AI Communications 35 (2022) 295–308
DOI 10.3233/AIC-220147
IOS Press

Conversational AI for multi-agent communication in Natural Language

Research directions at the Interaction Lab

Oliver Lemon
School of Mathematical and Computer Sciences, Heriot-Watt University, United Kingdom
E-mail: o.lemon@hw.ac.uk

Abstract. Research at the Interaction Lab focuses on human-agent communication using conversational Natural Language. The ultimate goal is to create systems where humans and AI agents (including embodied robots) can spontaneously form teams and coordinate shared tasks through the use of Natural Language conversation as a universal communication interface. This paper first introduces machine learning approaches to problems in conversational AI in general, where computational agents must coordinate with humans to solve tasks using conversational Natural Language. It also covers some of the practical systems developed in the Interaction Lab, ranging from speech interfaces on smart speakers to embodied robots interacting using visually grounded language. In several cases communication between multiple agents is addressed. The paper surveys the central research problems addressed here, the approaches developed, and our main results. Some key open research questions and directions are then discussed, leading towards a future vision of conversational, collaborative multi-agent systems.

Keywords: Conversational AI, Natural Language Processing, human-robot interaction, multi-agent communication

1. Introduction

Conversational interaction in Natural Language between humans and artificial agents is a long-standing goal of AI research, depicted in popular culture in many different ways, ranging from HAL to C3PO, and beyond. One of the first and most well-known conversational systems was ELIZA [69], a text-based psychotherapist bot, developed in the 1960s. Since then, conversational systems research has flourished and become a mainstream aspect of real-world applications of AI, with deployed speech systems such as Siri, Alexa, and Google Assistant. While these deployed real-world systems are not (yet) great conversationalists, since they often fail to maintain context and coherence over multiple dialogue turns, current research attempts to make such systems more useable, more natural and human-like, more accurate, and safer. While most conversational systems have been built with 2 agents in mind (human and AI), recent work also explores how to extend methods to multi-agent settings, for example several humans interacting with a robot, or a human interacting with multiple conversational agents, for example in a smart home. In addition, recent efforts attempt to make such systems multimodal, situated, and embodied, meaning that they should be aware of visual and spatial aspects of the interaction context. This paper will outline the main research challenges and approaches in these areas, focusing on recent work at the Interaction Lab,[1] Heriot-Watt University, Edinburgh.

Broadly, conversational AI research can be broken down into several types of systems, each generating a collection of related research questions:

(1) Goal-driven cooperative communication: humans and agents need to coordinate on shared tasks, driven by the human's goals e.g. book a flight, find a restaurant, get information.

[1] Interaction Lab website: http://www.macs.hw.ac.uk/InteractionLab.

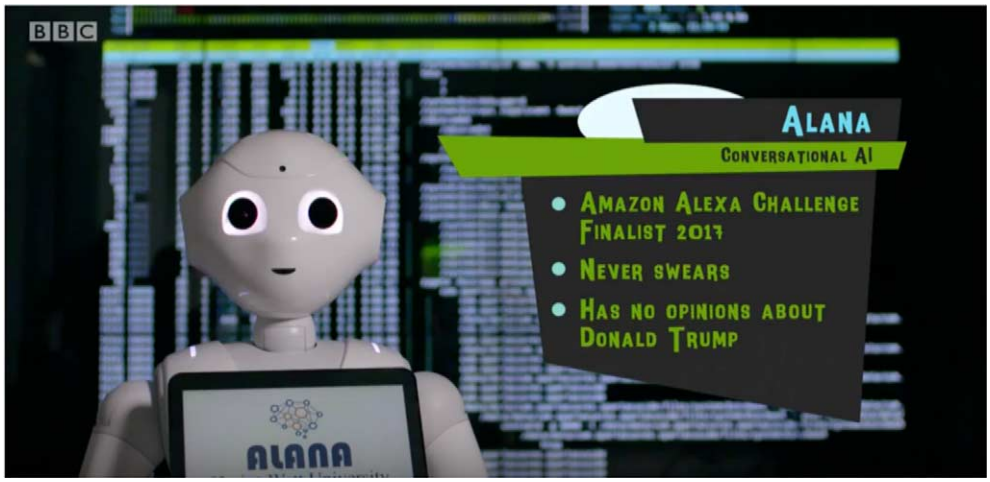

Fig. 1. The Alana system installed on a Pepper robot for the BBC's "Joy of AI" documentary. Alana is an open-domain social bot, also used in the MuMMER and SPRING projects [20,48].

(2) Situated and embodied communication: humans and agents coordinate to complete tasks within a shared visual and spatial environment e.g. a robot directing people around a building, a smart camera finding an object for a partially-sighted person.
(3) Social open-domain conversation: general conversation about topics of interest (movies, music, news, etc). This is not driven by any specific task goal, other than user engagement, information provision, and entertainment. See for example the open-domain conversational system Alana, in Figure 1.

This paper will largely focus on the first two types of system and the research problems and directions that they generate, as they are most closely related to research in Multi-Agent Systems.[2]

In general we model each human conversation partner as an agent which has goals, plans, and preferences, and which can send signals (usually Natural Language speech or text[3]) to other agents to convey its goals and request information, coordinate actions and plans etc.

The computational agents involved in the conversation then need to *decode* sequences of signals from humans to infer goals, plans, information needs, and so on. They then need to *decide* what to say or do next in the interaction, and they need to *encode* their decisions in Natural Language[4] output. These three processes – *decode, decide, encode* – are known as Natural Language Understanding (NLU), Dialogue Management (DM), and Natural Language Generation (NLG) respectively.[5] The field has largely moved from rule-based approaches for these three processes (up until approximately 2005) to developing statistical machine learning models for each one individually (from around 2005 until 2015), and has even more recently explored deep learning "end-to-end" models mapping directly from input symbol sequences to output sequences.

This paper will introduce research results and perspectives developed at the Interaction Lab, related to each of these topics.

