Corpus of Multimodal Interaction for Collaborative Planning

Miltiadis Marios Katsakioris[1], Atanas Laskov[2], Ioannis Konstas[1], Helen Hastie[1]
[1]School of Mathematical and Computer Sciences
Heriot-Watt University, Edinburgh, UK
[2]SeeByte Ltd, Edinburgh, UK
mmk11, i.konstas, h.hastie@hw.ac.uk
atanas.laskov@seebyte.com

Abstract

As autonomous systems become more commonplace, we need a way to easily and naturally communicate to them our goals and collaboratively come up with a plan on how to achieve these goals. To this end, we conducted a Wizard of Oz study to gather data and investigate the way operators would collaboratively make plans via a conversational 'planning assistant' for remote autonomous systems. We present here a corpus of 22 dialogs from expert operators, which can be used to train such a system. Data analysis shows that multimodality is key to successful interaction, measured both quantitatively and qualitatively via user feedback.

1 Introduction

Our goal is to create a collaborative multimodal planning system in the form of a conversational agent called VERSO using both visual and natural language interaction, which in our case will be images of the plan and messages. In this work, we focus on the domain of Autonomous Underwater Vehicles (AUVs). Experts in this domain typically create a plan for vehicles using a visual interface on dedicated hardware on-shore, days before the mission. This planning process is complicated and requires expert knowledge. We propose a 'planning assistant' that is able to encapsulate this expert knowledge and make suggestions and guide the user through the planning process using natural language multimodal interaction. This, we hope, will allow for more precise and efficient plans and reduce operator training time. In addition, it will allow for anywhere access to planning for in-situ replanning in fast-moving dynamic scenarios, such as first responder scenarios or offshore for oil and gas.

2 Previous work

Conversational agents are becoming more widespread, varying from social (Li et al., 2016) and goal-oriented (Wen et al., 2017) to multimodal dialog systems, such as the Visual Dialog Challenge (Das et al., 2017) where an AI agent must hold a dialog with a human in natural language about a given visual content. However, for systems with both visual and spatial requirements, such as situated robot planning (Misra et al., 2018), developing accurate goal-oriented dialog systems can be extremely challenging, especially in dynamic environments, such as underwater.

The ultimate goal of this work is to learn a dialog strategy that optimizes interaction for quality and speed of plan creation, thus linking interaction style with extrinsic task success metrics. Therefore, we conducted a Wizard of Oz (WoZ) study for data collection that can be used to derive reward functions for Reinforcement Learning, as in (Rieser, 2008).

Similar work is shown in (Kitaev et al., 2019), where the task involves two humans collaboratively drawing objects with one being the teller and the other the person who draws. The agents must be able to adapt and hold a dialog about novel scenes that will be dynamically constructed. However, in our scenario the agent must be capable of not only adapting but also identifying and editing specific attributes of the dynamic objects that are being created in the process.

Previous data collection on situated dialog, such as the Map Task Corpus (Anderson et al., 1991), tackle the importance of referencing objects while giving instructions on a drawn map with landmarks either for identification purposes or for displaying the perceived understanding of their shared environment. Our task is different in that it involves subjects collaboratively creating a plan

Proceedings of SpLU-RoboNLP, pages 1–6
Minneapolis, MN, June 6, 2019.

on a nautical chart rather than passively following instructions. In addition, our environment is dynamic. New objects are being created and the user with the agent, together, come up with the desired referring expressions (see Figure 3). A similar interactive method is described in (Schlangen, 2016), where they ground non-linguistic visual information through conversation.

In situated dialog, each user can perceive the environment in a different way, meaning that referring expressions need to be carefully selected and verified, especially if the shared environment is ambiguous (Fang et al., 2013). Our contributions include: 1) a generic dialog framework and the implemented software to conduct multiple wizard WoZ experiments for multimodal collaborative planning interaction; 2) available on request, a corpus of 22 dialogs on 2 missions with varying complexities and 3) a corpus analysis (Section 4) indicating that incorporating an extra modality in conjunction with spatial referencing in a chatting interface is crucial for successfully planning missions.

