities under discussion and indirect inference rules that concern unmentioned entities. The former are based on the user input and update the user model, while the latter are conditions that are checked against the user model if the system requires the corresponding information to construct an explanation. The system presented in this paper similarly aims to infer the grounding state of unmentioned entities, but realises it with a hierarchical probabilistic model.

Speaking of implemented systems adapting the social process of explanation generation, here are some more recent examples. Robrecht and Kopp's (2023) SNAPE model uses online planning in form of Monte Carlo Tree Search to solve a non-stationary Markov Decision Process for explanation generation, where transition probabilities depend on the level of understanding for concepts under discussion as observed by the system from user feedback. Axelsson and Skantze (2023) work on adaptive presentation. Their agent adapts its generation behaviour based on the grounding levels of various concepts as inferred from observed multimodal user feedback and stored in a knowledge graph.

2.2 Models of common ground in dialogue systems

As previously mentioned, the concept of common ground is important for adaptive explanation generation, as well as adaptive dialogue in general. Empirical evidence suggests that representations of common ground in humans are richer than a mere binary of grounded vs. ungrounded, however, these representations are still required to be efficient to support real-time language use (Brown-Schmidt, 2012). Stone and Lascarides (2010) distinguish between two types of grounding models: symbolic approaches based on discourse coherence and probabilistic approaches based on inference from observed evidence. Both of these approaches have been used in earlier-generations dialogue systems, a prominent example of the former is Traum and Larsson (2003), while the latter was pioneered by Paek and Horvitz (2000). Stone and Lascarides (2010), however, point out that both of these approaches have limitations. For instance, the probabilistic approaches were primarily used to predict whether the system had understood the user during slot-filling, i.e., collecting of the parameters of the user's query. Yet for cooperative dialogue, predicting whether the user had understood the system is equally important. Thus Stone and Lascarides (2010) integrate both types of approaches in a theoretical framework consisting of a dynamic Bayesian network (DBN) model of dialogue that represents the relationships between interlocutors' mental states, evolving dialogue context, discourse moves and observable evidence produced by interlocutors over time.

Buschmeier and Kopp (2018), too, use a DBN to represent the dependency of the probabilistic grounding state on the so-called *attributed listener state* (ALS) over time. The ALS consists of several variables based on communicative functions of linguistic feedback (Allwood et al., 1992; Kopp et al., 2008), namely contact, perception, understanding, acceptance and agreement which are inferred within the DBN based on incoming multimodal data and interaction context.

Axelsson and Skantze's (2023) adaptive presenter agent stores grounding as labels of properties in the domain knowledge graph, and these labels are updated based on the user feedback category obtained from a random forest classifier (positive, negative or neutral feedback). SNAPE (Robrecht and Kopp, 2023) similarly represents the grounding state via level of understanding (a concept can be either grounded or not) regarding relationships in a knowledge-graph-based domain model. Di Maro et al. (2021) focus on detecting conflicts during interaction leading to inconsistent state of common ground. They conceptualise their common ground representation in terms of personal common ground consisting of dialogue history, and communal common ground consisting of domain knowledge that is shared between the agent and the user. On the technical level, their common ground representation is implemented as a graph database. A similar approach is also pursued in this work.

2.3 Analogical problem-solving

The general principle of analogical reasoning lies in the concept of *mapping*, wherein correspondences are found between the *source* (also called *base*, i.e., known body of information) and the *target* (problem to be solved) of the analogy (Gick and Holyoak, 1983). Gentner (1983) defines the so-called *structure-mapping theory* describing interpretation rules for analogies. This theory postulates that an analogy is characterised by the mapping of structural relations between entities within base and target, rather than the surface-level similarity of their features, and that this mapping is governed by the principle of *systematicity*, i.e., the existence of related higher-order relations. The key concepts of the structure-mapping theory are supported by empirical evidence (Gentner and Maravilla, 2017).

As mentioned before, analogical reasoning is closely related to the process of schema induction, during which a generalised schema is extracted from specific examples. Gick and Holyoak (1983) found that, when given two analogy sources, the participants were able to derive the generalised problem schema as a byproduct of comparison of the sources, and that the quality of the generated schema was a positive predictor for the transfer of the analogy to the target. Similar results were obtained by Gentner et al. (2003) who additionally showed that increasing the degree of guidance during analogy training increased the rate of transfer during the exercise. These findings suggest that an adaptive explainer/tutoring agent may have a positive effect on the success of analogical transfer in problem-solving.