1.1. The Interaction Lab: A brief description

The Interaction Lab was founded in 2009 at the Department of Computer Science, Heriot-Watt University, Edinburgh, Scotland. We also work with the recently formed National Robotarium within the Edinburgh Centre for Robotics (ECR). The group currently consists of 8 faculty, 8 postdoctoral researchers, and 16 PhD students. We

[2] See [11] for our work on social open-domain conversation.
[3] Though sometimes also gesture and facial expressions etc.
[4] Combinations of language with gestures, movement, graphics etc are also considered in work on multimodal output generation.
[5] More recently NLU is sometimes called "semantic decoding' or "intent recognition', context maintenance and updating aspects of DM are sometimes called "state tracking", and NLG is also sometimes called "Response Generation".

are well-known for pioneering machine learning approaches in several different aspects of conversational AI, dialogue systems, and human-robot interaction. In particular we were one of the main groups working on Reinforcement Learning approaches to dialogue management and Natural Language Generation [24,38,51] – methods which have now become widespread. We have been involved in a number of EC projects developing machine learning approaches to conversational AI (FP6 TALK and CLASSiC) and human-robot interaction (e.g. FP7 JAMES, H2020 MuMMER, H2020 SPRING), and researchers in the Interaction Lab also work in the EPSRC ORCA Hub in Robotics and Autonomous Systems, which explores trust-worthy, interpretable, and explainable AI. We were also twice finalists in the Amazon Alexa Prize (2017, 2018), deploying an open-domain social conversational system to millions of Alexa users in the US for 2 years.We work with several robot platforms (ARI, Furhat) for HRI research, as well as smart speaker systems (Alexa, Google home etc), within a purpose-built instrumented experimental space. In terms of industry collaborations and applications we have worked with Amazon, Google, Apple, PAL Robotics, Softbank Robotics, France Telecom, and BMW, amongst others. The Interaction Lab works closely with its spin-out company Alana AI,[6] specialising in conversational AI.

2. Research questions

The research questions addressed in our work fall into the following broad areas, often associated with the particular processing modules (e.g. language understanding, dialogue management, language generation) needed to build a working conversational system (see below). In general, our deployed systems necessarily address all of these problems. As well as developing modular approaches, we also build end-to-end systems as described in e.g. [18,62]. Specific projects often focus on one of more of the following research areas:

2.1. Data collection and annotation

This work involves collecting audio, video, or text examples of human-human or human-machine interaction behaviour in the domain or task to be modelled, and is often a necessary pre-requisite to developing a system, although it may be possible to use existing datasets of which there is an increasing number [55,66]. The data collected is generally used to train models and develop components for Natural Language understanding and generation, and dialogue management, or sometimes to build simulations in which to train[7] a dialogue manager [51].

Research here also involves analysing and sometimes transcribing and/or annotating data for specific interactive phenomena of interest (e.g. negotiation, gaze, gesture, interruptions, pauses, disfluency, feedback, clarification ....) [76]. Designing effective data collection methods is a key topic [46], and "Wizard-of-Oz" data collection setups are often deployed where a hidden operator controls a system or robot in order to elicit complex interactions.

2.2. Natural Language Understanding (NLU)

This research area involves building semantic decoders – models which mapping from Natural Language user input to a formal meaning representation for use in state tracking and dialogue management (e.g. "I want to fly to Paris" could map to a formal representation such as `Destination = Paris`). This process often includes the notion of "intents" or information "slots" (`Destination`) which can be filled with different values (`Paris`). Filled slots are then often used in database queries [50] whose results are then presented to the user via DM and NLG (e.g. "There are 3 flights to Paris, one at 9am..."). Often the concept of 'Dialogue Act' is also used, for example the difference between a `question` ("which city do you want to fly to?) and an *explicit-confirmation* ("do you want to fly to Paris?").

In recent deep learning end-to-end approaches these sorts of explicit intermediate semantic representations are not used, and mappings are directly learned between input sequences (user utterances) combined with context (encodings of previous turns of the dialogue), and output sequences (system utterances). Here, semantic representations are distributed in hidden layers of deep neural networks, and a number of important research questions concern the properties of such representations [64].

[6] Alana AI Ltd: http://www.alanaai.com.

[7] User simulations are also useful in avoiding expensive and time-consuming evaluations with real users, although real-user evaluations are always the ultimate test of developed systems and modules.

2.3. State tracking

This topic concerns the representation and updating of dialogue or interaction state [70], which can contain several types of contextual information on which Dialogue Management decisions can later be made. State information usually includes representations of user goals (perhaps with confidence scores due to uncertainty in both speech recognition and/or NLU [21]) and can also contain aspects of interaction history (e.g. mentioned topics or entities, agreed information so far, open questions, etc). If using Reinforcement Learning approaches to dialogue management, the state must be encoded in a suitable manner [24]. An accurate representation of state needs to be maintained as new information and observations accumulate [68]. Again, in recent deep learning approaches, no explicit state representation is developed, and the state information is encoded using sequences of prior turns in the interaction [10,62,65].

2.4. Dialogue Management (DM)

Dialogue or Interaction Management concerns the decision of 'what to say' next in an interaction, based on the current state [34], in order to achieve an end goal. This is generally a mapping (either learned, or designed using rules, or approached using planning) from dialogue states S to dialogue acts A. For example the system could in some states decide to explicitly confirm the user's destination – a dialogue act such as `exp-confirm (destination)`. We have modelled this sequential decision process using Reinforcement Learning with MDPs [24,52], POMDPs [9,74], and using Deep Reinforcement Learning [10].

2.5. Natural Language Generation (NLG)

The decision of 'how to say it' [34] in Natural Language (sometimes including gestural or graphical output) given a dialogue act chosen by the DM. For example mapping `exp-confirm (destination = Paris)` to the output "So you want to fly to Paris?". This process can be based on simple templates, and Reinforcement Learning [27,34,53] but for increased performance in terms of naturalness, contextual adaptation,and variability, end-to-end systems have been developed in more recent years [18].

2.6. Evaluation

Here we develop datasets, methods, and metrics for evaluating performance of each of the above processes and modules, or of full end-to-end systems, or of deployed systems with end users. We have been involved in or have driven community-wide benchmarking shared tasks such as the Spoken Dialog Challenge [6], Dialogue State Tracking (DST) shared tasks [68], and the End-to-end NLG challenge (E2E-NLG) [18].