3 Method and Experiment Set-up

Our 'planning assistant' conversational agent will interface with planning software called SeeTrack provided by our industrial partner SeeByte Ltd. SeeTrack can run with real AUVs running SeeByte's Neptune autonomy software or in simulation and allows the planning of missions by defining a set of objectives with techniques described in (Lane et al., 2013; Miguelanez et al., 2011; Petillot et al., 2009). These can include, for example, searching for unexploded mines by surveying areas in a search pattern, while collecting sensor data and if, for example, a suspect mine is found then the system can investigate a certain point further (referred to as target reacquisition).

We used two wizards for our experiment, see Figure 1 for the set-up. We refer here to the wizards as 1) Chatting-Wizard (CW), who alone communicates with the subject getting information that is required to create the plan; and 2) the SeeTrack Wizard (SW), who sits next to the CW and implements the subject's requirements into a plan using SeeTrack and passes plan updates in the form of images to the CW to pass onto the subject. The subject was in a separate room to the wizards and interacted via a chat window for receiving text and images of the updated plan and sending text.

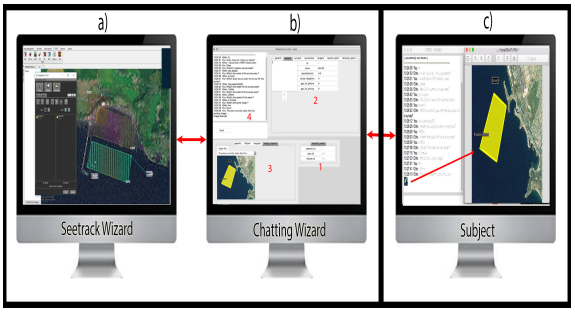

Figure 1: Experimental Set-Up, where a) SeeTrack Wizard, b) Chatting Wizard, and c) Subject console. Figure contains images from Seebyte's SeeTrack interface.

In order to establish the main actions and dialog act types for the system to perform, we recorded an expert planning a mission on SeeTrack whilst verbalizing the planning process and his reasoning. Similar human-provided rationalisation has been used to generate explanations of deep neural models for game play (Harrison et al., 2018). After analysing the expert video, we implemented a multimodal Wizard of Oz interface that is capable of sending messages, either in structured or free form as well as images of the plan. The Wizard Interface is made up of four windows (see Figure 2). The first has all the possible dialog acts (DA) the wizard can use together with predefined utterances for expedited responses. Once the DA is selected the predefined text appears in the chat window, from there the CW is able to modify as needed. The third window allows the CW to insert values (also referred to as 'slots') needed for the plan obtained through interaction from the user. Finally, the fourth window is for recording session details such as subject ID. The CW works collaboratively with the subject to develop a list of the necessary parameters that the SW needs to create the plan.

Each subject was given a short questionnaire to collect demographic information and instructions on how to approach the task of planning a mission using a conversational agent. A mission in our context is comprised of a nautical chart and a description of some objective that the subjects, together with the wizards, have to achieve. There are two main categories of missions A and B. The first (A) involves sending AUVs to survey areas of interest on the chart and is more time consuming. The second (B) category entails the reacquisition of a target, which overall can be achieved in less

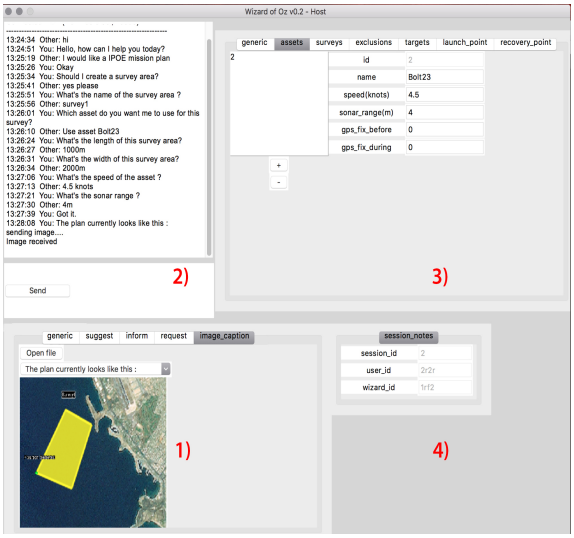

Figure 2: Wizard of Oz interface (Figure 1b) used by the Chatting Wizard, 1) dialog acts and structured prompts, 2) Chatting window, 3) State of the plan in the form of value-slots, 4) Session notes. Figure contains images from Seebyte's SeeTrack interface.