To be able to interpret this work in the bigger context of research on analogical reasoning, a set of frequently used problems from the experiments by Gick and Holyoak (1983) was chosen as the use case for the agent. The explainee is required to solve the *Radiation* problem first posed by Duncker (1945) with the help of various analogs from different domains. In the *Radiation* problem the user is asked to imagine they are a doctor and have to find appropriate treatment for a patient with an inoperable tumor. The tumor can be destroyed with high-intensity radiation, but such procedure would also destroy the healthy tissue the radiation would pass through on the way to the tumor. While there are several possible solutions to the *Radiation* problem, the desired one is the so-called *convergence solution* where multiple weaker forces converge on the target, such as several low-intensity radiation rays from different directions that will not damage the healthy tissue, but combined will destroy the tumor. Further information on the use case will be provided in section 4.

3 Agent architecture

The core components of the architecture facilitating predictive grounding state inference are depicted in figure 1:

1. the *dialogue manager* based on the `flexdiam` architecture described in Yaghoubzadeh and Kopp (2017), extended for grounding state prediction and explanation generation, and

2. the *memory* component in form of a graph database that stores multiple types of information, such as *domain model*, *conversational record* (i.e., interaction-related information that was made public to interlocutors) (Thomason, 2003) and *dialogue information state* (DIS) incorporating the agent's prediction about current grounding state of domain entities (Buschmeier and Kopp, 2012).

In this section, these will be described in more detail. Additionally, a subsection will be devoted to the natural language understanding (NLU) component of the architecture to discuss an example use of state-of-the-art large language models (LLMs) in adaptive dialogue interaction.

3.1 Memory component

The memory component stores all information that is available to the agent at runtime in the form of a graph defining relationships between various types of entities (figure 2). Currently, these include the following.

- DE nodes: structured representation of domain knowledge is important for the application of the *structure-mapping theory* in order to determine the best analogy for the target among the sources. This representation includes abstractions of relations and actions in the form of generalised schemata, as well as instances of these schemata in source and target examples. The model can support a higher granularity of domain knowledge representation if necessary. DE nodes are initialised at the start of the interaction and do not change throughout.

- DIS nodes for domain entities: DE nodes for schemata and schema instances have corresponding DIS nodes that store the parameters of the probability distribution describing the current belief of the agent about the grounding state G of an entity. These parameters are initialised when the entity first becomes significant, for instance, by being introduced by the system, and updated whenever relevant evidence of understanding is provided by the user.

- Conversational record nodes: these store information about employed dialogue moves and user feedback concerning a specific DIS node. New nodes in this category are continuously created throughout the interaction, but once added to the graph, they remain unchanged.

The memory component is implemented using the graph database framework Neo4j[1].