Evaluation can involve comparison against benchmarks [40], the development of new evaluation metrics and frameworks for evaluation [64], online and/or in-situ lab-based experiments [9,26] with e.g. smart speakers [11], multimodal interactive systems, or robots [30,44]. For evaluations of deployed systems we generally use efficiency metrics based on a combination of full or partial task completion (e.g. did the user book the right flight?/choose a suitable restaurant?/select the correct object in an image?, etc) along with some penalties for dialogue length, repetition, and errors of various types. However for social open-domain dialogue systems longer dialogues are generally considered to be better as they show more engagement from the user [11,56].

3. Main approaches and key results

The main approaches pursued in the Interaction Lab concern the following key themes and methods:

3.1. Machine learning models of language processing

Most of our projects involve the development of statistical machine learning approaches to the research questions described above. For example we learn mappings from Natural Language to meaning representations [5,32,40,67], and statistical models for Natural Language Generation (NLG) [18,27,53] and Dialogue Management (DM) [24,52, 53,56]. An important concern is to build data-efficient methods for learning goal-oriented conversational systems from small amounts of data [19,58,59].

We have generally used features from representations of the dialogue context (i.e. what has already been said in the conversation) as additional signal for model training and decision-making. One example of this is re-ranking speech recognition or semantic interpretation hypotheses based on predicted next dialogue acts from the user (e.g. based on the context, are we expecting a question or an answer?) [36]. Another example is using context to re-rank possible system responses – either at the level of DM or NLG decision-making [56]. A particular recent focus is on the use and adaptation of large pre-trained vision-and-language models in interactive systems [63,65].

3.2. Reinforcement Learning (RL) for DM and NLG

In particular, we have explored and developed the use of RL methods for decision-making in interactive systems. Decisions can be made at the level of dialogue acts [24,34,52], words [18], gestures, or robot actions, or combinations of these [28]. We have also used Deep RL for learning communication policies in strategic non-cooperative games[8] such as negotiations in the game *Settlers of Catan* [10].

3.3. User simulations

Since RL requires large amounts of data for training, a particular focus has been on the development of simulated users for training and testing RL policies. This work ranges from rule-based to statistical simulations of user behaviours [52]. We also developed a simulation for multi-agent conversation in [29]. Evaluation of user simulations is also a key topic [49].

3.4. Incremental processing

Human language is produced and understood word-by-word (it is 'incremental'), and people can interrupt each other and even complete each other's utterances. However, most systems need to wait until the perceived end of a user utterance (which is often not accurate since humans often pause mid-utterance) before they start processing its meaning or deciding on a next action. As well as increasing latency in a system, this is not natural or fluid dialogue behaviour for humans. Related to this, spontaneous human language production is often disfluent – it contains many fillers (*um, er, uh* etc), pauses, re-starts, and repairs. We therefore work on incremental language processing in understanding and generation, and the handling of disfluencies [57], often using the framework of Dynamic Syntax [2,19].

3.5. Vision and language

A recent research focus has been on systems that include visual as well as linguistic information in the interaction, and which in some cases learn visually-grounded word meanings [61,64,65], seeking to address the classic 'symbol-grounding' problem of AI [23]. Such setups are also known as 'multi-modal' interaction systems, as they combine the information modalities of language and vision. Figure 2 shows an example of a recently-developed deep learning model for learning visually grounded language [65]. Researchers in the Interaction Lab have shown that previous work on so-called 'Visual Dialog' does not really require taking dialogue context into account, and proposed new visual dialogue datasets where linguistic context matters [3]. We are currently working to further develop interactive systems for learning grounded language, for example within the 2022 Amazon Alexa SimBot challenge [47,63].

[8] See the STAC project http://www.irit.fr/STAC/index.html.

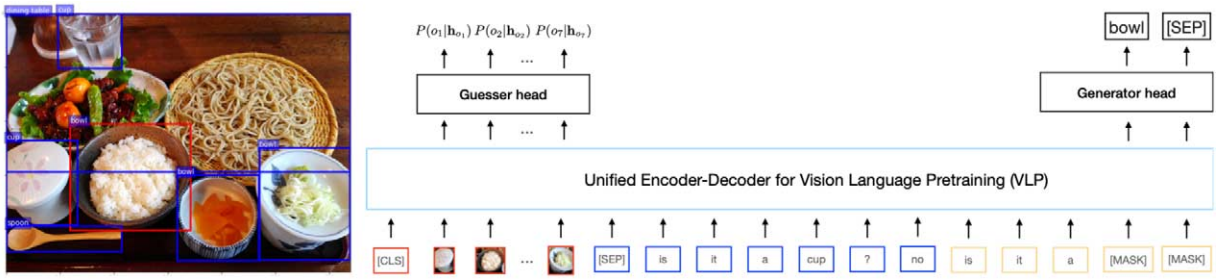

Fig. 2. Example of a vision-and-language conversational learning system (SPIEL) from [65]. Visual and conversation context (red, blue, yellow signals) are input to the VLP model. The Guesser head computes a probability distribution over possible target objects given the input, and the Generator head generates output language tokens word-by-word for an agent asking questions to find the target object e.g. "is it a bowl?"

3.6. Assistive visual conversational systems for blind and partially sighted people

We have developed several conversational systems where the interaction context includes visual and spatial information about objects and people. This allows us to create conversational systems which can assist blind and partially sighted people to find objects and access descriptions of places and scenes [4]. We have also developed interactive robots with the same capabilities [48]. Our spin-out company Alana AI is developing a visual conversational system for blind and partially sighted people in collaboration with the RNIB (Royal National Institute for the Blind).

3.7. Interactive visual language learning

We have developed self-play approaches to learning visually grounded language from data [65] as well as approaches that use conversational feedback from humans such as repairs ("no that is not red it's blue") to learn grounded word meanings [75]. In future human interactions with robots and smart devices, the ability to learn from such conversational feedback will be crucial [35]. Conversational feedback provides fine-grained signals which van be used to coordinate tasks and recover from mis-understandings, as well as in teaching agents the meanings of words.

3.8. Human-Robot Interaction (HRI)

Many of our systems are developed with HRI use cases as their ultimate goal. These systems involve multimodal information in interaction, for example considering user gestures and facial expressions as well as speech [44], and they are situated and embodied, meaning that visual and spatial information must be considered [48]. Some systems involve multiple humans, meaning that multi-agent social information and decision-making needs to be modelled.