time. This is because surveying an area requires extra interactions between the operators and typically more spatial commands. Subjects were told that they had to plan both missions within a total time of one hour. Table 1 shows that the first mission took more time to plan in terms of actual planning time and number of user/system turns. This is normal because the first mission was of category A, which is by design harder, and the second mission of category B. The wording of the mission was very high level, so as not to prime the subjects. There are many elements that make up a plan (e.g. survey areas, exclusion zones) and therefore many variations are possible. Once both missions were completed, a post-questionnaire was administered to obtain their subjective opinion on the planning process. Subjects were told at the end that they were interacting with a wizard.

3.1 Subject Group

Planning missions for AUVs is a complex task, especially in the case of sophisticated software, such as SeeTrack. For this reason, we decided to focus our study mostly on expert users, who are familiar with SeeTrack and AUVs. We recruited human operators from industry (10 male, M_{age}=30), who all had some experience with SeeTrack.

4 Corpus Analysis

We collected 22 dialogs between the wizards and the subjects, which were analysed by a single annotator (an author). Figure 3 shows an example of a dialog interaction with corresponding dialog acts. We split our analysis into objective and subjective measures.

1. USER: 'move t2 200m west of t1'[inform]
2. SYSTEM:'Could you repeat that in different words? t1? t2?'[repeat]
3. USER:'move target2 200m west of target1'[inform]

Figure 3: Dialog excerpt and the corresponding image, displaying 2 targets, a launch and a recovery point [dialog act]. Figure contains images from Seebyte's SeeTrack interface.

4.1 Objective Measures

Dialog act types were adopted from the ISO (24617-2:2012) standard for dialog act annotation. Figure 4 gives the distribution of dialog acts, which were categorized into five groups:

1. **Generic** (conversational acts): wait, ack, affirm, yourwelcome, thankyou, bye, hello, repeat, praise, apology
2. **Inform** (for informing of values for slots): inform, negate, delete, create, correction, plan_complete, plan_mission
3. **Request** (for requesting information): request, enqmore
4. **Suggest** (for making suggestion): suggest
5. **Image** (for interacting with images): image_caption, show_picture

The most frequent user DA is the "inform" dialog act (54%), which informs the system about the plan slot values. This dialog act is also used for

Measures	Mission 1	Mission 2
# of turns	26.4(9.1)	13.1(4.4)
# of system utterances	51.4(21.0)	27.0(7.5)
# of user utterances	36.4(14.4)	19.7(8.1)
# of produced images	8.8(3.8)	5.0(1.3)
Time-on-Task (min)	26.3(0.005)	14.5(0.004)

Table 1: Measures per dialog [mean(sd)]. One turn comprises one system and one user turn.

utterances that instruct the system to move objects around the chart by referring either to the object's position or to nearby objects. 53% of these "inform" acts contain referring expressions (see lines 1 and 3 of Figure 3 for examples). In addition, it is clear that, due to the spatial nature of the tasks, the extra modality of plan images is key to successful planning, as reflected by the frequency of 'Image' dialog acts (around 16% of the total dialog acts). These DAs include the user requesting a plan image 'show_picture' or 'image_caption' where the system, either proactively or as a response to a user request, sends an image of the plan. The most used DA by the wizard was "ack" 30%, used for acknowledging information (e.g. "okay").

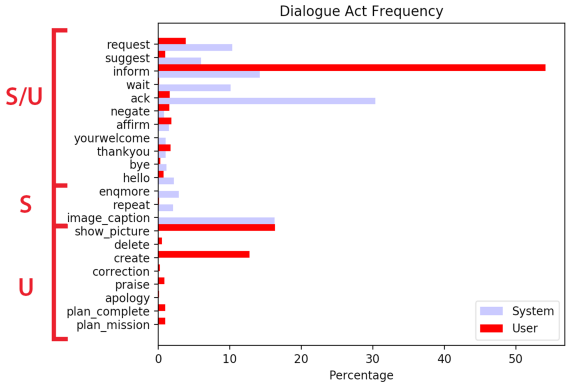

Figure 4: Dialog Act Frequency of the 22 DA types. "S" is for system only DA, "U" is for user only DA and "S/U" refers to DA that both the system and the user.