[1] `https://neo4j.com/`

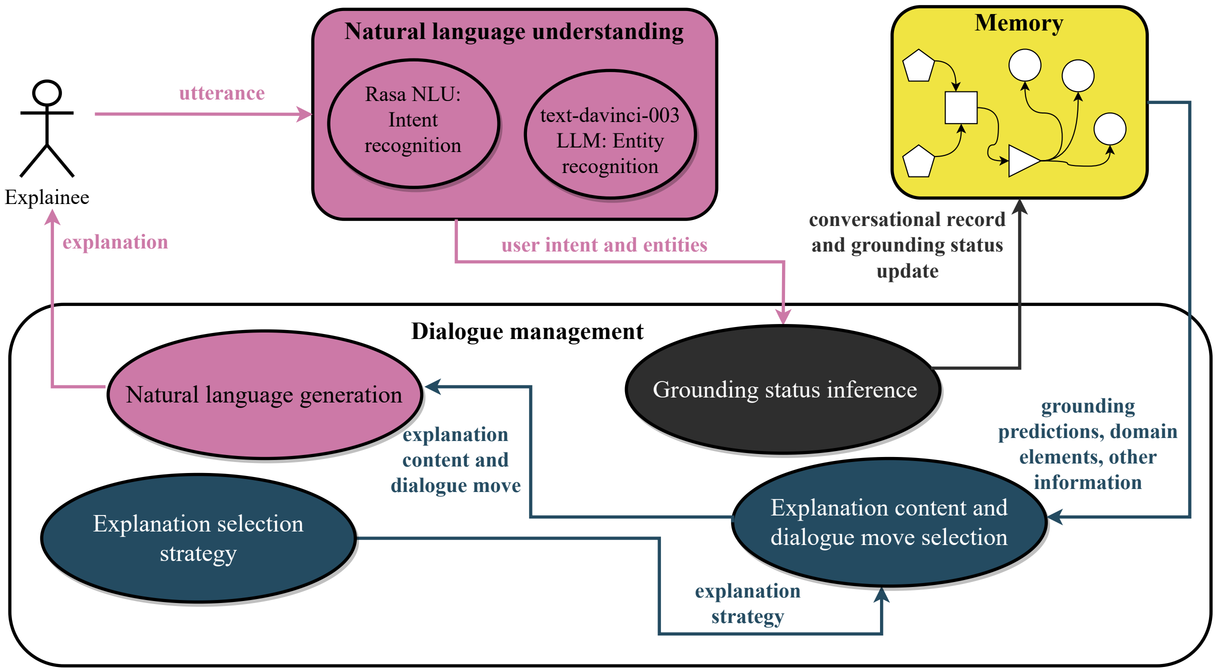

Figure 1: The architecture of the explainer agent.

3.2 Dialogue manager

As can be seen in figure 1, dialogue management essentially consists of two subsystems: *grounding state inference* and *explanation planning*. The dialogue manager is implemented in Python using the architecture called `flexdiam` (Yaghoubzadeh and Kopp, 2017) that was developed for spoken interaction in assistive settings utilising approaches tailored to dialogues with high degrees of uncertainty, which is also beneficial for a tutoring scenario.

Predictive grounding state inference

As previously mentioned, the belief of the agent about the grounding state G of a DE is described by a probability distribution. The parameters of this distribution are initialised when the entity becomes relevant for the first time during the explanation process. This initial distribution constitutes a uniform prior over the grounding state belief $P(G)$. When evidence of understanding U relevant to the entity is observed by the agent, it is used to calculate the posterior distribution $P(G|U)$ based on the Bayes' theorem:

$$P(G|U) \propto P(G) \times P(U|G) \tag{1}$$

Once the posterior is computed, it becomes the new prior distribution for the grounding state belief. In order to make the calculation of the posterior tractable at interaction time, the system uses conjugate priors for corresponding evidence likelihoods (Lambert, 2018). As the model for grounding state inference is hierarchical, two pairs of likelihoods and conjugate priors are used in the system, depending on the type of DE they are assigned to.

The lower level of the inference model deals with beliefs about the grounding state of schema instances. A belief about the grounding state of a schema instance is thus described by the beta distribution with probability density function (PDF) defined as

$$f(g;\alpha,\beta) = \frac{g^{\alpha-1}(1-g)^{\beta-1}}{B(\alpha,\beta)} \tag{2}$$

where $g \in [0;1]$ is the realisation of the random variable G representing the grounding state of a DE, $\alpha, \beta > 0$ are the shape parameters of the distribution, and $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function acting as the normalisation constant (where Γ is the gamma function defined for positive integers as $\Gamma(y) = (y-1)!$).

When the explainee reacts with positive or negative feedback to the agent's utterance, this feedback is interpreted by the system as evidence of understanding or non-understanding, respectively. This binary outcome is modelled using Bernoulli likelihood to which the beta distribution is the conjugate prior. Thus, the posterior is also a beta distribution with updated parameters

$$\alpha' = \alpha + \sum_{i=1}^{n} u_i \quad \text{and} \quad \beta' = \beta + n - \sum_{i=1}^{n} u_i \tag{3}$$

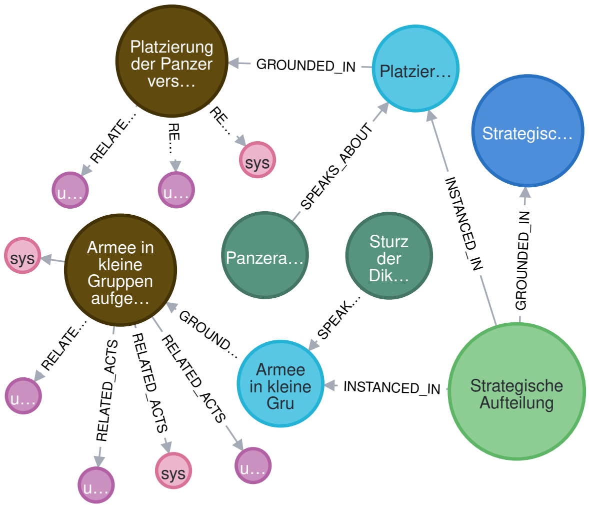

Figure 2: Part of the graph associated with the schema "strategic division" ("Strategische Aufteilung", big light green node). Following relationships and node types are represented here: `:GROUNDED_IN` as relationship between a DE and a DIS node, `:INSTANCED_IN` as relationship between schema and its instances, `:SPEAKS_ABOUT` as relationship between example and its schema instances, `:RELATED_ACTS` as relationship between a DIS and a conversational record node.

where $u \in \{0;1\}$ is the evidence of non-understanding ($u = 0$) or understanding ($u = 1$). As maximum of one instance of evidence per DE can be observed each turn, $n = 1$.

