Knowledge Modelling for Establishment of Common Ground in Dialogue Systems (with debate)

Lina Varonina*
Bielefeld University

Stefan Kopp**
Bielefeld University

The establishment and maintenance of common ground, i.e. mutual knowledge, beliefs and assumptions, is important for dialogue systems in order to be seen as valid interlocutors in both task-oriented and open-domain dialogue. It is therefore important to provide these systems with knowledge models, so that their conversations could be grounded in the knowledge about the relevant domain. Additionally, in order to facilitate understanding, dialogue systems should be able to track the knowledge about the beliefs of the user and the level of their knowledgeability, e.g., the assumptions that they hold or the extent to which a piece of knowledge has been accepted by the user and can now be considered shared. This article provides a basic overview of current research on knowledge modelling for the establishment of common ground in dialogue systems. The presented body of research is structured along three types of knowledge that can be integrated into the system: (1) factual knowledge about the world, (2) personalised knowledge about the user, (3) knowledge about user's knowledge and beliefs. Additionally, this article discusses the presented body of research with regards to its relevance for the current state-of-the-art dialogue systems and several ideal application scenarios that future research on knowledge modelling for common ground establishment could aim for.

1. Introduction: Why do we need to model knowledge in dialogue systems?

When speaking about engaging in conversations with machines, most people would refer to interactions with proprietary voice-based assistants (VA), such as Amazon Alexa, Google Assistant, Siri, etc. Since their introduction to the market in the previous decade, these systems have consistently been on the rise. According to a survey conducted in the U.S. in 2020, more than a third of the adult population of the country possesses a smart speaker (Kinsella 2020). Therefore, for many people experiences with these voice-based assistants will influence their perception and expectations with regards to language-based interactions with machines. However, despite recent advances in natural language processing (NLP) the capabilities of these wide-spread VAs to lead human-like conversations are rather limited, resulting in a mismatch between expectations and reality, which is especially prevalent among users with little technical knowledge who cannot form adequate judgements about the capabilities of the system and rely on their experiences from human-human communication when interacting with VAs (Luger and Sellen 2016). Failure to engage with a voice-based assistant in a meaningful way can cause users to change their communicative behaviour, e.g., by limiting their vocabulary

* Social Cognitive Systems Group - Inspiration 1, 33619 Bielefeld, Germany.
E-mail: lvaronina@techfak.uni-bielefeld.de
** Social Cognitive Systems Group - Inspiration 1, 33619 Bielefeld, Germany.
E-mail: skopp@techfak.uni-bielefeld.de

and simplifying utterances or reducing the interactions with the system to a range of simple tasks that the users trust the system to perform correctly (Luger and Sellen 2016). The implications of such communicative failures and lessons that can be learnt from them for the design of conversational agents is one of the research topics of the project *IMPACT*[1] (The implications of conversing with intelligent machines in everyday life for people's beliefs about algorithms, their communication behaviour and their relationship-building), a cooperation of various universities and disciplines that the authors of this paper are a part of.

It is interesting to note that some researchers reject the notion of classifying interactions with VAs as *conversations* due to their fundamental differences with *actual human conversations*. Porcheron and colleagues (Porcheron et al. 2018) discussed this idea in the context of the findings of their study on everyday use of voice-based assistants in families. For instance, they argue that the predefined request-response format of the interaction with VAs cannot be equated with interactively emerging adjacency pairs, such as question-response, that serve as the basic organisational unit of many of our everyday conversations. The responses of voice-based assistants sometimes fail to coherently follow the requests of their users, which is usually treated by the users as incorrect output, rather than a reaction of an equal conversation partner. Overall, the findings of the study suggest that smart devices with voice-based assistants are not treated as interlocutors by family members, even though the interactions with them are embedded in conversational situations within a family.

These differences in treatment were also seen in open-question interviews about the nature of conversations conducted by Clark and colleagues (Clark et al. 2019). While the interviewees acknowledged the importance of similar concepts in both human-human and human-agent conversations, they operationalised them differently, as conversations with humans were characterised to have both social and transactional purposes, but descriptions of conversations with agents (which were influenced by interviewees' experiences with voice-based assistants) were mainly focused on the transactional aspect. So, for example, establishing *common ground* was identified as one of the most important parts of a good conversation with other humans. However, in a human-agent setting the interviewees rather preferred to speak about *personalisation* where certain information is used by the system to tailor user experience, which, in a long-term perspective, could create an illusion of common ground between the human and the machine. The interviewees also did not view this process as co-constructed as they would the establishment of common ground in human-human communication.

While it is not necessary to strive for the ideal of "human-like" conversation in every domain where conversational systems are used, in certain use cases, it is necessary to endow these systems with qualities that would allow them to be increasingly treated as valid conversation partners by humans (cf. (Kopp and Krämer 2021)). On the one hand, these may be the use cases in which the social aspect of the conversation is of importance, e.g., in social care or robot companionship. On the other hand, even in task-oriented, primarily transactional interactions the inclusion of certain aspects of human-human conversation is needed: the one-shot request-response format of interaction currently provided by the voice-based assistants is not sufficient. Clark and colleagues (Clark et al. 2019) offer an apt goal for task-oriented conversational systems: service desk interactions between humans. In these types of conversations, the concepts of

1 https://www.impact-projekt.de/

common ground and facilitation of understanding for all conversation partners become crucial for successful accomplishment of tasks.

According to the definition of *common ground* as established by Clark and Brennan, it entails "mutual knowledge, mutual beliefs and mutual assumptions" (Clark and Brennan 1991, p. 222). Thus, modelling these categories in a conversational system is a prerequisite for its capability to establish and maintain common ground. Various types of knowledge can be of relevance here, e.g., knowledge about the domain, but also knowledge about the user and their beliefs, their level of expertise in the domain, known facts and possible preconceptions, as well as the ability to track how these change as the conversation progresses, what kind of knowledge becomes *grounded* and can be used for future reference.

The goal of this article is to provide an overview of current research on knowledge modelling for the establishment of common ground in dialogue systems. Roughly, it is possible to divide this body of research into three major categories based on the type of knowledge integrated into the system. Each of these has its own research focus and use cases. These categories will be discussed in the following order:

1. factual knowledge about the world,
2. personalised knowledge about the user,
3. knowledge about user's knowledge and beliefs.

This article is by no means a comprehensive collection of work on these topics, but strives to provide a basic overview of the directions state-of-the-art research takes. Additionally, the attention currently devoted to the aforementioned topics within the research community will be discussed, along with the perspectives for and the impact of the realisation of common ground on future dialogue systems.

2. Factual knowledge about the world

The first category of knowledge that can be integrated into dialogue systems is factual knowledge about the world and the elements of the so-called commonsense knowledge, e.g., information such as *"A dog has four legs"* (Zhou et al. 2018).

The emergence of data-driven neural language models in the field of machine translation inspired the creation of end-to-end dialogue systems where similar approaches could be used, which offered an alternative to the traditional multi-component dialogue systems with separate modules for natural language understanding, generation and synthesis and dialogue management (Ritter, Cherry, and Dolan 2011; Sordoni et al. 2015; Serban et al. 2016; Gao, Galley, and Li 2019). These new systems, of course, had their own challenges, such as uninformativeness and the lack of diversity of utterances generated, which was addressed in different areas of research. Amongst them an idea was born to introduce *knowledge-based grounding* to neural conversational systems in order to make their responses more diverse, specific and "human-like" (Han et al. 2015; Yin et al. 2016; Zhu et al. 2017; Ghazvininejad et al. 2018; Zhou et al. 2018). This type of grounding allows the dialogue system to talk about entities not seen in the training data and also reflect changes in the domain within their responses through updates of the knowledge base (Gao, Galley, and Li 2019). *Knowledge-aware* dialogue systems are applied in both open-domain as well as task-oriented dialogue.