4.2 Quality of the plan

The subjective quality of the plans were measured by an expert, who has worked for years on planning missions, using a 5-point scale (see Table 2). The quality of a plan was measured according to the completeness, appropriateness and if it was operationally successful. At least 45% of the plans for both missions were measured "High Quality", with a greater number of lower rated plans for Mission 1. No correlation was found between time-on-task and quality of plans, however, subjective feedback indicates that subjects would have liked more time to improve their plans. The response time of the Wizard was slower than natural interaction (average 15sec), which is a typical issue in WoZ studies. The dataset contains plans of varying quality, which we hope will enable the system to learn better strategies for creating optimal plans, as well as, coping strategies. There is a medium-strong positive correlation of $r = 0.59$ (Spearman's Correlation) between the expertise of the subjects (as determined by the pre-questionnaire) and the quality of the plan for the first mission indicating that, perhaps unsurprisingly, the higher the expertise, the better the quality of plans.

Quality	Mission 1	Mission 2
Very High Quality	0%	9%
High Quality	45%	45%
Neutral	9%	36%
Low Quality	18%	9%
Very Low Quality	27%	0%

Table 2: The quality of all 22 plans measured by an expert using a 5-point Likert scale.

4.3 Subjective Measures

The post-task questionnaire measured the subjective scores for User Satisfaction (US), the pace of the experiment and the importance of multimodality. Specifically the following questions were asked on a 5-point Likert Scale:

- Q1: I felt that VERSO understood me well
- Q2: I felt VERSO was easy to understand
- Q3: I knew what I could say at each point in the interaction
- Q4: The pace of interaction of VERSO was appropriate
- Q5: VERSO behaved as expected
- Q6: It was easy to create a plan with VERSO
- Q7: From my current experience with using VERSO, I would use the system regularly to create plans
- Q8: The system was sluggish and slow to respond (reversed)
- Q9: The screen shots of the plan were useful
- Q10: The screen shots of the plan were sent frequently.

Mean US is 3.5 out of 5, calculated as an average of Q1-7, which are questions adapted from the PARADISE evaluation framework (Walker et al., 1997). Q8 reflects the speed of the interaction with the mean/mode/median as 4/4/4. This score is reversed and so these high scores indicate high perceived slowness. As mentioned above, this is a common problem with wizarding set-ups and will not be a problem for the final implemented system. Q9 and Q10 refer to the images sent and we can see from the mean/mode/median of 4.6/5/5 for Q9 that images were clearly useful but perhaps could be sent more frequently (3/4/3 for Q10). The users' preference for images of plans may be related to their cognitive styles being mostly spatial.

After both tasks, we collected perceived workload using NASA Task Load Index (NASA-TLX) (Dickinson et al., 1993), where low scores indicate low cognitive workload. Our mean Raw TLX score was 46/100 ($SD = 9.08$). This mean score is comparable to a study for remote controlling robots through an interface as reported in (Kiselev and Loutfi, 2012). Further analysis and data collection would be needed to understand the user workload with respect to interaction phenomena observed in the corpus.

4.4 Qualitative Feedback

Subjects were asked two open questions of what they liked or not about VERSO. An inductive, thematic analysis was done using grounded theory with open coding (Strauss, 1987). Themes identified include:

Theme 1 Suggestions for extra functionality: Due to delays some subjects were not sure if the program crashed. We had a dialog act "wait" but feedback indicated it would be better to have a visual indicator as well. Note, in the actual future working system, we will not have the same delays as in the WoZ experiment.

Theme 2 Chart meta-data: Some subjects (P5 most specifically) desired more meta-data on the plan images they were receiving when referring to an object. When performing spatial tasks on the chart, clear referring expressions are crucial and meta-data on the chart, such as entity names (as with the Map Task (Anderson et al., 1991) landmarks), would help establish grounded referring expressions.

In our case, some of the referring expressions were names decided on between the Wizard and the subject, e.g. survey3. However, if the subject uses such objects as points of reference, e.g. "place target1 near survey3", this can become problematic when the object ("survey3") could measure up to a mile width because the exact location for "target1" is ambiguous.