The higher level of the inference model deals with beliefs about the grounding state of generalised schemata. A new posterior for grounding state belief distribution of a schema is calculated if the distribution parameters of at least one of its instances were updated. The general update rule defined by the Bayes' theorem (equation 1) is applied here as follows. The mean value μ of the newly calculated posterior distribution $P(G|U)$ for the related schema instance is assigned to categories "low", "medium" and "high". These categories are defined in an overlapping fashion to express uncertainty within the model, for instance, μ that equals 0.45 is categorised as both "low" and "medium". The evidence of understanding is then defined by a categorical variable $\boldsymbol{u} = (u_{low}, u_{medium}, u_{high})$ where u is the number of occurrences of each category. So, for μ equals 0.45, the evidence of understanding used on the higher level of inference is $\boldsymbol{u} = (1, 1, 0)$.

The conjugate prior to the categorical likelihood is the Dirichlet distribution with PDF defined as

$$f(g_1, ..., g_K; \alpha_1, ..., \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} g_i^{\alpha_i - 1} \quad (4)$$

where $g_i \in [0;1]$ for all $i \in \{1;K\}$ and $\sum_{i=1}^{K} g_i = 1$ is the realisation of the random variable G representing the grounding state of a DE, $\boldsymbol{\alpha} > 0$ is the vector of concentration parameters of the distribution and $B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}$ is the multinomial beta function where the gamma function is expressed for positive integers in the same way as above. In the inference model, $K = 3$ for the categories "low", "medium" and "high".

Considering the definitions above, the parameter update rule for the Dirichlet distribution is

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \boldsymbol{u} \quad (5)$$

A special case of feedback regarding a schema instance can occur if the agent poses an open question to the user in order to encourage them to apply a schema with a high grounding state belief to a new example by themselves, similarly to Cawsey (1993). This is a way to obtain high-quality evidence of understanding. If the user manages to successfully generate the schema instance, the update rules for the lower level of the inference model defined in equations 3 are superseded in order to distinguish such maximising feedback from regular positive feedback such as responding with "yes" to an agent's utterance. In this case, the parameters of the distribution are directly adjusted so that the mean of the distribution lies exclusively within the "high" category. A special label is added to the corresponding DIS node in the memory graph as well, denoting that its DE was generated by the user. The update of the higher level of the inference model then proceeds normally with $\boldsymbol{u} = (-1, -1, 1)$ to increase the impact of the evidence of understanding resulting from a user-generated utterance.

Explanation planning

The planning of explanations in the architecture is also hierarchical. On the higher level of abstraction, the agent can implement different general strategies that define the principles for explanation content and dialogue move selection, while on the lower level of abstraction, it selects new content and dialogue moves for every explanation turn based on predictions of the grounding state of DEs and rules defined by the high-level strategies. High-level planning can thus be seen as an instance of the *cognitive process* of explanation generation as defined by Dazeley et al. (2021), while low-level planning belongs more to the *social process* of explanation generation, and was therefore the primary focus of research so far.

Currently, high-level explanation planning is kept constant by predefined rules. For instance, the agent always starts with examples that are most similar to the target in terms of the *structure-mapping theory*. In future research, however, it can be attempted to formalise high-level strategies as adaptable *pathways*, building up on the definition of El-Assady et al. (2019), and explore the impact of this level of adaptation in empirical studies.

Concerning low-level planning, first, the main content of the next explanation turn is determined according to principles predefined by the high-level strategy. When the agent needs a new example, it is selected based on its structural similarity to the target. It is calculated using the Jaccard similarity coefficient:

$$J(S,T) = \frac{|S \cap T|}{|S \cup T|} \tag{6}$$

where S and T are sets of analogy-relevant relationships within the source and target example, respectively. For instance, all relationships of the type `:SPEAKS_ABOUT` (figure 2).

Schema instances within an example are selected based on the high-level strategy. The memory component is hereby queried for corresponding grounding state predictions to inform the system's dialogue move selection via predictive inference.