In such systems, external collections of knowledge are usually used. These can have varying representations, e.g., as textual data or structured knowledge bases or knowl-

edge graphs. Examples of the textual data approach can be found in (Ghazvininejad et al. 2018) where the researchers used data from social networks such as Twitter and Foursquare indexed by relevant entities, or in (Dinan et al. 2019) where Wikipedia articles organised as documents structured into paragraphs and sentences were utilised. When it comes to structured knowledge bases (Han et al. 2015; Yin et al. 2016; Zhu et al. 2017; Zhou et al. 2018; Zhang et al. 2020; Wu et al. 2020), the researchers usually use large knowledge graphs that are well-established and have been maintained by the Semantic Web community throughout the years, such as the multi-language common sense knowledge graph *ConceptNet*[2] (Speer, Chin, and Havasi 2017) or *DBpedia*[3] (Lehmann et al. 2012) that represents the information created in Wikipedia and other Wikimedia projects. Another advantage of these graphs is that they can also be connected with each other to leverage knowledge about terms and concepts across domains as part of the *Linked Data* standard (Berners-Lee 2006). The relations in such knowledge bases are typically represented by subject-predicate-object `(s, p, o)` triples, e.g., the piece of information *"a puppy can become a dog"* is represented in *ConceptNet* as a triple (`/c/en/puppy`, `/r/CapableOf`, `/c/en/become_dog`) where `/c/en/` and `/r/` are graph-specific namespaces used for distinction of identifiers.

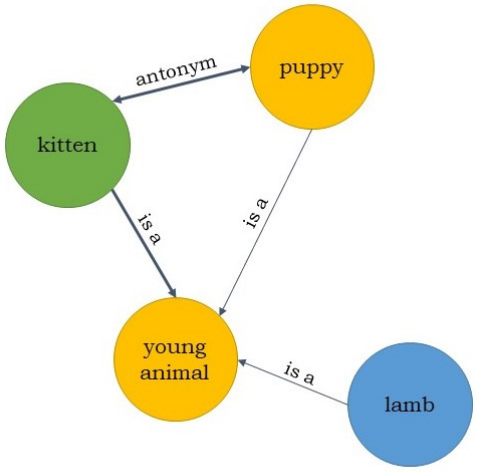

Figure 1
Part of a concept graph as defined in (Zhang et al. 2020). It is centered around the concept of `kitten` and based on relations from ConceptNet. Yellow-coloured nodes are one-hop concepts and the blue-coloured node is a two-hop concept.

Overall, the research on grounding of dialogues in factual or commonsense knowledge is primarily concerned with the inclusion of already available factual and commonsense knowledge bases into conversational models, i.e., selecting and extracting the knowledge relevant to the entities mentioned in user utterances and encoding this

2 `http://conceptnet.io/`
3 `https://www.dbpedia.org/`

knowledge and leveraging it for response generation, not knowledge modelling in itself. These topics are outside of the scope of this paper and will not be expanded upon here. However, it is noteworthy that some approaches to knowledge integration go beyond static entity matching and acknowledge the fact that humans reference related concepts in conversations and usually shift their focus to different topics as the exchange progresses, which is modelled as *attentional state* by Grosz and Sidner (Grosz and Sidner 1986). Thus, methods are researched that would enable dialogue systems to introduce new related concepts into conversation. Consider a dialogue about *kittens* that may develop into a conversation about other *young animals*, such as *puppies* or perhaps even *lambs* if the people speaking live in the countryside. In *ConceptNet*, `young animal` and `puppy` are one-hop concepts and `lamb` is a two-hop concept with regards to `kitten` as illustrated in figure 1. It is possible to include this type of concept shift into conversational models with approaches such as the conversation generation model *ConceptFlow* that constructs a so-called *concept graph* which is a local part of the main knowledge graph centered around the currently grounded concept of the conversation (`kitten`) and extended to its one-hop (`young animal` and `puppy`) and two-hop concepts (`lamb`). This concept graph is later used for conversation modelling and response generation (Zhang et al. 2020).

3. Personalised knowledge about the user

As mentioned in the introduction, according to interviews conducted by Clark and colleagues (Clark et al. 2019), some people reject the notion of having *common ground* with machines and prefer to speak of *personalisation* in the context of human-agent conversations, i.e. the adaptation of user experience based on the information about the user collected by the system, which can create an illusion of common ground over time. However, important differences exist between the concepts of *user adaptation* or *personalisation* of dialogue systems and the establishment of *common ground* as defined by Clark and Brennan (Clark and Brennan 1991).

First, the concept of personalisation is very broad. In their survey on empathetic systems, Ma and colleagues (Ma et al. 2020) distinguish between two types of dialogue systems with regards to personalisation: *personality-aware* and *personality-infused*. While the former type only considers the personality (or certain distinct features thereof) of the *user* when composing responses, the latter type additionally infuses the *agent* with its own personality. Personality-infused systems are out of scope of this paper and will not be discussed further.

Second, personalisation is not co-constructed, as it is the system that is burdened with the collection of information about the user. Nevertheless, this collected information can be applied in the context of ensuring mutual understanding between the user and the dialogue system, e.g., by allowing the system to better understand user's intentions and react appropriately, which would be consistent with the way information exchanged for the establishment of common ground is used.

According to the aforementioned survey of Ma and colleagues (Ma et al. 2020), there are two categories of methods that can be applied to user modelling in personalised dialogue systems: *identity-based* and *knowledge-based*, while some hybrid systems also exist. Identity-based systems model the user via a set of attributes that define their basic characteristics, e.g., gender, age group, profession. The required attributes vary based on the interaction context and are oftentimes collected during the first interaction with the user. On the other hand, knowledge-based personalisation uses structured knowledge bases with facts about the user, mostly represented by subject-predicate-object triples,

and can be seen as a special case of knowledge-aware dialogue systems as described in the previous chapter. For both of these approaches, unstructured data from past interactions can also be leveraged to extract either attribute values for the identity-based models or facts to be placed in the knowledge base for knowledge-based models.

3.1 Identity-based systems

Apart from profile-building via "get-to-know" sessions during the first interaction or analysis of previously acquired interaction data, it is also possible for a dialogue system to utilise profiles of similar users in order to make assumptions about the current user. This could be beneficial for situations when the system has not yet had many interactions with the user or the profile required for personalisation is too extensive to be explicitly requested. Pei and colleagues (Pei, Ren, and de Rijke 2021), for example, propose an architecture called *Cooperative Memory Network* for this purpose, a part of which is a user profile enrichment module which maintains the profile and the dialogue memory that are represented as embeddings of the profiles of the current user and users similar to them and their dialogue history respectively. Individual user profiles are represented as numerical vectors and the utterances in dialogue history are represented as a bag-of-words. Missing values in the current profile are then inferred based on these embeddings and memory components get updated. These enriched profiles are then used to update the representation of the current user query and find the appropriate response.

Before that, Luo and colleagues (Luo et al. 2019) proposed another memory-based architecture called *Personalized MemN2N* for task-oriented dialogue systems. This architecture also leverages conversational data embeddings from similar users along with the current user profile in order to generate personalised response candidates. Of special interest here is that the researchers also use profile information to infer user preferences over entities in a knowledge base that contains facts about the task domain, e.g., whether the user would like to contact the restaurant they want to eat at via phone or social media.