Theme 3 Mixed initiative & Handling multiple requests: The WoZ interface was designed as a mixed-initiative dialog system, capable of suggesting actions and the subjects seem to like this type of interaction. Also noted was the 'system's' ability to handle multiple requests in a single utterance, which will need to be implemented in the final system.

5 Discussion and Future work

This paper presents a two-wizard WoZ study for collecting data on a collaborative task, identifying the importance of mixed modalities and object referencing, for successful interaction during mission planning. Further data collection on Amazon Mechanical Turk using Open Street Maps will be conducted in order to reach a wider audience and compensate for the gender imbalance.

Deep learning methods have surpassed human performance in a variety of tasks and one crucial factor for this achievement is the amount of data used to tune these models. However, to be able to learn from limited amounts of data will be key in moving forward (Daugherty and Wilson, 2018).

In future work, the corpus described here will be used in the development of a mixed-initiative data-driven multimodal conversational agent, for planning missions collaboratively with a human operator. With the collected WoZ data, we can capture the main strategies of how to plan a mission and make data-driven simulations possible. Therefore, we can train a Reinforcement Learning agent on simulated dialogs that are fully data-driven with the reward function being derived from our subjects' preferences, optimizing for plan quality and speed. Moreover, supervised approaches that require less data to learn, such as the Hybrid Code Networks (HCN) (Williams et al., 2017), could be used for the creation of such a system. Finally, the system will be compared to a baseline in a further human evaluation study.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra. 2017. Visual dialog. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.

Paul R. Daugherty and H. James Wilson. 2018. *Human + machine : reimagining work in the age of AI / Paul R. Daugherty, H. James Wilson*. Harvard Business Review Press Boston, Massachusetts.

John Dickinson, Winston D. Byblow, and L.A. Ryan. 1993. Order effects and the weighting process in workload assessment. *Applied Ergonomics*, 24(5):357 – 361.

Rui Fang, Changsong Liu, Lanbo She, and Joyce Y. Chai. 2013. Towards situated dialogue: Revisiting referring expression generation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 392–402. Association for Computational Linguistics (ACL).

Brent Harrison, Upol Ehsan, and Mark O. Riedl. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 Conference on Artificial Intelligence, Ethics and Society*.

Andrey Kiselev and Amy Loutfi. 2012. Using a mental workload index as a measure of usability of a user interface for social robotic telepresence. *2nd Workshop of Social Robotic Telepresence in Conjunction with IEEE International Symposium on Robot and Human Interactive Communication 2012*.

Nikita Kitaev, Jin-Hwa Kim, Xinlei Chen, Marcus Rohrbach, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. *International Conference on Learning Representations*.

David Lane, Keith Brown, Yvan Petillot, Emilio Miguelanez, and Pedro Patron. 2013. *An Ontology-Based Approach to Fault Tolerant Mission Execution for Autonomous Platforms*, pages 225–255. Springer New York, New York, NY.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 994–1003. Association for Computational Linguistics.

Emilio Miguelanez, Pedro Patron, Keith E Brown, Yvan R Petillot, and David M Lane. 2011. Semantic knowledge-based framework to improve the situation awareness of autonomous underwater vehicles. *IEEE Transactions on Knowledge and Data Engineering*, 23(5):759–773.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. 2018. Mapping instructions to actions in 3d environments with visual goal prediction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2667–2678, Brussels, Belgium. Association for Computational Linguistics.

Yvan Petillot, Chris Sotzing, Pedro Patron, David Lane, and Joel Cartright. 2009. Multiple system collaborative planning and sensing for autonomous platforms with shared and distributed situational awareness. In *Proceedings of the AUVSI's Unmanned Systems Europe, La Spezia, Italy*.

Verena Rieser. 2008. *Bootstrapping Reinforcement Learning-based Dialogue Strategies from Wizard-of-Oz data*. Ph.D. thesis, Saarland University, Saarbruecken Dissertations in Computational Linguistics and Language Technology.

David Schlangen. 2016. Grounding, Justification, Adaptation: Towards Machines That Mean What They Say. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue (JerSem)*.

Anselm L Strauss. 1987. *Qualitative analysis for social scientists*. Cambridge University Press.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, ACL '98, pages 271–280, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 438–449. Association for Computational Linguistics.

Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 665–677, Vancouver, Canada. Association for Computational Linguistics.