Consider the general update principle of the grounding state belief of a DE in equation 1. This equation can be used to estimate the posterior distribution $P(G_m|U_m)$ given the most likely evidence of understanding $u*_m$ the agent would receive after a dialogue move m. The dialogue move resulting in the highest posterior distribution is selected by the explanation planner. The system currently supports two dialogue moves relating to introduction of new schema instances: "elicit generation" and "present alignment". Section 4 shows how the system chooses between these alternatives using predictive grounding state.

Determining $u*_m$ is not trivial and ideally requires a model of explainee's feedback generation. Right now, this value is defined by a set of rules for each available dialogue move. It is decided based on the category with the highest expected value in the grounding state belief distribution of the schema corresponding to the instance selected for the explanation turn. However, data of interactions with real users that will be collected in future empirical studies could be used to construct a generative model of evidence of understanding that can be used to estimate $u*_m$.

3.3 Natural language understanding

Previously, the `flexdiam` dialogue management architecture used the Rasa NLU[2] framework for intent and entity recognition. The language model employed there is based on word vectors that worked well for use cases with more structured user input where entity recognition was used primarily for slot-filling. However, in order to allow the users to answer open questions freely and use diverse expressions to refer to complex concepts and schemata, a different type of NLU component was required. This component should be capable of reformulating and summarising user utterances to obtain DEs that can be easily matched to the definitions in the agent's domain model. This kind of task is highly suitable for a pre-trained large language model (Yang et al., 2023), especially in the absence of high-quality training data.

These requirements led to a hybrid approach for NLU where intent recognition is still done with the Rasa NLU framework for a higher degree of control, while entity recognition is done with a pre-trained large language model based on the transformer architecture (Vaswani et al., 2017), namely, `text-davinci-003` from the GPT 3.5 family. Once the intent has been recognised by Rasa NLU, a prompt corresponding to the required entity recognition task is constructed. Currently, the pre-trained model is used "as-is", taking advantage of the LLMs' capabilities for few-shot learning from a small amount of handcrafted examples (Brown et al., 2020). However, the authors are preparing to evaluate the use of a smaller open-source model instead of `text-davinci-003` and are currently creating a data set for model fine-tuning.

While using an LLM can lead to unpredictable output such as hallucinations (i.e., undesirable text generation) (Ji et al., 2023), these risks were deemed acceptable, as the adaptive nature of the agent is expected to mitigate potential downstream errors caused by undesired language model out-

[2] https://rasa.com/docs/rasa/nlu-only

put through interaction, serving a function similar to repair of miscommunication in human-human interaction (Albert and de Ruiter, 2018).

4 Worked examples

This section offers more details about the use case for the agent, as well as some dialogue excerpts showcasing its behaviour in response to different types of user feedback. These are real conversations a user can have with the agent as it is implemented at the moment. Natural language generation is currently done with templates that were pre-generated using the `text-davinci-003` language model and manually edited. The possibility of using an LLM for online natural language generation is currently being evaluated. The agent converses with the user in German, however, for illustration purposes, the dialogues were translated into English by the author of the paper.

As mentioned in section 2.3, the use case chosen for the agent is based on experiments by Gick and Holyoak (1983). The user is required to find the *convergence solution* to the *Radiation* problem. The desired solution can be learned from analogous examples from other domains adapted from Gick and Holyoak (1983, Appendix II). These examples include various stories such as *Fall of the Dictatorship* (originally *The General*), in which the attacking army needed to be divided into smaller groups that converged on the fortress for a coordinated attack in order to avoid triggering the mines on the roads to the fortress.

Example interactions in the following dialogue boxes illustrate the adaptivity of the agent regarding presentation of instances of a known schema in previously unseen stories. The evidence of understanding for the schema instance in the story under discussion is used to update the probability distributions for the grounding state of the instance, as well as that of the underlying schema. These new probabilities are later used by the system to adapt dialogue moves happening potentially several turns after the evidence was provided by the user. In the example dialogues, the adaptation happens in the last turn of the agent when a new story containing a known schema is introduced to the user. For reasons of simplicity, the exchanges in the dialogue boxes happen in the system where the prior probability distributions for grounding state have just been initialised and were not changed by previous turns. For update rules and descriptions of nodes from the memory graph, see sections 3.2 and 3.1, respectively.

Without user-generated schema instance

AGENT: Can you name an important point from the story "Fall of the Dictatorship" that describes the problem in the story or its solution?

USER: No.