3.2 Knowledge-based systems

In their position paper, Balog and Kenter (Balog and Kenter 2019) define the concept of the so-called *personal knowledge graph* (PKG), as opposed to publicly available knowledge graphs such as *DBpedia* that include knowledge about entities that are publicly significant. Despite the fact that various researchers have previously used concepts similar to a *personal knowledge base* (PKB) or graph (Kim et al. 2014; Li et al. 2014; Bang et al. 2015), Balog and Kenter (Balog and Kenter 2019) establish the key properties of PKGs and identify important research questions with regards to these. According to them, three key aspects of a PKG are:

1. inclusion of entities that are of personal interest to the user,
2. the "spiderweb" layout centered around the user,
3. possible integration with other knowledge graphs as part of the *Linked Data* idea.

The population and maintenance of these personal knowledge graphs should occur automatically, as no designated human editors exist to curate the graphs. The authors

of the article present this as one of the challenges and research questions to be explored: how to transfer the data-driven state-of-the-art neural approaches to link prediction to PKGs for which the availability of data is very limited (Balog and Kenter 2019). Previously, other approaches to personal knowledge graph population were suggested, such as the combination of support vector machines (SVM) and conditional random fields (CRF) for the classification of personal facts in dialogue data, relation extraction and subsequent slot filling to complete user-related triples that are then added to the PKG (Li et al. 2014). However, when it comes to conversational data as information source, the assertions that need to be captured in the knowledge base are rarely stated explicitly (Tigunova et al. 2019). Instead, a person who works as a teacher might often talk about school, grades and homework without explicitly saying that they are a teacher. In (Tigunova et al. 2019) a neural architecture called *Hidden Attribute Model* is presented that is trained on triples to predict scores for different objects that could complete a given subject-predicate pair by using attention both within and across user utterances, e.g., it could predict the scores for different professions `X` to complete the triple `(user, employedAs, X)`.

With regards to maintenance of personal knowledge graphs, it needs to be taken into consideration that PKGs are inherently more dynamic than general-purpose knowledge graphs that store information about the world and place value on established assertions that will unlikely change fast (Balog and Kenter 2019) (consider the Wikipedia-based *DBpedia* graph and the dynamics of knowledge that you can find on Wikipedia as opposed to how fast your own preferences, possessions, etc. change). To model these temporal dynamics when it comes to user-related knowledge, Kim and colleagues (Kim et al. 2014) integrate a personal knowledge base with a *forgetting model* endowing their dialogue system with a long-term memory about the user. Each entry in the PKB has two properties: retention, which models the degree of user interest in this fact, and strength, which prevents the retention value from decaying too quickly. Both of these values change over time depending on the occurrences of the respective entity in user utterances. The forgetting model used by Kim and colleagues applies Ebbinghaus's forgetting curve and spacing effect (Kim et al. 2014; Ebbinghaus 2011).

Another interesting idea that utilises a knowledge-based approach, yet concerns itself not with personalised response generation but rather with a memory-based personal question answering, is proposed in (Moon et al. 2019). In their paper, the authors represent episodic memories concerning the user, such as events they attended, as a *memory graph* consisting of the entities related to a memory connected by corresponding edges. The entities are nodes of a knowledge graph that models the related domain knowledge. An example of such a *memory graph* can be seen in figure 2. Consider that the user knows that they have once eaten at a venue in the city district Bielefeld-Mitte (the centre of the city of Bielefeld), but they do not remember in what year it was. So they can query the system with the question *When have I been to a venue in Bielefeld-Mitte?* and the system can use the proposed approach of *Memory Graph Networks* to expand memory slots with external knowledge via attention-based memory graph traversal. That way, it can eventually obtain the result that the restaurant the user has been to in 2020 for Mary's birthday is in fact located in the desired city district. The authors also mention the possibility of memory graph extraction from social media posts and tagged photo albums of a particular user.

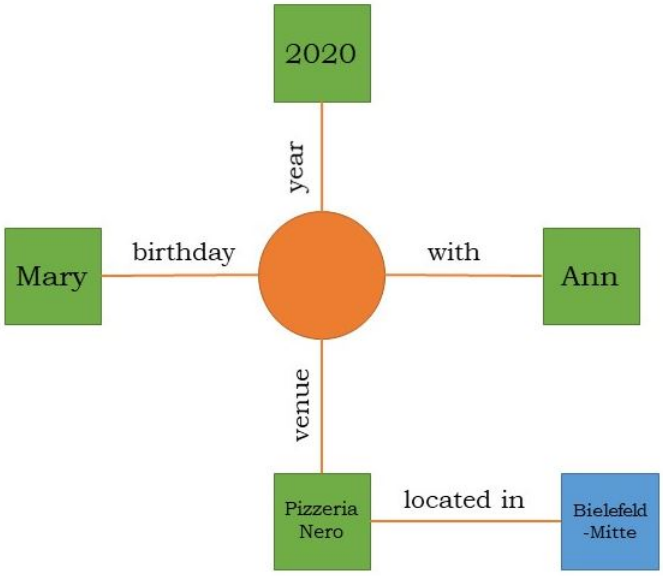

Figure 2
An episodic memory in a memory graph as defined in (Moon et al. 2019). The orange circle is the memory slot and the green squares are knowledge graph entities that are related to the memory, i.e. the birthday party of Mary in 2020 at Pizzeria Nero where the user went with their friend Ann. The blue square is the knowledge graph node that was activated after the expansion of the graph with regards to the user query *When have I been to a venue in Bielefeld-Mitte?*

4. Knowledge about user's knowledge, beliefs and mental states

The definition of *common ground* by Clark and Brennan that was mentioned throughout this paper refers to "mutual knowledge, mutual beliefs and mutual assumptions" (Clark and Brennan 1991, p. 222). However, how do the conversation partners know that they possess mutual knowledge or mutual beliefs between each other?

In an attempt to clarify the existing definitions with regards to *common ground*, Lee (Lee 2001) uses the terms of *common* and *shared knowledge* and *belief*. The author defines the concept of *common* as the information that people assume to have in common with others because of their similar background of up-bringing and the concept of *shared* as the information negotiated during a mutual interaction, while the difference between *knowledge* and *belief* lies in the certainty of truth of the information as perceived by the person. According to these definitions, in order to understand what kinds of knowledge or beliefs are common between conversation partners, they have to make assumptions about (1) the other person's background and (2) the extent to which they have understood or remembered the negotiated information. Both of these are arguably made possible by the so-called *theory of mind* (ToM).

4.1 Modelling knowledge about beliefs for ToM in human-agent interaction

One could define a theory of mind as "a basic cognitive and social characteristic that enables us to make conjectures about each others' minds through observable or latent behavioural and verbal cues" (Wang et al. 2021, p. 2). These conjectures allow humans to act accordingly in order to lead successful conversations and collaborations with others. The concept of theory of mind was also adapted for the design of human-agent interactions (Krämer, Rosenthal-von der Pütten, and Eimler 2012), primarily in the area of robotics and task-oriented collaboration (Wang et al. 2021; Scassellati 2002; Peters 2005; Devin and Alami 2016; Dissing and Bolander 2020), as perception and sensory-motor expression such as gestures are a part of the framework of ToM (Baron-Cohen 1995). Studies show that implementing ToM in robots leads to positive effects such as reduction of unnecessary communication during collaborative tasks (Devin and Alami 2016) or the perception of robots as more intelligent and natural in interaction (Hiatt, Harrison, and Trafton 2011).

As voice-based assistants fail in dialogues beyond one-shot interactions, there is a growing need and motivation to adapt aspects of the ToM concept for conversational assistants (Wang et al. 2021; Kopp and Krämer 2021). Existing neural models for question answering do not succeed at false-belief tasks, such as the classic *Sally-Anne-Experiment* (Baron-Cohen, Leslie, and Frith 1985), as was shown in an article by Nematzadeh and colleagues (Nematzadeh et al. 2018), where the researchers created a dataset of tasks that can be used for the evaluation of question answering neural models (such as memory networks, the examples of which were shown in chapter 3.1) with regards to belief reasoning. They tested several of such models and found that they make reasoning mistakes in false-belief tasks due to not having the ability to track mental states of agents that are inconsistent with the state of the real world. This might be a potential motivation to develop models that can explicitly incorporate theory of mind in conversational contexts.

Different approaches to the implementation of theory of mind in artificial agents exist and a brief overview will be given in this subsection. In general, one could divide the existing approaches into three groups:

1. models based on logic and symbolic reasoning,
2. probabilistic models, and
3. models based on machine learning.

With regards to the first group, one example could be the work of Devin and Alami (Devin and Alami 2016) that deals with the execution of shared plans in human-robot teams. Their proposed architecture features a ToM manager that maintains the mental state of the robot and other agents. The mental state is defined as (1) a set of facts about the current world state, (2) the state of current goals, (3) the state of plans, and (4) the state of actions, all of these from the respective agent's perspective. The states 2-4 denote, e.g., whether the current goal is achieved or whether a certain action has already been requested. The ToM manager then can utilise a symbolic reasoning process to make assumptions from and update the mental states of the agents.

The approach of Dissing and Bolander (Dissing and Bolander 2020) advocates for the usage of *dynamic epistemic logic* (DEL) (Bolander 2018) for theory of mind models in order to facilitate higher-order belief attribution, i.e. beliefs of agents about other agents' beliefs, as opposed to first-order belief attribution, i.e. beliefs of agents about the state

of the world. Their system maintains an epistemic state consisting of a representation of the actual world and an epistemic model over a set of possible worlds. This epistemic state is updated based on a set of rules when new actions take place in the context and can be queried with regards to a belief by using speech. The robot endowed with this approach was able to successfully pass first- and second-order false-belief tasks in an experimental setting. However, the authors state some of the limitations of their approach, e.g., the assumption that the robot is considered omniscient and cannot have false beliefs on its own or their model not accounting for agent intentions, which would be an important aspect while applying theory of mind to conversational scenarios.

In terms of probabilistic models, the most prominent one is arguably the *Bayesian theory of mind* (BToM). This approach views mentalising about the mental state of the other as Bayesian inference of the agent's hidden mental state given their behaviour in a specific context (Baker et al. 2017). The candidate mental states are defined by the agent's beliefs and desires. The beliefs are hereby represented as a probability distribution over world states in all possible worlds and their update is modelled as rational Bayesian state estimates given what was perceived by the agent and their prior beliefs. The agent's desires are represented by a utility function over situations and possible actions. The BToM adds a prior over these candidate mental states in form of a probability distribution.

BToM models can get very complex depending on the scenario they are used in. To integrate the Bayesian ToM into social agents acting in contexts when quick reaction is of importance, e.g., conversational situations, it could be beneficial to simplify these models in a way that would retain sufficient accuracy, while producing reasonable costs. In their work, Pöppel and Kopp (Pöppel and Kopp 2018) investigate the potential to simplify BToM models based on various sets of assumptions about uncertainties the acting agent faces in the environment. This results in specialised models matching a specific type of uncertainty. However, they also propose a combination model capable of switching between these specialised models according to the metric of surprise which describes how well the current model explains the behaviour the agent is observing. The authors have tested their approach, comparing the simplified models, the full BToM model and the combination model by applying them to inferences over human behavioural data in situations with various degrees of uncertainty. This data was collected by letting participants complete a set of maze traversal tasks in different uncertainty conditions, e.g., uncertainty about the structure of the maze. The results show that simplified specialised models have the ability to perform both well and badly depending on the condition they were applied to, thus leading to the necessity of the flexible combination model that achieved best performance across conditions, and, importantly, in a short enough time to facilitate online behaviour evaluation, unlike the full BToM model (Pöppel and Kopp 2018).

Lastly, machine learning methods started being involved in the implementation of ToM in agents in recent time to forgo the explicit modelling of mental states and beliefs. A prominent work here is the concept of *machine theory of mind* pioneered in the article by Rabinowitz and colleagues (Rabinowitz et al. 2018) who consider the construction of a theory of mind as a meta-learning problem. Here, in a sequence of training episodes an observer gets a set of behavioural data for a novel agent in order to make predictions about their future actions. As training progresses, the observer should learn to make better predictions about new agents from the limited set of data it receives. The architecture proposed for the observer contains three neural networks: a character net, a mental state net and a prediction net. The character net parses the historical behavioural data of the agent into a character embedding, while the mental

state net creates an embedding of their mental state based on the agent's behavioural data from the current episode. Both embeddings are then given to the prediction net to form predictions over possible next steps of the agent.

In many papers in this subsection, when it comes to theory of mind, agents usually exist in the real physical world and can observe this world and the actions carried out in it in order to update the state of the world and the mental states of others. Alternatively, it can also be an artificial world that is analogous to the real world by virtue of having specific rules and laws, and the agents in the scenario at hand act within the confines of this world. However, when it comes to social interaction, it might not be enough to update mental states based on explicit actions of others in the world. People can change their mental state because of dialogues they have with others and it is important for conversational agents to be able to capture that as well (Kopp and Krämer 2021).

Qiu and colleagues (Qiu et al. 2021) have recently introduced a *hybrid mental state parser* that can transform both continuous dialogue data and discrete action data into a graphical representation of agent's beliefs about their environment and other agents in it. Their work is based on the research of Adhikari and colleagues (Adhikari et al. 2020) who developed a graph-aided transformer agent that is capable of learning to construct and update a graph representing their beliefs about the environment of a text-based game in an end-to-end fashion from textual data by using a combination of reinforcement and self-supervised learning. Inspired by this approach and aiming to design a method that can construct belief representations from dialogue data, Qiu and colleagues (Qiu et al. 2021) also situate their agent in a text-based game (however, this type of game additionally allows dialogues between players) and apply a graph-based representation of agent's beliefs in their system. In the *belief graph*, all agents and objects along with their descriptions are represented as nodes and the belief of the agent about the current state of the environment is represented in edges that define relations between the entities and can have varying strengths. The vocabulary of entities and relation types is known in this domain by virtue of it being a game. The topology of the graph is, however, unknown and needs to be learned by the agent. It is updated as new actions and dialogue history are observed. Discrete actions carried out in the game, e.g., *put* or *give*, can be mapped onto combinations of graph update operations to add or remove specific edges in the graph. Meanwhile, continuous dialogue data is used to update the graph via a recurrent neural network.

4.2 Modelling knowledge about knowledge and beliefs in dialogue systems

One important area for belief modelling in dialogue systems are argumentative dialogues, as accounting for the perspectives of those engaged in an argument is crucial here. Additionally, a lot of uncertainty exists in this type of dialogue: with regards to beliefs of your conversation partner, the completeness of information known to them and the extent of their rationality, as well as with regards to the strength of own arguments and their influence on the beliefs of the other. Hunter and colleagues (Hunter, Polberg, and Thimm 2020) aim to create a new formalism for argumentation dialogues and reasoning that could provide solutions to these challenges: the *epistemic graph*. They describe an epistemic language that can be used to define logical formulae to specify belief in arguments and relations between them given a directed argument graph, e.g., as seen in figure 3.

The beliefs of the agent are represented with probabilities: an agent believes a term (a propositional formula of an argument) to some degree if its probability is higher than 0.5, disbelieves it to some degree if its probability is lower than 0.5 and neither believes

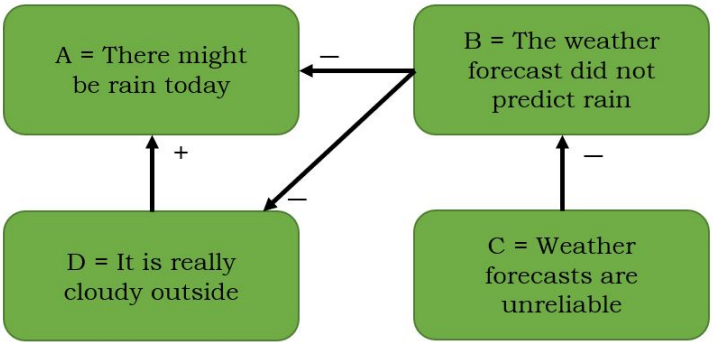

Figure 3
An example of an argument graph. Edges labelled with - and + represent attack and support respectively.

nor disbelieves it if its probability is equal to 0.5. These belief probabilities can later be used to form constraints that reflect complex beliefs, perspectives and choices. These constraints can then be reasoned with based on the logical framework developed and proved by the authors (Hunter, Polberg, and Thimm 2020). Many directions for future work are also proposed, oriented towards a practical application of epistemic graphs, for example, in computational persuasion, amongst them collection of constraints in a data-driven fashion by applying machine learning methods to crowd-sourced data on beliefs in arguments, or development of methods for belief updates during dialogues.

Another graph-based model of reasoning are *Bayesian networks*, where every node is a random variable representing a proposition and edges express statistical dependencies of one variable on another. These influences can change the belief in the target node either in a positive or a negative way, which makes Bayesian networks similar to epistemic graphs.

Bayesian networks are also used in dialogue systems when it comes to mental state representation. Buschmeier and Kopp (Buschmeier and Kopp 2011) describe the so-called *attributed listener state* (ALS) that is the assumption the speaker forms over the mental state of the listener with regards to basic communicative functions according to listener's communicative feedback the speaker receives. For example, from the listener's feedback the speaker can infer their level of understanding and form a belief about it or, more specifically, a belief about the listener's perception of their own current level of understanding. However, as this mental state attribution process is subject to uncertainty, it is necessary to understand the speaker's belief about listener's mental states in terms of their subjective *degree of belief*, i.e. the subjective confidence that this belief holds true at a given point in time, which is modelled as a probability. From this, the speaker's belief state about the listener's mental state can be defined in terms of their degree of belief in all possible worlds (Buschmeier and Kopp 2012).

Overall, the attributed listener state is modelled as a set of five discrete random variables representing the graded beliefs of the speaker about five aspects of the listener's mental state, namely, (1) them being in contact with the speaker, (2) them being able to perceive, (3) understand, (4) accept and (5) agree with the speaker. The interactions between these random variables in the ALS could be expressed with a joint

probability distribution, however, due to independence assertions for these variables it is possible to represent them in terms of five conditional distributions which is a much simpler representation that can also be expressed in terms of a graphical probabilistic model which would allow reasoning with the resulting data structures: the *Bayesian network* (Buschmeier and Kopp 2012). In fact, as seen in figure 4, *attributed listener state* is a sub-network of the larger Bayesian network of the listener where it mediates between the *conversational context* and the *information state* of the dialogue. The conversational context consists of fully observable variables, some of which are inferred from listener's feedback, such as modality, and abstract concepts, such as difficulty of the speaker's utterance. On the other hand, the information state (IS) of the dialogue denotes the level of grounding in the current conversation.

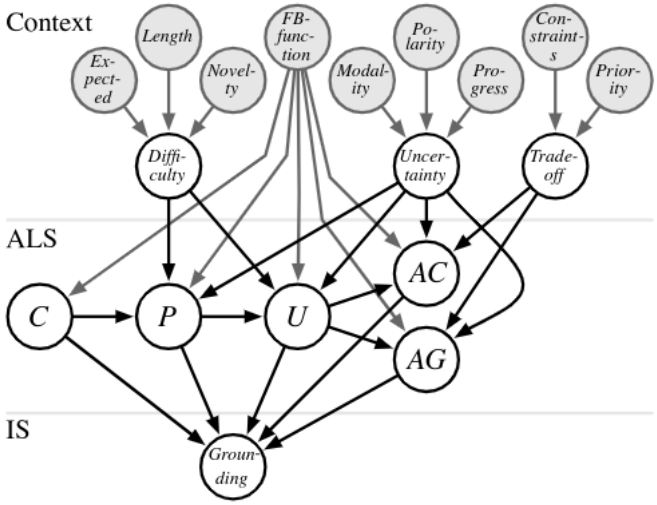

Figure 4
Structure of the Bayesian model of the listener. The variables shaded in grey are fully observable to a speaker (*FB function, modality, polarity,* and *progress* are derived from the listener's feedback signal). Source of the picture: (Buschmeier and Kopp 2012).

The ALS model was later made part of the *attentive speaker agent* that is able to adapt its communicative behaviour based on user feedback. A study was conducted investigating the willingness of human listeners to provide communicative feedback in an interaction with an *attentive speaker agent* and the ability of these humans to notice the collaborative communicative behaviour of the agent (Buschmeier and Kopp 2018). In the study, the observation of the properties of feedback was done by a human who entered the corresponding context values (cf. figure 4) into the system that autonomously interpreted this feedback and in turn adapted its own communicative behaviour, including elicitation of feedback from the listener with verbal and non-verbal cues. Also, two baseline conditions were added in which the agent did not analyse the feedback, but followed a fixed strategy instead: to either always ask for feedback after presenting a unit of information or not ask at all. In general, the findings show that the participants provided feedback to the *attentive speaker agent* in a form similar to human-human interaction and stopped providing feedback to the agents that did not analyse it. Additionally, the participants recognised the attentiveness and adaptiveness of the

attentive speaker agent and also of the agent that was constantly requesting their feedback, yet only the former was ascribed a desire to be understood and helpful.

5. Discussion

In this paper we have presented an overview of methods for knowledge modelling for the representation of common ground in artificial conversational agents. Three categories of knowledge with corresponding representation formalisms were discussed: factual knowledge about the world, personalised knowledge about the user and knowledge about user's knowledge, beliefs and mental state, each of these serving its unique purpose in various types of dialogue systems. Knowledge-awareness in general allows these systems to generate more informative and helpful responses.

With the emergence of neural conversational models, many researchers in the field of dialogue systems have abandoned the classical plan-based approach to dialogue management in favor of the data-driven approach. However, as was mentioned in chapter 4.1, even advanced state-of-the-art neural networks lack the necessary representations of mental states, which results in them struggling with false-belief tasks. These representations would also allow conversational systems to establish *common ground* with the users, to model what knowledge can be considered *shared* throughout the process of interaction. Surprisingly, the availability of novel research on this topic is rather low when it comes to conversational agents and many approaches are dating back to the older plan-based systems (cf. (Kopp and Krämer 2021)).

Concepts such as theory of mind are mostly adapted for human-robot teams and are heavily grounded in observations about the world which conversational agents might not have direct access to, only learning about it indirectly through the information exchanged with the user. So if the robotic theory of mind cannot be transferred to the conversational domain one-to-one, special adaptation is necessary, as the representation of mental states and beliefs is very important for dialogue. Additionally, one needs to consider the aspect of interactivity of dialogues. If dialogue systems are supposed to make inferences about the user's mental state and beliefs by using theory of mind models, these are required to be efficient enough to be deployed online, while maintaining reasonable accuracy. The approach of Pöppel and Kopp (Pöppel and Kopp 2018) described in chapter 4.1 could be beneficial in this case, however, one needs to account for the complexity of conversational tasks. This complexity makes it challenging to identify properties of the task that can serve as the adequate basis for the creation of simplified specialised ToM models which then could be integrated into the combination model able to switch between them in order to best explain the observed behaviour.

In chapter 4.2, the domain of argumentative dialogues presented conversational scenarios where it is crucial to be able to recognise and understand the perspective of others. However, perhaps a more general and more sought-after domain where perspectives also play a major role are explanation dialogues.

Explainable AI is on the rise now and researchers argue for the social nature of explanations (Miller 2019) that should not be ignored. Explanations of the same machine learning algorithm provided to an AI expert, an elderly person with no technical experience, and a 30 y.o. technology enthusiast with a smart home would all be different. These differences can apply not only to the vocabulary used, but also potentially to dialogue structure. Consider delivering the explanation to the expert in one long turn, or allowing the technical enthusiast more room to chime in with "what-if" questions, or asking the elderly person for more feedback to ensure understanding. Ideally, in order to make these explanations different, the system needs to not only have a good factual

model of their explanandum, which was talked about in chapter 2, and not only to know the user and their personality, for which a plethora of methods were discussed in chapter 3, but also to know their mental state, to know what users believe and be able to reason about the dynamics of belief updates during the process of explanation, to know what sort of knowledge has been negotiated enough to be considered shared and can be freely referenced in the future. These are valid research areas that can be tackled, and the methods described in this paper can build a foundation for the discovery of further approaches.

To close the loop with the introduction to this paper, let us consider how voice-based assistants such as Alexa and Google Assistant could be enhanced with mental state modelling. One of the findings of the study by Porcheron and colleagues (Porcheron et al. 2018) was that the families that used smart speakers embedded them in conversational situations within the family, yet ultimately did not recognise them as interlocutors. However, if a voice-based assistant was able to maintain mental models for all family members, and to bring this to bear recognisably in dialogue, they would be able to actively participate in those conversational situations. Further, they would become able to cooperatively resolve communicative issues, for example, in case of misunderstanding or present family members having conflicting goals and desires with regards to the way they wish to use the voice-based assistant.

It would be very interesting to study how the VA would be perceived in such a case and what new group dynamics would emerge during interactions.

Acknowledgments

This research is partially supported by the Volkswagen Foundation that is funding the project IMPACT. Project IMPACT is a cooperation of several research groups belonging to University of Duisburg-Essen, Bielefeld University, University of Kassel and Lutheran University of Applied Sciences in Nuremberg.

References

Adhikari, Ashutosh, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, and William L. Hamilton. 2020. Learning Dynamic Belief Graphs to Generalize on Text-Based Games. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, December.

Baker, Chris L., Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(64).

Balog, Krisztian and Tom Kenter. 2019. Personal Knowledge Graphs: A Research Agenda. In *Proceedings of the 2019 ACM SIGIR Conference on Theory of Information Retrieval (ICTIR 2019)*, Santa Clara, CA, USA, October.

Bang, Jeesoo, Hyungjong Noh, Yonghee Kim, and Gary Geunbae Lee. 2015. Example-based Chat-oriented Dialogue System with Personalized Long-term Memory. In *Proceedings of the 2015 International Conference on Big Data and Smart Computing (BIGCOMP 2015)*, pages 238–243, Jeju City, South Korea, February.

Baron-Cohen, Simon. 1995. *Mindblindness*. MIT Press, Cambridge, MA, USA.

Baron-Cohen, Simon, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21:37–46.

Berners-Lee, Tim. 2006. Linked data. Online: `https://www.w3.org/DesignIssues/LinkedData.html`. Accessed on 06.05.2021.

Bolander, Thomas. 2018. Seeing Is Believing: Formalising False-Belief Tasks in Dynamic Epistemic Logic. In van Ditmarsch H. and Sandu G., editors, *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, volume 12 of *Outstanding Contributions to Logic*. Springer, Cham.

Buschmeier, Hendrik and Stefan Kopp. 2011. Towards Conversational Agents That Attend to and Adapt to Communicative User Feedback. In *Proceedings of the 11th International Conference*

on Intelligent Virtual Agents (IVA-2011), pages 169–182, Reykjavik, Iceland, September.
Buschmeier, Hendrik 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, September.
Buschmeier, Hendrik 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)*, Stockholm, Sweden, July.
Clark, Herbert H. and Susan E. Brennan. 1991. Grounding in Communication. In L. B. Resnick, J. Levine, and S. D. Behrend, editors, *Perspectives on Socially Shared Cognition*. American Psychology Association, Washington, D.C., pages 222–233.
Clark, Leigh, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*, Glasgow, Scotland, UK, May.
Devin, Sandra and Rachid Alami. 2016. An implemented theory of mind to improve human-robot shared plans execution. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot-Interaction (HRI 2016)*, Christchurch, New Zealand, March.
Dinan, Emily, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019)*, New Orleans, LA, USA, May.
Dissing, Lasse and Thomas Bolander. 2020. Implementing Theory of Mind on a Robot Using Dynamic Epistemic Logic. In *Proceedings of the 29th International Joint Conference of Artificial Intelligence (IJCAI-20)*, Yokohama, Japan, January.
Ebbinghaus, Hermann. 2011. *Memory: A contribution to experimental psychology*. Martino Fine Books, Eastford, CT, USA, reprint of 1913 edition.
Gao, Jianfeng, Michel Galley, and Lihong Li. 2019. Neural Approaches to Conversational AI: Question Answering, Task-Oriented Dialogues and Social Chatbots. *Foundations and Trends ® in Information Retrieval*, 13(2-3):127–298.
Ghazvininejad, Marjan, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A Knowledge-Grounded Neural Conversation Model. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, LA, USA, February.
Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.
Han, Sangdo, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting Knowledge Base to Generate Responses for Natural Language Dialog Listening Agents. In *Proceedings of the 16th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2015)*, Prague, Czech Republic, September.
Hiatt, Laura M., Anthony M. Harrison, and J. Gregory Trafton. 2011. Accomodating Human Variability in Human-Robot Teams through Theory of Mind. In *Proceedings of the 22nd International Joint Conference of Artificial Intelligence (IJCAI-11)*, Barcelona, Spain, July.
Hunter, Anthony, Sylwia Polberg, and Matthias Thimm. 2020. Epistemic Graphs for Representing and Reasoning with Positive and Negative Influences of Arguments. *Artificial Intelligence*, 281.
Kim, Yonghee, Jeesoo Bang, Junhwi Choi, Seonghan Ryu, Sangjun Koo, and Gary Geunbae Lee. 2014. Acquisition and Use of Long-Time Memory for Personalized Dialogue Systems. In *Proceedings of the International Workshop on Mulitmodal Analyses Enabling Artificial Agents in Human-Machine Interaction (MA3HMI 2014)*, Singapore, Singapore, September.
Kinsella, Bret. 2020. Nearly 90 million U.S. adults have smart speakers, adoption now exceeds one-third of consumers. Online: `shorturl.at/iDJ69`. Accessed on 28.04.2021.
Kopp, Stefan and Nicole Krämer. 2021. Revisiting human-agent communication: The importance of joint co-construction and understanding mental states. *Frontiers in Psychology*, 12:597.
Krämer, Nicole, Astrid M. Rosenthal-von der Pütten, and Sabrina Eimler. 2012. Human-Agent and Human-Robot Interaction Theory: Similarities to and Differences from Human-Human Interaction. In M. Zacarias and de Oliviera J.V., editors, *Human-Computer Interaction: The Agency Perspective*, volume 396 of *Studies in Computational Intelligence*. Springer, Berlin, Heidelberg.

Lee, Benny P.H. 2001. Mutual knowledge, background knowledge and shared beliefs: Their roles in establishing common ground. *Journal of Pragmatics*, 33:21–44.

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2012. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, 1:1–5.

Li, Xiang, Gokhan Tur, Dilek Hakkani-Tür, and Qi Li. 2014. Personal Knowledge Graph Population from User Utterances in Conversational Understanding. In *Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT 2014)*, South Lake Tahoe, NV, USA, December.

Luger, Ewa and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2016)*, pages 5286–5297, San Jose, CA, USA, May.

Luo, Liangchen, Wenhao Huang, Qi Zeng, Zaiqing Nie, and Xu Sun. 2019. Learning Personalized End-to-End Goal-Oriented Dialog. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 6794–6801, Honolulu, HI, USA, January.

Ma, Yukun, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A Survey on Empathetic Dialogue Systems. *Information Fusion*, 64:50–70.

Miller, Tim. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267:1–38.

Moon, Seungwhan, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Memory Graph Networks for Explainable Memory-grounded Question Answering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL 2019)*, pages 728–736, Hong Kong, China, November.

Nematzadeh, Aida, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating Theory of Mind in Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 2392–2400, Brussels, Belgium, November.

Pei, Jiahuan, Pengjie Ren, and Maarten de Rijke. 2021. A Cooperative Memory Network for Personalized Task-oriented Dialogue Systems with Incomplete User Profiles. In *Proceedings of the Web Conference 2021 (WWW 2021)*, Online, April.

Peters, Christopher Edward. 2005. Foundations of an agent theory of mind model for conversation initiation in virtual environments. In *Proceedings of the AISB-05 Joint Symposium on Virtual Social Agents*, Hatfield, UK, April.

Pöppel, Jan and Stefan Kopp. 2018. Satisficing Models of Bayesian Theory of Mind for Explaining Behavior of Differently Uncertain Agents. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, Stockholm, Sweden, July.

Porcheron, Martin, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2018)*, Montreal, QC, Canada, April.

Qiu, Liang, Yizhou Zhao, Yuan Liang, Pan Lu, Weiyan Shi, Zhou Yu, and Song-Chun Zhu. 2021. Towards Socially Intelligent Agents with Mental State Transition and Human Utility. *Computing Research Repository (CoRR)*, arXiv:2103.07011.

Rabinowitz, Neil, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine Theory of Mind. In *Proceedings of the 35th International Conference of Machine Learning (ICML 2018)*, Stockholm, Sweden, July.

Ritter, Alan, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, Edinburgh, Scotland, UK, July.

Scassellati, Brian. 2002. Theory of Mind for a Humanoid Robot. *Autonomous Robots*, 12:13–24.

Serban, Iulian Vlad, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative Deep Neural Networks for Dialogue: A Short Review. In *Proceedings of 30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, December.

Sordoni, Alessandro, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *Proceedings of 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2015)*, Denver, CO, USA, June.

Speer, Robyn, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*

(AAAI-17), San Francisco, CA, USA, February.
Tigunova, Anna, Andrew Yates, Paramita Mirza, and Gerhard Weikum. 2019. Listening between the Lines: Learning Personal Attributes from Conversations. In *Proceedings of the 28th The Web Conference (WWW 2019)*, San Francisco, CA, USA, May.
Wang, Qiaosi, Koustuv Saha, Eric Gregori, David A. Joyner, and Ashok K. Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive about a Virtual Teaching Assistant. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI 2021)*, Yokohama, Japan, May.
Wu, Sixing, Ying Li, Dawei Zhang, and Zhonghai Wu. 2020. Improving Knowledge-Aware Dialogue Response Generation by Using Human-Written Prototype Dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, Online, November.
Yin, Jun, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2016. Neural Generative Question Answering. In *Proceedings of the Twenty-Fifth Joint Conference on Artificial Intelligence (IJCAI-16)*, New York City, NY, USA, July.
Zhang, Houyu, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graph. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online, July.
Zhou, Hao, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *Proceedings of the Twenty-Seventh Joint Conference on Artificial Intelligence (IJCAI-18)*, Stockholm, Sweden, July.
Zhu, Wenya, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible End-to-End Dialogue System for Knowledge Grounded Conversation. *Computing Research Repository (CoRR)*, arXiv:1709.04264.

Knowledge Modelling for Establishment of Common Ground in Dialogue Systems

Discussion points raised by
Francesco Cutugno
and Maria Di Maro

The paper deals with the process of grounding in dialogue systems, modelled in terms of factual knowledge of the world, knowledge concerning the user, and the hypothesis of mental knowledge state of the user, i.e., theory of mind. The difficulty of describing and modelling this pragmatic process in conversational agents emerges here in the necessity to refer to and integrate other cognitive theories. Specifically, considering that there are diverse types of shared sets of knowledge, the question that can be addressed refers to their possible different modelling strategy. More in detail, how can they be differently modelled according to their functions? As these sets of knowledge can be partially represented as different aspects of the Common Ground (GC) (Clark 2020), it would be worth exploring how they also interact with one another to successfully communicate. The processes described in the paper, which reflect the state-of-the-art, point out how the success of such grounding applications requires a consistent number of interactions or dialogue turns to efficiently ground information to be used to personalise a dialogue or to infer user's mental states. In this sense, corpus-based training processes, with or without probability-based methods, could be considered as a good starting point. Citing the author's abstract, "[. . .] this article provides a basic overview of current research on knowledge modelling for the establishment of common ground (henceforth CG) in dialogue systems." The overview is by far more than "basic" and covers a wide range of issues related to CG deepening how to integrate three types of knowledge (i.e., factual, personalised, and beliefs about user knowledge) into any form of automatic system able to manage with task oriented (and not only with them) dialogues. Even if it is not clearly noted in the paper, the introduction of a module able to introduce and represent CG in the general architecture of an automatic dialogue system manager, needs to be strictly "synchronised" with another fundamental module in the architecture: the Dialogue State Tracker (henceforth DST), which, in the recent literature, is more and more becoming the real "pulsing hearth" of these systems. Provided that DST systems have been deeply transformed by the application of Deep Neural Networks, contextual (in a very wide sense) embeddings, inexplicable procedures whose details we all are trying to explain, it could be worth exploring how this is reflected into CG module design. In other words, provided that automatic CG representation processes are called to interact with DST at any time, it is interesting to know what the authors' vision is on the evolution of CG technologies faced to DST systems affected by a high level of complexity. More specifically, under which constraints is it imaginable that also CG technologies can go "into deep"? Natural dialogues, both task oriented and general, have a temporal dynamic. Dialogue state evolves with time and so CG does. We have found very few literature references on evolving systems, able, for example, to find inconsistencies, or re-align dialogue states along with the dialogic situations

that can appear during interaction. CG and knowledge representation can be thought "static" and encyclopaedic but some pointer, indexes, should be active and varying with time, or better, with turns advancements. What is authors' idea on this matter? In conclusion, it appears almost clear that in the next future, online learning techniques will be introduced more and more pervasively into dialogue systems. Again, temporal evolution awareness and state tracking will take an advantage by this injection. But what about CG? And how Deep Neural Network and online learning will be integrated?

References

Clark, Herbert H. 2020. Common ground. *The International Encyclopedia of Linguistic Anthropology*, pages 1–5.

Knowledge Modelling for Establishment of Common Ground in Dialogue Systems

Response to the discussion points
by Lina Varonina
and Stefan Kopp

Here, we would like to respond to the questions and discussion points raised by Francesco Cutugno and Maria Di Maro in the wake of our paper on knowledge modelling for common ground establishment in dialogue systems. These points are concerned with the connection between the modules for common ground (CG) modelling and dialogue state tracking (DST), and how recent developments with the introduction of deep learning methods to DST can influence CG and knowledge modelling.

Recent research on DST has started to recognise the importance of connecting dialogue context with background information about a domain (Zhou and Small 2019; Ouyang et al. 2020; Chen et al. 2020; Liao et al. 2021). While this is by no means a novel insight, is has not been incorporated much into technical modelling approaches to DST. Further, even in smaller task-oriented domains it is often necessary to look beyond one turn to understand the user and successfully accomplish the task, as one turn will most likely not carry all the information the system needs. In contrast, contemporary voice assistants focus on one-shot request-response interactions with limited needs for context understanding. This is one of the problems users encounter with commercially available voice assistants, which usually expect users "by design" to provide all the necessary information and to ensure that it is understood by the system. When integrating background knowledge about the domain with the information provided by the user during dialogue, however, a system can resolve under-specified requests by making assumptions about user goals (Ouyang et al. 2020): If the user booked a hotel and a restaurant for two people and then wants to also book a taxi, it is highly likely that the taxi will be required for the same two people to transfer from the hotel to the restaurant.

The bigger point behind this argument is that communicative goals of the conversation partners are part of their mental state and the ability to infer mental state facilitates the construction of CG as per the definition of (Clark and Brennan 1991; Lee 2001) that is used in our paper, i.e. mutual knowledge, beliefs and assumptions that the parties have in common either due to similar background or because they were negotiated during the interaction. We argue that to solve the DST challenge one needs to re-recognise the importance of such mental state modelling for human-agent conversational interaction in the future. The current focus of research seems to lie on inference and prediction, while one of the main aspects of CG seen in the above-presented definition is the negotiation of knowledge as the interlocutors cannot be sure that their interpretation of the other's mental state is correct (Kopp and Krämer 2021). Thus, future research should aim to extend the capabilities of dialogue systems to include representations of knowledge that go beyond taking the information recovered from dialogue history as "objective truth". Instead, these representations should incorporate aspects of the interlocutor's mental state, such as epistemic stances or degrees of belief, in order to

account for different degrees of user's understanding or agreement with regards to a particular piece of information.

As we describe in our paper, graphs are an important representation of knowledge in the context of dialogue systems and research on DST often uses this form of representation in combination with deep neural networks (Zhou and Small 2019; Chen et al. 2020; Liao et al. 2021). However, we do believe that there is a need for bringing these methods closer together and that graph-based representations can be embedded into neural dialogue state trackers to enhance the quality of the dialogues through the introduction of cross-turn grounded context and general domain knowledge as argued above. Cutugno and Di Maro correctly note that it is crucial to account for the dynamic nature of knowledge in the context of building CG. Different types of knowledge can exhibit different temporal dynamics of change. Those distinctions can be traced back to Description Logics (Baader, Horrocks, and Sattler 2008) with its concepts of T-Box (terminological box) and A-Box (assertional box). The T-Box contains descriptions of properties and roles of general domain concepts and the relationships between these concepts (comparable with a database schema). The A-Box, on the other hand, contains properties of and relationships between individual instances of these concepts (comparable with data within a database). Looking at modern knowledge representation approaches, one can see parallels of the T-Box and A-Box with the concepts of ontology and knowledge graphs, respectively. In fact, ontology languages such as OWL (Bock et al. 2012), a widely-used web ontology language developed by W3C OWL Working Group, are often based on Description Logics.

Many examples of knowledge modelling discussed in our paper consider knowledge that is static within the use case, e.g., factual domain knowledge. However, we argue that in order to build truly conversational dialogue systems capable of co-constructing CG with their human user, the dynamic aspect of knowledge cannot be discarded. That is, characterising a dialogue state only based on static domain knowledge is insufficient because, even if the topic of a conversation is not changing, user's stance with regards to it may. These notions are currently being introduced into DST research. The work in (Zhou and Small 2019) features a dynamically changing knowledge graph for DST to represent relationships between slots and their values. In their work, they also consider the labels "not mentioned" and "user doesn't care" with regards to possible slot values. This can be considered a basic expression of dynamically changing stance about knowledge within conversation.

The other way around, even the system can have its own stance that evolves throughout the discourse. Depending on the type of the dialogue and the communicative goals of the human and the agent, it may then be necessary to align their beliefs about the domain through interaction, for instance by explanation or argumentation. Hereby it is important to separate representations of the agent's beliefs about the domain and its beliefs about the user's beliefs about the domain. Especially challenging here is that changes in user's beliefs about the domain are never directly observable and can only be inferred under uncertainty from communicative responses or feedback signals. An interesting research question is thus whether a conversational system can reduce this uncertainty with specific dialogue strategies and feedback elicitation in order to more efficiently infer the mental state of the user.

References

Baader, Franz, Ian Horrocks, and Ulrike Sattler. 2008. Description Logics. In Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter, editors, *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*. Elsevier, pages 135–179.

Bock, Conrad, Achille Fokoue, Peter Haase, Rinke Hoekstra, Ian Horrocks, Alan Ruttenberg, Uli Sattler, and Michael Smith. 2012. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition). Online. Accessed on 26.10.2021.

Chen, Lu, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, New York, NY, February 7-12.

Clark, Herbert H. and Susan E. Brennan. 1991. Grounding in Communication. In L. B. Resnick, J. Levine, and S. D. Behrend, editors, *Perspectives on Socially Shared Cognition*. American Psychology Association, Washington, D.C., pages 222–233.

Kopp, Stefan and Nicole Krämer. 2021. Revisiting human-agent communication: The importance of joint co-construction and understanding mental states. *Frontiers in Psychology*, 12:597.

Lee, Benny P.H. 2001. Mutual knowledge, background knowledge and shared beliefs: Their roles in establishing common ground. *Journal of Pragmatics*, 33:21–44.

Liao, Lizi, Le Hong Long, Yunshan Ma, Wenqiang Lei, and Tat-Seng Chua. 2021. Dialogue state tracking with incremental reasoning. *Transactions of the Association for Computational Linguistics*, 9:557–569.

Ouyang, Yawen, Moxin Chen, Xinyu Dai, Yinggong Zhao, Shujiang Huang, and Jiajun Chen. 2020. Dialogue state tracking with explicit slot connection modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July.

Zhou, Li and Kevin Small. 2019. Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering. *arXiv preprint*, (arXiv:1911.06192).