For example, the JAMES project developed a robot bartender, and focussed on multi-user social interaction such as managing queuing behaviour [28]. The MuMMER project developed a socially entertaining conversational robot for single users at a time, adapting our Alexa Prize system Alana [11] on a Pepper robot (see Fig. 1) in a shopping mall scenario, and it was also able to give spatial and visual directions involving gestures as well as speech [20]. The SPRING project currently works on socially intelligent HRI in a multi-user scenario, in the context of a hospital waiting room. Here we deal with visual dialogue about the room (where to sit, lost objects etc) [48] as well as multi-party conversations tracking the individual goals of different users (e.g. one needs a coffee, another wants to check in) within the interaction context. The NLU, DM, and NLG processes then need to consider multiple agents in combination (see Section 4.1).

3.9. Ethical issues: Safety and factuality in language generation and conversation

Recent years have seen the exploration of large language models such as GPT-3 for generation of linguistic output in conversational systems. However, such models are known to hallucinate information, provide non-factual information, and encode various types of negative bias – providing serious ethical barriers to real-world deployment. Researchers in the Interaction Lab therefore work on methods to enhance the safety of such models [16]. In particular work on summarization and data-to-text generation seeks to make NLG systems more accurate and controllable [73]. On the other hand, human users can often be offensive towards conversational agents, and researchers in the Interaction Lab also work on ways to handle this, relating to the design of conversational personas [1].

3.10. Embodied interaction

Related to visual and spatial systems, as well as to HRI, we need to take into account conversational interaction with embodied systems which can take actions on objects situated in the world. We have developed methods that allow for conversationally interactive monitoring, negotiation, and execution of plans and activities with duration and resource use [17,37]. More recently, researchers in the Interaction Lab have developed deep learning systems such as 'Embodied BERT' (EmBERT) [62] which combine video streams and language to learn grounded language and action execution. Related to this work, we are currently the only European team participating in the Amazon Alexa SimBot challenge[9] (2022) which works on the TEACh dataset [47] of videos combined with conversations about household tasks (see Fig. 3).

3.11. New evaluation methods and metrics

We also work on fundamental issues regarding how the performance of various conversational AI models, modules, and whole systems should be evaluated. For example researchers in the Interaction Lab have proposed new evaluation methods and metrics for NLG [18,45,72], user simulations [49], open-domain conversation [11,71], and visually grounded language learning [64]. In general, the proposed NLG metrics seek to better measure accuracy and faithfulness with respect to source documents or data-structures, reducing "hallucination" of false or unsupported information. Work on metrics for open-domain conversation seeks to promote coherence, diversity, and the user-engagement of the conversations. Our work on visually grounded language learning proposes a multi-task evaluation framework [64] where aspects of learned representations such as attribute learning, generalizability, and compositionality are also evaluated as well as task performance.

4. Open problems/future work

We next survey some of the open research directions that researchers in the Interaction Lab are pursuing in conversational human-agent interaction.

4.1. Multi-party and multi-agent interaction

As mentioned above, the great majority of work on conversational AI has focused on bi-lateral interaction, between a human and a system. However, this perspective has broadened in recent years, with some recent work on multi-agent conversational interaction. We have worked on so-called "multi-party" dialogue systems, where a computational agent needs to interact with two or more humans (see [7] for some of the earliest work on this problem). Such a system cannot simply be modelled as several bi-lateral systems running in parallel, since there are important interactions between the humans (discussed below) which the system will need to track accurately – for example maintaining dialogue states not only for each individual but also representing information which has been agreed between them, or is still under discussion. There are several important research problems to be addressed here:

[9] http://sites.google.com/site/hwinteractionlab/home/amazon-simbot-challenge

Fig. 3. Example of a simulation environment for data collection and evaluation: the SimBot challenge using the TEACh dataset, from [47]. An Instructor (top) guides a Follower (with egocentric perspective in the main window) in a realistic 3D home environment to complete tasks such as "put all the forks in the sink". Our EMMA system competes in this challenge [63].

(1) Multi-party and multi-agent data collections – for model development and training we need to collect significant amounts of realistic data on multiple humans interacting with conversational systems, or more generally, mixtures of AI and human agents collaborating on tasks. This is more challenging than standard bilateral conversational data collections, since humans (and AI agents) may have conflicting or complementary tasks, goals, and information. For example we have recently designed a multi-party wizard-of-oz data collection with 2 humans and an ARI robot, where the humans can have shared goals (e.g. both want a coffee), or complementary information (e.g. A wants a coffee and B knows where the cafe is), or even conflicting information (e.g. A thinks the next meeting is in room 15, B thinks it is in room 17). Such considerations require careful design of tasks and methods to elicit natural, spontaneous, and complex multiparty dialogue phenomena. Recent online tools such as SLURK [22] also support multi-party data collections.
(2) Speaker diarization: *"who is saying what"?* – the problem of determining which human is speaking. Speech recognition systems are able to perform speaker recognition to some level of accuracy but multimodal information may also be used to assign speech signals to a particular speaker. Researchers in the Interaction Lab have recently performed an evaluation of current speech recognisers to perform this task [2].
(3) Multi-party NLU and addressee identification: *"who did they say it to?"* – once we know who said what, we need to determine which agent the utterance was addressed to. This can sometimes be explicit in the speech signal ("Jimi can you pass me the guacamole?") but can also be signalled by gaze, body pose, and gesture (I look at Jimi, point towards the table and say "Guacamole please?"). Utterances can also be broadcast to multiple people, or overheard by people that were not addressed – and in such cases the content of the utterance still updates their dialogue states.
(4) Incrementality: a challenging further aspect of multi-party NLU is that humans can complete each other's utterances. For example, in a hospital waiting room scenario from the SPRING project[10] – Patient: "I'm looking for..."; Companion: "Room 15 please". In such cases the user goal needs to be constructed by combining signals from two speakers.

[10]See http://spring-h2020.eu

(5) Multi-party state tracking: *"Who has which goals and information? What is agreed so far between participants? What issues are under discussion?"* – Following on from speaker diarization and NLU, multi-party state tracking needs to maintain an accurate representation of each individual agent's goals, plans, and information. Moreover, new representations need to be built and tracked for issues which have been agreed between 2 agents (or between a group of agents), and issues which are still 'live' in the conversation. For example Patient and Companion may have agreed that they both want a coffee, while a new person just joining the conversation will not be aware of this context. A conversational system needs to track all of this state.