[*(1) Update $P(G|U)$ for all DIS nodes for schemata instances in "Fall of the Dictatorship" with $u = 0$, leading to μ' in category* ***"low"***
(2) Update $P(G|U)$ for all DIS nodes for schemata in "Fall of the Dictatorship" with $\boldsymbol{u} = (1, 0, 0)$]

AGENT: No problem, I will explain the important points of the story to you in a moment!

AGENT: One important aspect of the story "Fall of the Dictatorship" was that the big army of the general had to be divided in smaller groups, so that the mines on the road don't get triggered. Thus we see that the principle of "strategic division" was important here for the successful resolution of the problem.

USER: Okay.

[*(1) Update $P(G|U)$ for the DIS node for schema instance with schema name "strategic division" and story name "Fall of the Dictatorship" with $u = 1$ leading to μ' in categories* ***"low"*** *and* ***"medium"***
(2) Update $P(G|U)$ for the DIS node for schema called "strategic division" with $\boldsymbol{u} = (1, 1, 0)$]

[*(1) After several unrelated turns, when the next example containing "strategic division" gets presented, the DIS node for this schema has the highest expected value for $P(G|U)$ in category* ***"low"***
(2) Apply rules to determine $u*_m$ for category ***"low"****: when m is "elicit generation", $u*_m$ is expected to be negative, predicting* ***"low"*** *posterior; when m is "present alignment" $u*_m$ is expected to be positive, predicting* ***"medium"*** *posterior*]

AGENT: A relevant feature of the story "Tank Assault" was that the available tanks were divided and positioned on different bridges, so that they could cross these narrow bridges. It means that the principle of "strategic division" is important here. Similarly, in the story "Fall of the Dictatorship" the troops were divided into smaller groups, so that mines on the road don't get triggered.

When a story is introduced to the explainee, they are asked whether they can identify any "important points", i.e., schemata present in it. In response, the explainee can describe any concepts that in their opinion contributed to the solution of the problem in the story. The system then evaluates whether valid schemata instances were named and/or described and updates the probability distributions of related DEs. This results in different states of the system and different behaviour later on.

In the first example, the user does not identify any schemata. Using this negative feedback and the grounding state update rules presented previously, new posterior distributions $P(G|U)$ are calculated for the grounding state of all schemata present in the story under discussion. The system is then required to introduce a schema (e.g., *strategic division*) to the user by describing its instance in the current story. The user can give feedback to signal their level of understanding. "Okay" is interpreted as positive feedback by the system. This leads to updates of grounding state distributions on relevant memory graph nodes. Later in the dialogue, when a new story containing the schema *strategic division* is presented to the user, the system chooses the appropriate dialogue move by predicting the most likely evidence of understanding $u{*}_m$ for each of the available dialogue moves m considering the expected grounding state category of the schema and the uniform prior over the grounding state corresponding to the unseen schema instance. Here, the expected grounding state category is "low" based on previous interactions, so the system chooses the dialogue move *present alignment*, i.e., to explicitly tell the user about the instance of *strategic division* in the new story and compare it with the instance of this schema from the previous story.

In the second example, the user correctly identifies the schema *strategic division* in the current story, which is recognised by the system using the approach described in section 3.3. In this case, the grounding state distributions related to the generated schema and its instance in the story under discussion are updated using previously defined special rules for high-quality evidence of understanding. Later on, when a new story containing the schema *strategic division* is introduced, the system, as described above, chooses the appropriate dialogue move by predicting $u{*}_m$. Now, however, the expected category of the schema grounding state is "high", and so the system predicts that the dialogue move *elicit generation* will result in successful generation of the schema instance in the new story by the user, giving the system the best possible evidence of understanding. Given the uniform prior over the grounding state of the new schema instance, this would result in the best expected value of the posterior distribution, therefore, the move *elicit generation* is chosen.

With user-generated schema instance

AGENT: Can you name an important point from the story "Fall of the Dictatorship" that describes the problem in the story or its solution?

USER: Yes, it was an important point in the story that the army was divided into smaller groups so that the mines on the road don't get blown up.

[*(1) Recognise that user is speaking about the schema called "strategic division"*
*(2) Update $P(G|U)$ for the DIS node for schema instance with schema name "strategic division" and story name "Fall of the Dictatorship", so that μ' is in category **"high"***
(3) Update $P(G|U)$ for the DIS node for schema called "strategic division" with $\boldsymbol{u} = (-1, -1, 1)$]

AGENT: You have correctly identified one of the important principles of this story. It is called "strategic division". Now I will help you identify other concepts that played an important role in the solution.

[*(1) After several unrelated turns, when the next example containing "strategic division" gets presented, the DIS node for this schema has the highest expected value for $P(G|U)$ category **"high"***
*(2) Apply rules to determine $u{*}_m$ for category **"high"**: when m is "elicit generation", $u{*}_m$ is expected to be maximising, predicting **"high"** posterior; when m is "present alignment" $u{*}_m$ is expected to be positive, predicting **"medium"** posterior*]

AGENT: Can you recognise the concept of "strategic division" in the story "Tank Assault"? If so, then please tell me how it is expressed here.

5 Conclusion and future work

This paper presents a *hierarchical grounding state prediction model* that allows an explainer agent to adapt dialogue moves regarding previously unmentioned domain elements.

This is necessary in scenarios such as instructions in analogical problem-solving, as through the principle of schema induction users can learn generalised schemata and apply them to new domains autonomously. Following the principle of *scaffolding* (Rohlfing et al., 2021), the agent has to be able to predict the grounding state of relevant domain entities in order to build up on the available knowledge in the explanation process and make it more engaging and cooperative. While these are the expectations placed on the agent, the system can only be comprehensively evaluated in a series of empirical studies. Preparing for these is the next step in the project. Interaction data with real users needs to be collected in order to construct a generative model for evidence of understanding to move away from the rule-based approach currently employed for posterior prediction of grounding state after a specific dialogue move. Additionally, it would be interesting to expand the research on high-level strategies for explanation planning and investigate whether and how those could/should be adapted.

Limitations

Inference of common ground in humans incorporates complex cognitive processes the exact combination of which is not fully understood. A computational model dealing with these processes naturally features a lot of limitations. A system that strives to be co-constructive in conversations with humans also needs to be efficient and interpretable. For reasons of efficiency, the proposed grounding state prediction model uses conjugate priors for Bayesian inference. However, conjugate priors often do not capture the full complexity of real-life data and events. Only results of an empirical study can show whether they are sufficiently good for the intended application.

Another limitation relates to the interpretation of feedback. While feedback fulfils a variety of communicative functions in interaction, such as signalling contact or perception (Allwood et al., 1992), the proposed agent interprets it as evidence of understanding. However, the system should react differently to a user signalling negative perception than to a user signalling negative understanding.

Additionally, the interpretability requirement makes the use of "black box" machine learning models inside the system problematic. Tools such as LLMs are powerful and can allow the system to engage in more complex dialogues, where tasks such as text summarisation and paraphrasing are required on behalf of the system. However, the risk posed by hallucinations of the language model cannot be eliminated completely. Even though the adaptive nature of the developed agent should mitigate it, potential downstream errors might still have a negative effect on understanding, overall success in the problem-solving task or acceptance of the agent. Despite this, multiple applications of LLMs should be considered in more detail, for example, automatic generation of domain knowledge representations, including abstractions of generalised schemata and analogous concepts from text descriptions of problems, as well as generation of training data for classic NLU approaches.

Ethics Statement

This research does not have any particular ethical implications, apart from general considerations on the use of artificial agents in dialogue interactions with humans. The authors are aware of risks associated with the use of large language models, however, some of them such as generation of potentially harmful or misleading content are contained through downstream processing in the architecture, as the output of the language model is not directly present in the utterances of the agent.

This article does not contain any studies involving animal or human participants performed by any of the authors. The authors also declare no conflict of interest.

Acknowledgements

This work was mostly done during the project IMPACT[3] (*The Implications of Conversing with Intelligent Machines in Everyday Life for People's Beliefs about Algorithms, Their Communication Behaviour and Their Relationship-Building*) which was funded by the Volkswagen Foundation and ended in March 2023. It is now continued in the project SAIL[4] (*Sustainable Life-Cycle of Intelligent Socio-Technical Systems*) that is funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia under the grant number NW21-059A.

References

Saul Albert and J. P. de Ruiter. 2018. Repair: The Interface Between Interaction and Cognition. *Topics in Cognitive Science*, 10(2):279–313.

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1):1–26.

Agnes Axelsson and Gabriel Skantze. 2023. Do You Follow?: A Fully Automated System for Adaptive Robot Presenters. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 102–111, Stockholm, Sweden.

[3]https://www.impact-projekt.de/
[4]https://www.sail.nrw/

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada.

Sarah Brown-Schmidt. 2012. Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, 27(1):62–89.

Hendrik Buschmeier and Stefan Kopp. 2012. Using a Bayesian Model of the Listener to Unveil the Dialogue Information State. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2012)*, pages 12–20, Paris, France.