(6) Multi-party DM and turn-taking: *"Who is going to speak next? What should I say? Who should I say it to?"* – as well as the state tracking aspects described above, a key aspect of state for interaction management with multiple agents is turn-taking i.e. who gets to take the conversational floor and how this turn-taking is managed and signalled. This is due to the constraint of there being only one audio channel resource available for signalling. Sometimes the current speaker will explicitly signal who the turn is being passed to ("Jimi do you know?"). Speakers can also bid for the floor and try to take the next turn though gaze and gesture. These aspects of multimodal turn-taking are very challenging and the community is only beginning to work on them. Further aspects from the DM decision point of view are whether or not to remain silent and allow others to speak, and who to address when speaking.

(7) Social behaviours: related to the above multi-party DM decisions, social behaviours in turn-taking sometimes need to be modelled. For example, in the JAMES project,[11] we modelled queuing behaviour at a drinks bar. Here the robot learned to ask users to wait until it had finished serving the current customer. For example if a new customer B was bidding for attention during an ongoing conversation with customer A, the system learned to ask B to wait their turn [21,28].

(8) Multi-party simulations – in order to train and test models without expensive and time consuming human interactions and evaluations, we will need facilities to simulate realistic multi-party user behaviours. This will involve extending bilateral approaches, such as in the Multi-User Simulation Environment (MUSE) model of [29], using tools and methods such as [60] or [31].

(9) Multi-party NLG – conversely to NLU, we will also need to ensure that system NL output is adapted to be able to explicitly address one or more humans, and there may need to be additional verbalisations to handle turn-taking and interruptions (e.g. "Excuse me, I think I know where you can get a coffee"). Template-based approaches to NLG can be fairly easily adapted on a case-by-case basis, but more general learned response generation models will require more extensive collections of multi-party training data.

4.2. Multi-agent conversational AI

The considerations mentioned in Section 4.1 are important for multi-agent communication consisting of a single agent and multiple humans, but some different issues arise when we consider multiple agents interacting with humans. Possible scenarios for such communication could be multiple autonomous vehicles, or smart home devices, which communicate amongst themselves (perhaps in some inter-agent protocol) but where there are also practical, legal, and/or ethical requirements for explainable AI such that humans be kept "in the loop" regarding their information and/or decisions.

Here some aspects of the agent-agent communication will need to be transparently communicated to the humans (e.g. if the agents have adopted a plan or collated some information which the humans need to be informed of [37]). In some simple cases this could be done by visual or gestural means (e.g. indicator lights on vehicles) but in general we will need mechanisms for agent-agent communication to be 'translated' into Natural Language. This perspective implies that at least some agent-agent communication must be explainable in Natural Language, and some recent research begins to explore this possibility [33,35].

[11] https://www.macs.hw.ac.uk/~rpp6/projects/james/

4.3. New data collections for multi-agent conversational collaboration

A central reason for the lack of work on conversational collaborative interaction in multi-agent systems is that current datasets do not contain semantic coordination phenomena of the type that will allow such skills to be learned. This is because (as argued by [8,25,35,41,54]) these datasets do not focus on agents with different goals and/or knowledge of the task that needs to be coordinated. The few exceptions, such as Cups [41] and MeetUp [25], have only small volumes of data, and almost no datasets go beyond pairs of agents. Moreover, while work on visually grounded language learning commonly uses shared real images [14,61], prior work on coordination in conversational grounding in visual tasks (such as [75]) has almost exclusively used simulated, artificial, and abstract data (e.g. using abstract shapes [39,43,77]). Note that the 'Visual Dialog' work of [12] does use agents with different visual information (one agent can see the image, one cannot) but there is no shared task here, making the whole dataset problematic [3,42].

Further, AI systems trained and tested only in simulation and on abstract images may not transfer well to real-world use cases – the 'sim-to-real' problem [15]. Therefore, an important first objective for our future work is to collect new data and create shared tasks in more realistic environments and to ensure that results and models are ecologically valid [13], i.e. involving more realistic tasks such as [47] which should better transfer to real-world settings. This argument is further developed in [35].

5. Conclusion

This paper surveyed the main research questions in conversational AI in relation to multi-agent interaction, which are being explored at the Interaction Lab. We first described the key concepts and research questions in such work, ranging over Natural Language Understanding, State Tracking, Dialogue Management, and Natural Language Generation, and we also described key research questions in data collection and evaluation methods. We then surveyed recent research methods and results from the Interaction Lab, involving the development of machine learning approaches for Natural Language Processing, visual conversational systems, grounded language learning, incremental language processing systems, end-to-end NLG, explainable AI, and safety. Along the way we briefly described several of our deployed systems, for example the Amazon Alexa Prize system 'Alana' and several embodied human-robot conversational systems in the social robots of the JAMES, MuMMER, and SPRING projects.

We then focused on conversational AI issues that are most important from a multi-agent point of view, specifically the issues involved in 'multi-party' conversational systems and our approaches to them.

In terms of our future perspective on important open challenges related to multi-agent systems research, we highlight two main areas: embodied conversation with visual context, and multi-agent communication in Natural Language. We are ultimately working towards systems where humans and AI agents (including embodied robots) can spontaneously form teams and coordinate shared tasks through the use of Natural Language conversation as a flexible and transparent common communication interface.

Acknowledgements

This work is partially funded by the EU Horizon 2020 program under grant agreement no. 871245: the SPRING project.[12] This paper has benefited from discussions with many members of the Interaction Lab, both past and present. The multi-agent perspective on conversational AI presented here in particular has been informed by discussions with Christian Dondrup, Arash Eshghi, Ioannis Konstas, Alessandro Suglia, and Verena Rieser.

[12] http://spring-h2020.eu/

References

[1] G. Abercrombie, A.C. Curry, M. Pandya and V. Rieser Alexa, *Google, Siri: What Are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants*, 2021.

[2] A. Addlesee, Y. Yu and A. Eshghi, A comprehensive evaluation of incremental speech recognition and diarization for conversational AI, in: *Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics*, Barcelona, Spain, 2020, pp. 3492–3503, https://aclanthology.org/2020.coling-main.312. doi:10.18653/v1/2020.coling-main.312.

[3] S. Agarwal, T. Bui, J.-Y. Lee, I. Konstas and V. Rieser, History for visual dialog: Do we really need it? in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 8182–8197, https://aclanthology.org/2020.acl-main.728. doi:10.18653/v1/2020.acl-main.728.

[4] K. Baker, A. Parekh, A. Fabre, A. Addlesee, R. Kruiper and O. Lemon, The spoon is in the sink: Assisting visually impaired people in the kitchen, in: *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021), Association for Computational Linguistics*, Gothenburg, Sweden, 2021, pp. 32–39, https://aclanthology.org/2021.reinact-1.5.

[5] V.S. Bastianelli and Rieser, SLURP: A spoken language understanding resource package, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2020, pp. 7252–7262, https://www.aclweb.org/anthology/2020.emnlp-main.588.

[6] A.W. Black, S. Burger, A. Conkie, H.W. Hastie, S. Keizer, O. Lemon, N. Merigaud, G. Parent, G. Schubiner, B. Thomson, J.D. Williams, K. Yu, S.J. Young and M. Eskénazi, Spoken dialog challenge 2010: Comparison of live and control test results, in: *Proceedings of the SIGDIAL 2011 Conference, the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Oregon Science & Health University*, June 17–18, 2011, Portland, Oregon, USA, The Association for Computer Linguistics, 2011, pp. 2–7, https://aclanthology.org/W11-2002/.

[7] D. Bohus and E. Horvitz, Facilitating multiparty dialog with gaze, gesture, and speech, in: *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, ICMI-MLMI'10*, Association for Computing Machinery, New York, NY, USA, 2010, ISBN 9781450304146. doi:10.1145/1891903.1891910.

[8] K.R. Chandu, Y. Bisk and A.W. Black, Grounding 'grounding' in NLP, in: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics*, Online, 2021, pp. 4283–4305, https://aclanthology.org/2021.findings-acl.375. doi:10.18653/v1/2021.findings-acl.375.

[9] P. Crook, S. Keizer, Z. Wang, W. Tang and O. Lemon, Real user evaluation of a POMDP spoken dialogue system using automatic belief compression, *Computer Speech and Language* **28**(4) (2014), 873–887. doi:10.1016/j.csl.2013.12.002.

[10] H. Cuayáhuitl, S. Keizer and O. Lemon, *Strategic Dialogue Management via Deep Reinforcement Learning, in: NIPS Workshop on Deep Reinforcement Learning*, 2015.

[11] A. Curry, I. Papaioannou, A. Suglia, S. Agarwal, I. Shalyminov, X. Xinnuo, O. Dusek, A. Eshghi, I. Konstas, V. Rieser and O. Lemon, Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking, in: *Proceedings of Alexa Prize*, 2018.

[12] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J.M.F. Moura, D. Parikh and D. Batra, Visual Dialog, in: *Proceedings of CVPR*, 2017.

[13] H. de Vries, D. Bahdanau and C. Manning, *Towards Ecologically Valid Research on Language User Interfaces*, 2020.

[14] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle and A. Courville, GuessWhat?! Visual object discovery through multi-modal dialogue, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4466–4475. doi:10.1109/CVPR.2017.475.

[15] M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford, L. Weihs, M. Yatskar and A. Farhadi, RoboTHOR: An Open Simulation-to-Real Embodied AI Platform, 2020, CoRR, https://arxiv.org/abs/2004.06799 arXiv:2004.06799.

[16] E. Dinan, G. Abercrombie, A. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau and V. Rieser, SafetyKit: First aid for measuring safety in open-domain conversational systems, in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics*, Dublin, Ireland, 2022, pp. 4113–4133, https://aclanthology.org/2022.acl-long.284.

[17] C. Dondrup, I. Papaioannou and O. Lemon, Petri Net Machines for Human-Agent Interaction, 2019, https://arxiv.org/abs/1909.06174. doi:10.48550/ARXIV.1909.06174.

[18] O. Dušek, J. Novikova and V. Rieser, Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge, *Computer Speech & Language* **59** (2020), 123–156, https://www.sciencedirect.com/science/article/pii/S0885230819300919. doi:10.1016/j.csl.2019.06.009.

[19] A. Eshghi, I. Shalyminov and O. Lemon, Bootstrapping incremental dialogue systems from minimal data: The generalisation power of dialogue grammars, in: *EMNLP*, 2017.

[20] M.E. Foster, B.G.W. Craenen, A.A. Deshmukh, O. Lemon, E. Bastianelli, C. Dondrup, I. Papaioannou, A. Vanzo, J. Odobez, O. Canévet, Y. Cao, W. He, Á. Martínez-González, P. Motlícek, R. Siegfried, R. Alami, K. Belhassein, G. Buisan, A. Clodic, A. Mayima, Y. Sallami, G. Sarthou, P. Singamaneni, J. Waldhart, A. Mazel, M. Caniot, M. Niemelä, P. Heikkilä, H. Lammi and A. Tammela, MuMMER: Socially Intelligent Human-Robot Interaction in Public Spaces, 2019, CoRR, http://arxiv.org/abs/1909.06749 arXiv:1909.06749.

[21] M.E. Foster, S. Keizer and O. Lemon, Action selection under uncertainty for a socially aware robot bartender, in: *Proceedings of Human-Robot Interaction (HRI)*, 2014.

[22] J. Götze, M. Paetzel-Prüsmann, W. Liermann, T. Diekmann and D. Schlangen, The SLURK Interaction Server Framework: Better Data for Better Dialog Models, 2022.
[23] S. Harnad, The symbol grounding problem, *Physica D: Nonlinear Phenomena* **42**(1–3) (1990), 335–346. doi:10.1016/0167-2789(90)90087-6.
[24] J. Henderson, O. Lemon and K. Georgila, Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets, *Computational Linguistics* **34**(4) (2008), 487–511, https://aclanthology.org/J08-4002. doi:10.1162/coli.2008.07-028-R2-05-82.
[25] N. Ilinykh, S. Zarrieß and D. Schlangen, MeetUp! A corpus of joint activity dialogues in a visual environment, in: *Proceedings of SEMDIAL 2019*, 2019.
[26] S. Janarthanam and O. Lemon, Adaptive referring expression generation in spoken dialogue systems: Evaluation with real users, in: *Proceedings of SIGDIAL*, 2010.
[27] S. Janarthanam and O. Lemon, Adaptive generation in dialogue systems using dynamic user modeling, *Computational Linguistics* **40**(4) (2014), 883–920. doi:10.1162/COLI_a_00203.
[28] S. Keizer, M. Ellen Foster, Z. Wang and O. Lemon, Machine Learning for Social Multiparty Human–Robot Interaction, *ACM Trans. Interact. Intell. Syst.* **4**(3) (2014). doi:10.1145/2600021.
[29] S. Keizer, M.E. Foster, O. Lemon, A. Gaschler and M. Giuliani, Training and evaluation of an MDP model for social multi-user human-robot interaction, in: *Proceedings of SIGDIAL*, 2013.
[30] S. Keizer, P. Kastoris, M.E. Foster, A. Deshmukh and O. Lemon, User evaluation of a multi-user social interaction model implemented on a nao robot, in: *Proceedings of the ICSR 2013 Workshop on Robots in Public Spaces*, 2013.
[31] T.E. Kim and A. Lipani, A multi-task based neural model to simulate users in goal-oriented dialogue systems, in: *Proc. SIGIR*, 2022. doi:10.1145/3477495.3531814.
[32] I. Konstas, S. Iyer, M. Yatskar, Y. Choi and L. Zettlemoyer, Neural AMR: Sequence-to-sequence models for parsing and generation, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics*, Vancouver, Canada, 2017, pp. 146–157, https://aclanthology.org/P17-1014. doi:10.18653/v1/P17-1014.
[33] A. Lazaridou, A. Potapenko and O. Tieleman, in: *Multi-Agent Communication Meets Natural Language: Synergies Between Functional and Structural Language Learning, in: ACL, Association for Computational Linguistics*, 2020, pp. 7663–7674.
[34] O. Lemon, Learning what to say and how to say it: Joint optimization of spoken dialogue management and natural language generation, *Computer Speech and Language* **25**(2) (2011), 210–221. doi:10.1016/j.csl.2010.04.005.
[35] O. Lemon, Conversational grounding in emergent communication – data and divergence, in: *Emergent Communication (EmeComm) Workshop at ICLR*, 2022, https://openreview.net/forum?id=BbG-m-0Xbq.
[36] O. Lemon and A. Gruenstein, Multithreaded context for robust conversational interfaces: Context-sensitive speech recognition and interpretation of corrective fragments, *ACM Transactions on Computer-Human Interaction (ACM TOCHI)* **11**(3) (2004), 241–267. doi:10.1145/1017494.1017496.
[37] O. Lemon, A. Gruenstein, A. Battle and S. Peters, in: *Multi-Tasking and Collaborative Activities in Dialogue Systems, in: SIGDIAL Workshop, the Association for Computer Linguistics*, 2002, pp. 113–124.
[38] O. Lemon and O. Pietquin, Machine learning for spoken dialogue systems, in: *INTERSPEECH, ISCA*, 2007, pp. 2685–2688.
[39] P.P. Liang, J. Chen, R. Salakhutdinov, L. Morency and S. Kottur, On Emergent Communication in Competitive Multi-Agent Teams, 2020, CoRR, https://arxiv.org/abs/2003.01848 arXiv:2003.01848.
[40] X. Liu, A. Eshghi, P. Swietojanski and V. Rieser, Benchmarking Natural Language Understanding Services for building Conversational Agents, 2019, CoRR, http://arxiv.org/abs/1903.05566 arXiv:1903.05566.
[41] S. Loáiciga, S. Dobnik and D. Schlangen, Reference and coreference in situated dialogue, in: *Proceedings of the Second Workshop on Advances in Language and Vision Research*, Association for Computational Linguistics, 2021, pp. 39–44, https://aclanthology.org/2021.alvr-1.7. doi:10.18653/v1/2021.alvr-1.7.
[42] D. Massiceti, P.K. Dokania, N. Siddharth and P.H.S. Torr, Visual dialogue without vision or dialogue, in: *proceedings of CRACT workshop at NeurIPS 2018*, 2019.
[43] A. Narayan-Chen, P. Jayannavar and J. Hockenmaier, Collaborative dialogue in minecraft, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, Florence, Italy, 2019, pp. 5405–5415, https://aclanthology.org/P19-1537. doi:10.18653/v1/P19-1537.
[44] J. Novikova, C. Dondrup, I. Papaioannou and O. Lemon, Sympathy begins with a smile, intelligence begins with a word: Use of multimodal features in spoken human-robot interaction, in: *Proceedings of the First Workshop on Language Grounding for Robotics*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 86–94, https://aclanthology.org/W17-2811. doi:10.18653/v1/W17-2811.
[45] J. Novikova, O. Dusek, A.C. Curry and V. Rieser, Why we need new evaluation metrics for NLG, in: *EMNLP*, 2017.
[46] J. Novikova, O. Lemon and V. Rieser, Crowd-sourcing NLG data: Pictures elicit better data, in: *Proceedings of INLG*, 2016, http://arxiv.org/abs/1608.00339.
[47] A. Padmakumar, J. Thomason, A. Shrivastava, P. Lange, A. Narayan-Chen, S. Gella, R. Piramuthu, G. Tur and D. Hakkani-Tur, 2021, TEACh: Task-driven Embodied Agents that Chat.
[48] J. Part, D. Hernandez-Garcia, Y. Yu, N. Gunson, C. Dondrup and O. Lemon, Towards visual dialogue for human-robot interaction (demonstration), in: *Human-Robot Interaction (HRI)*, 2021.
[49] O. Pietquin and H. Hastie, A survey on metrics for the evaluation of user simulations, *The Knowledge Engineering Review* **28** (2012), 59–73. doi:10.1017/S0269888912000343.
[50] V. Rieser and O. Lemon, Does this list contain what you were searching for?: Learning adaptive dialogue strategies for interactive question answering, *Natural Language Engineering* **15**(1) (2008), 55–72. Special Issue on Interactive Question Answering, MP.

[51] V. Rieser and O. Lemon, *Reinforcement Learning for Adaptive Dialogue Systems*, 1st edn, Springer, 2011.
[52] V. Rieser and O. Lemon, Learning and evaluation of dialogue strategies for new applications: Empirical methods for optimization from small data sets, *Computational Linguistics* **37**(1) (2011), 153–196. doi:10.1162/coli_a_00038.
[53] V. Rieser, O. Lemon and S. Keizer, Natural language generation as incremental planning under uncertainty: Adaptive information presentation for statistical dialogue systems, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **22**(5) (2014), 979–994. doi:10.1109/TASL.2014.2315271.
[54] D. Schlangen, Grounded Agreement Games: Emphasizing Conversational Grounding in Visual Dialogue Settings, 2019.
[55] I.V. Serban, R. Lowe, P. Henderson, L. Charlin and J. Pineau, A Survey of Available Corpora for Building Data-Driven Dialogue Systems, 2015, https://arxiv.org/abs/1512.05742. doi:10.48550/ARXIV.1512.05742.
[56] I. Shalyminov, O. Dušek and O. Lemon, Neural response ranking for social conversation: A data-efficient approach, in: *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI, Association for Computational Linguistics*, 2018, https://doi.org/10.18653%2Fv1%2Fw18-5701. doi:10.18653/v1/w18-5701.
[57] I. Shalyminov, A. Eshghi and O. Lemon, Multi-task learning for domain-general spoken disfluency detection in dialogue systems, 2018, arXiv:1810.03352.
[58] I. Shalyminov, S. Lee, A. Eshghi and O. Lemon, Data-efficient goal-oriented conversation with dialogue knowledge transfer networks, in: *EMNLP*, 2019.
[59] I. Shalyminov, S. Lee, A. Eshghi and O. Lemon, Few-shot dialogue generation without annotated data: A transfer learning approach, in: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, Association for Computational Linguistics*, Stockholm, Sweden, 2019, pp. 32–39, https://aclanthology.org/W19-5904. doi:10.18653/v1/W19-5904.
[60] W. Shi, K. Qian, X. Wang and Z. Yu, How to build user simulators to train RL-based dialog systems, in: *Proceedngs of EMNLP*, 2019.
[61] A. Suglia, Y. Bisk, I. Konstas, A. Vergari, E. Bastianelli, A. Vanzo and O. Lemon, An empirical study on the generalization power of neural representations learned via visual guessing games, in: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Association for Computational Linguistics, 2021, pp. 2135–2144, https://aclanthology.org/2021.eacl-main.183.
[62] A. Suglia, Q. Gao, J. Thomason, G. Thattai and G. Sukhatme, Embodied bert: A transformer model for embodied, language-guided visual task completion, 2021, arXiv preprint arXiv:2108.04927.
[63] A. Suglia, B. Hemanthage, M. Nikandrou, G. Pantazopoulos, A. Parekh, C.G. Arash Eshghi, I. Konstas, O. Lemon and V. Rieser, Demonstrating EMMA: Embodied MultiModal Agent for Language-guided Action Execution in 3D Simulated Environments, in: *Proceedings of SIGDIAL*, 2022.
[64] A. Suglia, I. Konstas, A. Vanzo, E. Bastianelli, D. Elliott, S. Frank and O. Lemon, CompGuessWhat?!: A multi-task evaluation framework for grounded language learning, in: *Proceedings of ACL*, 2020.
[65] A. Suglia, A. Vergari, I. Konstas, Y. Bisk, E. Bastianelli, A. Vanzo and O. Lemon, Imagining grounded conceptual representations from perceptual information in situated guessing games, in: *Proceedings of COLING*, 2020.
[66] A. Sundar and L. Heck, Multimodal Conversational AI: A Survey of Datasets and Approaches, 2022, https://arxiv.org/abs/2205.06907. doi:10.48550/ARXIV.2205.06907.
[67] A. Vanzo, E. Bastianelli and O. Lemon, Hierarchical multi-task natural language understanding for cross-domain conversational AI: HERMIT NLU, in: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, Association for Computational Linguistics*, Stockholm, Sweden, 2019, pp. 254–263, https://www.aclweb.org/anthology/W19-5931. doi:10.18653/v1/W19-5931.
[68] Z. Wang and O. Lemon, A simple and generic belief tracking mechanism for the dialogue state tracking challenge: On the believability of observed information, in: *Proc. SIGDIAL*, 2013.
[69] J. Weizenbaum, ELIZA – a computer program for the study of natural language communication between man and machine, *Commun. ACM* **9**(1) (1966), 36–45. doi:10.1145/365153.365168.
[70] J. Williams, A. Raux and M. Henderson, The dialog state tracking challenge series: A review, *Dialogue Discourse* **7** (2016), 4–33. doi:10.5087/dad.2016.301.
[71] X. Xu, O. Dusek, I. Konstas and V. Rieser, Better conversations by modeling, filtering, and optimizing for coherence and diversity, in: *EMNLP*, 2018.
[72] X. Xu, O. Dušek, J. Li, V. Rieser and I. Konstas, Fact-based content weighting for evaluating abstractive summarisation, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 5071–5081, https://aclanthology.org/2020.acl-main.455. doi:10.18653/v1/2020.acl-main.455.
[73] X. Xu, O. Dušek, V. Rieser and I. Konstas, AggGen: Ordering and aggregating while generating, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2021, pp. 1419–1434, https://aclanthology.org/2021.acl-long.113. doi:10.18653/v1/2021.acl-long.113.
[74] S. Young, M. Gašić, B. Thomson and J.D. Williams, POMDP-based statistical spoken dialog systems: A review, in: *Proceedings of the IEEE*, Vol. 101, 2013, pp. 1160–1179. doi:10.1109/JPROC.2012.2225812.
[75] Y. Yu, A. Eshghi and O. Lemon, Learning how to learn: An adaptive dialogue agent for incrementally learning visually grounded word meanings, in: *Proceedings of the First Workshop on Language Grounding for Robotics*, 2017, http://dx.doi.org/10.18653/v1/W17-2802. doi:10.18653/v1/w17-2802.
[76] Y. Yu, A. Eshghi, G. Mills and O. Lemon, The BURCHAK corpus: A challenge data set for interactive learning of visually grounded word meanings, in: *Proceedings of the Sixth Workshop on Vision and Language*, 2017, http://dx.doi.org/10.18653/v1/W17-2001. doi:10.18653/v1/w17-2001.

[77] S. Zarrieß, J. Hough, C. Kennington, R. Manuvinakurike, D. DeVault, R. Fernández and D. Schlangen, PentoRef: A corpus of spoken references in task-oriented dialogues, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 125–131, https://aclanthology.org/L16-1019.