Hendrik Buschmeier and Stefan Kopp. 2018. Communicative Listener Feedback in Human–Agent Interaction: Artificial Speakers Need to Be Attentive and Adaptive. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, pages 1213–1221, Stockholm, Sweden.

Alison Cawsey. 1993. User modelling in interactive explanations. *User Modeling and User-Adapted Interaction*, 3:221–247.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, MA, USA.

Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. 2021. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299.

Maria Di Maro, Antonio Origlia, and Francesco Cutugno. 2021. Cutting melted butter? Common Ground inconsistencies management in dialogue systems using graph databases. *Italian Journal of Computational Linguistics*, 7(1 | 2):157–190.

Karl Duncker. 1945. On problem-solving. *Psychological Monographs*, 58(5):1–113.

Mennatallah El-Assady, Wolfgang Jentner, Rebecca Kehlbeck, Udo Schlegel, Rita Sevastjanova, Fabian Sperrle, Thilo Spinner, and Daniel Keim. 2019. Towards XAI: Structuring the Processes of Explanations. In *Proceedings of Human-Centered Machine Learning Perspectives Workshop at CHI'19*, Glasgow, Scotland, UK.

Chris D. Frith and Uta Frith. 2006. How we predict what other people are going to do. *Brain Research*, 1079(1):36–46.

Dedre Gentner. 1983. Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7(2):155–170.

Dedre Gentner, Jeffrey Loewenstein, and Leigh Thompson. 2003. Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, 95(2):393–408.

Dedre Gentner and Francisco Maravilla. 2017. Analogical Reasoning. In L. J. Ball and V. A. Thompson, editors, *International Handbook of Thinking & Reasoning*, pages 186–203. Psychology Press, New York, NY.

Mary L. Gick and Keith J. Holyoak. 1983. Schema induction and analogical transfer. *Cognitive Psychology*, 15(1):1–38.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12).

Stefan Kopp, Jens Allwood, Karl Grammer, Elisabeth Ahlsen, and Thorsten Stocksmeier. 2008. Modeling Embodied Feedback with Virtual Humans. In I. Wachsmuth and G. Knoblich, editors, *Modeling Communication with Robots and Virtual Humans*, volume 4930 of *Lecture Notes in Computer Science (LNAI)*, pages 18–37. Springer Berlin Heidelberg.

Stefan Kopp and Nicole Krämer. 2021. Revisiting Human-Agent Communication: The Importance of Joint Co-construction and Understanding Mental States. *Frontiers in Psychology*, 12.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective.

Ben Lambert. 2018. *A Student's Guide to Bayesian Statistics*. SAGE Publications, Inc., London, United Kingdom.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in Human-AI Systems: A Literature Meta-Review. Synopsis of Key Ideas and Publications and Bibliography for Explainable AI.

Tim Paek and Eric Horvitz. 2000. Conversation as Action Under Uncertainty. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI'00)*, pages 455–464, Stanford, CA, USA.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Amelie Robrecht and Stefan Kopp. 2023. SNAPE: A Sequential Non-Stationary Decision Process Model for Adaptive Explanation Generation. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, pages 48–58, Lisbon, Portugal.

Katharina J. Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M. Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Hab-Umbach, Ilona Horwath, Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. 2021. Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13(3):717–728.

Maayan Shvo, Toryn Q. Klassen, and Sheila A. McIlraith. 2020. Towards the Role of Theory of Mind in Explanation. In *Proceedings of EXTRAAMAS 2020: Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, volume 12175 of *Lecture Notes in Computer Science (LNAI)*, pages 75–93, Online. Springer International Publishing.

Kacper Sokol and Peter Flach. 2020. One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. *KI - Künstliche Intelligenz*, 34:235–250.

Matthew Stone and Alex Lascarides. 2010. Coherence and Rationality in Grounding. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2010)*, Poznań, Poland.

Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan.

Richmond H Thomason. 2003. Accommodation, Meaning, and Implicature: Interdisciplinary Foundations for Pragmatics. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 325–363. MIT Press, Cambridge, MA, USA.

David Traum and Staffan Larsson. 2003. The Information State Approach to Dialogue Management. In J. van Kuppevelt and R. W. Smith, editors, *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology (TLTB)*, pages 325–353. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need.

Ramin Yaghoubzadeh and Stefan Kopp. 2017. Enabling robust and fluid spoken dialogue with cognitively impaired users. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 273–283, Saarbrücken, Germany.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond.