Challenges in Designing Natural Language Interfaces for Complex Visual Models

Henrik Voigt[1,2], Monique Meuschke[1], Kai Lawonn[1] and Sina Zarrieß[2]
[1]University of Jena [2]University of Bielefeld
[1]first.last@uni-jena.de [2]first.last@uni-bielefeld.de

Abstract

Intuitive interaction with visual models becomes an increasingly important task in the field of Visualization (VIS) and verbal interaction represents a significant aspect of it. Vice versa, modeling verbal interaction in visual environments is a major trend in ongoing research in NLP. To date, research on Language & Vision, however, mostly happens at the intersection of NLP and Computer Vision (CV), and much less at the intersection of NLP and Visualization, which is an important area in Human-Computer Interaction (HCI). This paper presents a brief survey of recent work on interactive tasks and set-ups in NLP and Visualization. We discuss the respective methods, show interesting gaps and conclude by suggesting neural, visually grounded dialogue modeling as a promising potential for NLIs for visual models.

1 Introduction

In recent years, research in NLP has become more and more interested in data sets, tasks and models that pair Language and Vision, cf. work on image Captioning (Vinyals et al., 2015; Herdade et al., 2019; He et al., 2020), Visual Question Answering (Antol et al., 2015; Goyal et al., 2017; Kazemi and Elqursh, 2017), or Instruction Following and -Generation in visual domains (Fried et al., 2017, 2018). This new area is generally called Vision & Language (Mogadala et al., 2019), but it is actually based mostly on combining methods from NLP (like e.g. language models) and Computer Vision, e.g. visual analysis and recognition models for encoding visual input like images. Methods and models from the research area of Visualization – which investigates solutions for modelling, exploring, analyzing and communicating data by using visual technologies and can be seen as the field of visual synthesis – are, to the best of our knowledge, less well known in the NLP community.

In the VIS community, interaction with visual models plays an important role and natural language interaction represents a big part of it (Bacci et al., 2020; Kumar et al., 2020; Srinivasan et al., 2020). Natural Language Interfaces (NLIs) that support interactive visualizations based on language queries have found increasing interest in recent research (Narechania et al., 2020; Yu and Silva, 2020; Fu et al., 2020). However, from an NLP point of view, the methods applied in these recent interfaces, mostly rely on established methods for implementing semantic parsers that map natural language instructions to symbolic data base queries, which are consecutively visualized by a visualization pipeline. In this paper, we argue that there is space for further support of intuitive interaction with visual models using state-of-the-art NLP methods that would also pose novel and interesting challenges for both domains.

We focus this brief overview on a selection of methods for modeling interaction in the fields of NLP and VIS based on recent submissions to the top conferences ACL, EACL, VIS and EuroVIS. First, we briefly describe how interaction is understood in the respective fields (Section 2). We provide a short overview of recent, state-of-the-art systems related to interaction with visual models or in visual environments (Section 3). Finally, we discuss potential research gaps and challenges that could be addressed in future work on modelling interaction with visual models (Section 4). As interaction is a major research topic in both NLP and VIS, we do not aim for a complete survey, but we hope to make readers from both communities aware that there could be fruitful directions for collaboration.

2 Interaction in NLP and VIS

We refer to *interaction* as the "mutual determination of behaviour" between different entities, like humans, digital agents or interfaces, following

Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 66–73

Hornbæk and Oulasvirta (2017). Work on interaction in NLP typically investigates verbal communication between human dialogue partners and models dialogue systems that interact with users via natural language, but also recognizes the fact that verbal communication typically happens in combination with other modalities, like touch, movements and gestures in embodied dialogue or gaze and visual stimuli in visual dialogue (Cuayáhuitl et al., 2015). In HCI and VIS, interaction via multiple modalities plays a very prominent role, i.e. involves gestures, movements, different input controllers, screens, gazes, modalities and more. The interaction between a user and a visual model is a key aspect of many VIS tasks and applications and impacts on the user evaluation of a visual model to a significant degree (Yi et al., 2007; Tominski, 2015; Figueiras, 2015).

2.1 Interaction in VIS

Generally speaking, the field of VIS is interested in the development of techniques for creating visual models (Brehmer and Munzner, 2013; Liu et al., 2014; Amar et al., 2005). A visual model is data that is mapped into a visually perceivable space by representing concepts in the data through visual concepts to make them easily perceivable and understandable by humans. This supports research and education in many aspects as well as data exploration and understanding of big data sets. Research on interaction in VIS often addresses the design of appropriate human-computer interfaces and the abilities they need to offer for interacting with a visual model. Natural Language Interfaces (NLIs), in this context, can be seen as one possible solution of enabling interaction with a visualization. Dimara and Perin (2019) provide a comprehensive study on how interaction is seen in VIS by defining it as "the interplay between a person and a data interface involving a data-related intent, at least one action from the person and an interface reaction that is perceived as such". The authors deliberately distinguish their view from the HCI definition of interaction as stated in Hornbæk and Oulasvirta (2017), by making the importance of the data related intent of the user the focus in VIS. As a conclusion, the authors observe that approaches towards interaction in VIS currently lack two points, i.e. *flexibility* and a better understanding of the *user goal*. The lack of these currently leads to interfaces that are too predictable, unsatisfying in their capacities to act or risk misdirected interactions with visual models. Despite that, Hornbæk and Oulasvirta (2017) and Dimara and Perin (2019) both argue that interaction foremost represents a form of *dialogue* which the authors evaluate in terms of its "naturalness" and its mutual "strong sense of understanding".

This highlights the point that interaction with a visual model is fundamentally conceived as a multi-modal process that leverages various different interface modalities for communication and information exchange. As discussed below, from an NLP perspective, interactions with systems in VIS can be seen as multi-modal dialogues between a system and a user having data-related goals.

2.2 Interaction in NLP

Work in NLP often aims at understanding and modeling how dialogue partners collaborate and achieve common ground by exchanging verbal utterances (potentially in combination with different modalities, like e.g. vision). This typically involves language understanding, dialogue management (reasoning over latent user goals) and language generation (Young et al., 2010, 2013). Recent work on dialogue has turned more and more to so-called neural end-to-end-dialogue systems that do not separate processes of understanding, reasoning and generation, and aim for more flexibility and adaptiveness (Santhanam and Shaikh, 2019). Santhanam and Shaikh (2019) distinguish between goal-driven and open dialogue systems as they address fundamentally different interaction and evaluation set-ups. Goal- or task-oriented systems are typically designed towards helping the user to achieve a very specific goal in a given context. For instance, in instruction-following and -generation (Fried et al., 2017, 2018), a user or system needs to reach a specific position in an environment by following navigation instructions. Here, the interaction is often asymmetric in the sense that the modalities to be used by the partners are very restricted (the instruction follower acts, the giver speaks). Open-domain dialogue systems, like Li et al. (2017); Adiwardana et al. (2020) are not bound to a goal and therefore require a high awareness of context, personality and variety of the dialogue system as Santhanam and Shaikh (2019) point out.

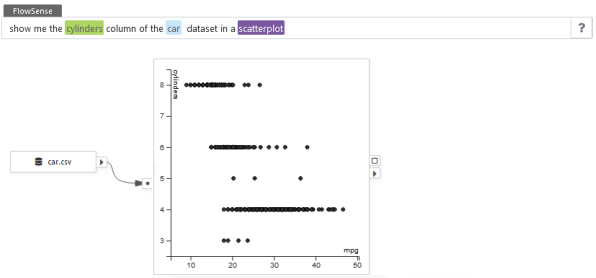

Figure 1: Goal-oriented NLI as used in Yu and Silva (2020), created from: https://visflow.org/demo/

3 Existing Work

3.1 Natural Language Interfaces in VIS

A range of recent papers have looked into integrating NLP in VIS systems, by implementing NLIs that translate a natural language query to a visualization command in some programming language. This allows users to "talk to some dataset", as illustrated in Figure 1. Existing NLIs are applied in systems that create and manipulate, e.g., chart visualizations (Shao and Nakashole, 2020). Similar interfaces are proposed in Narechania et al. (2020); Huang et al. (2019); Yu and Silva (2020); Fu et al. (2020); Chowdhury et al. (2021); Setlur et al. (2016)

These existing NLIs are mostly applied in the field of visual analytics. Here, the user has a concrete goal in mind, i.e. some manipulation of the underlying data table (e.g. aggregation, filtering). Dimara and Perin (2019) point out that the exact understanding of the users' goal is important in these interfaces and one current approach for improving the inference of the users' intent is to predict it based on activity logs. Setlur and Kumar (2020)'s work suggests that the handling of vague subjective modifiers in utterances can be improved by using sentiment analysis techniques.

In terms of NLP methods, these interfaces are mostly based on manually engineered grammars that parse user input to queries and then generate an appropriate visualization output, as e.g. in (Yu and Silva, 2020). These grammars are relatively easy to set-up even for non-experts (of NLP) and do not require large amounts of training data, as most state-of-the-art dialogue systems developed in NLP. An important limitation of this approach, however, is that such semantic grammars are designed to translate directly between a given user query in natural language and some underlying data query language like e.g. SQL. This means that, in longer multi-turn interactions between a user and a system, users are not able to formulate short, intuitive queries that implicitly refer to the context (e.g. "now, make this a bit bigger" where "this" refers to an aspect of the visual model discussed in the preceding context) or multi-modal queries (e.g. "increase the volume of this particle here" *while user points to a region on the screen*). Finally, and most importantly, they assume that the user can precisely formulate or describe the action or manipulation that is needed to obtain a certain visualization or information from the visual model, as shown for instance in Figure 1.

Beyond NLIs for visual analytics (Narechania et al., 2020; Yu and Silva, 2020), we see further potential for other NLP methods in visualization tasks that require more than plots of data for a specific, precisely formulated goal. In visual exploration, the goal is typically rather vague and developed during the exploration process itself, in an iterative fashion while interacting with the system. Moreover, applications in augmented and virtual reality entail new possibilities of immersive experiences and interaction for supporting performance as in (Butcher et al., 2020) or to enhance retention (Yang et al., 2020), which clearly includes multiple modalities. We argue that users of visual models in such exploratory setups could greatly benefit from natural language interaction, if the NLI would allow for more context-sensitive and situated querying of the model. Moreover, beyond querying, we expect that users would highly appreciate verbal system feedback or suggestions and explanations (see Figure 2). Ideally, this back-and-forth between the user and the system should support the user not only in realizing his goal, but also in establishing his goal or refining his initially vague goal.

Thus, we currently see a lot of interest in systems that enable interaction with an underlying data set via natural language, but the query-based approach used in many NLIs still seems to lack flexibility.

3.2 Language & Vision

A lot of recent work in NLP tackles dialogue modelling (Shuster et al., 2020; Qin et al., 2020; Ham et al., 2020; Rameshkumar and Bailey, 2020) or question answering (Baheti et al., 2020; Liu et al., 2020). Goal-based dialogue covers navigation Zhu et al. (2020), manipulation (Jayannavar et al., 2020) or classical information presentation tasks (Andreas et al., 2020). A central problem in these mod-

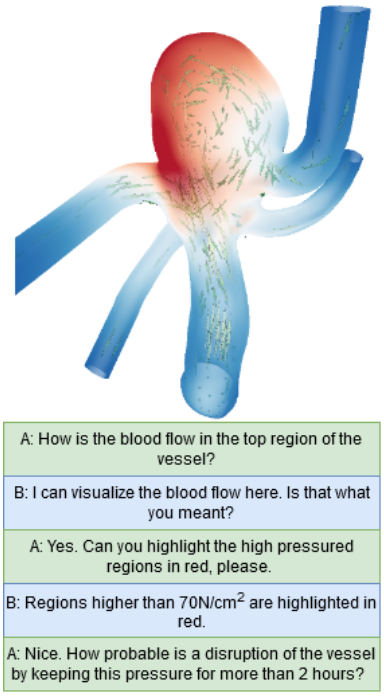

Figure 2: Visual exploration of a visual model of an aneurysm and an exemplary mixed-initiative dialogue

els is the fact that at each point in an interaction, there is uncertainty with respect to the understanding of the user goal. The predominant approach to handle reasoning under uncertainty is (Deep) Reinforcement Learning (RL) where an agent learns a dialogue policy (Jaques et al., 2020; Li et al., 2020, 2016). RL optimizes the utterance understanding and generation in the system with respect to a certain reward function in the given environment.

Next to these improvements on the level of dialogue modeling, recent developments in Language & Vision focus on grounding verbal utterances in visual inputs as, for instance, in visual question answering (Huang et al., 2020; Khademi, 2020). Visual dialogue (Das et al., 2017; Wang et al., 2020) extends the dialogue modelling task to the visual modality. Mixed-initiative visual dialogue, as e.g. in Ilinykh et al. (2019), aims at modeling interactions in which both dialogue partners can talk and act, which could be an interesting setting for visual exploration tasks. We believe that these successes in neural dialogue modelling and the integration of different modalities as in visual dialogue can lead to new possibilities for interactive systems in VIS, as we will discuss below.

4 Future Work

Uncertainty We believe that a fruitful direction for more flexible NLP-based systems in VIS is to look at scenarios where users might not have a concrete manipulation task or goal in mind, but want to explore a complex visual model. Numerous applications, such as in medicine (Meuschke et al., 2016, 2017) or cultural-technical scenarios (Lawonn et al., 2016), require the visual exploration of complex models. Figure 2 shows an example of a 3D-mesh of an aneurysm and a corresponding, made-up dialogue that would support the user in exploring the model (Meuschke et al., 2018). On the NLP side, this setting involves a high degree of uncertainty. The user investigates a certain region of the model she is interested in, develops an understanding of the visual landscape and/or just learns how to handle it best. Thus, we argue that complex visual models like in Figure 2 probably call for different and more flexible types of interactions, as compared to NLIs discussed in Section 3.1. A user analyzing a barplot might be interested in minima, maxima, trends or outliers, which correspond to fixed goals. In contrast to that, a neurosurgeon inspecting an aneurysm in virtual reality is much more interested in the *how* than in the *what* and the goal might not be precisely formulated beforehand by the surgeon but evolving through the back and forth of the interaction with the visual model.

An important question that arises from that is, if users would really use natural language for exploring a visual model or if they would rather prefer the use of e.g. a controller or touch gesture. For many cases this might indeed be true, but we argue, that certain scenarios in visual exploration especially demand for verbal problem solution: recommendation of possibilities ("show me how to reach the largest vessel from here", "how to achieve a blood pressure increase in this region"), problem-solution-suggestions, tutorial-like action descriptions the user has to mimic, e.g. in educational scenarios or future state simulations that are highly hypothetical ("what would the vessel behave if we changed the blood flow drastically to ..."). These cases are in fact simulations of possible solutions helping the user to visualize and explore the solution space, which are much more convenient and intuitive expressed using natural language which can be supported by strong dialogue models that adapt to the context.

Visual Grounding Visual language grounding in these scenarios captures not only the grounding of words into the scene (e.g. "vessel") but also the grounding of movements and gazes like pointer gestures (e.g. "here") which shows that the interplay of context awareness and multi-modal visual grounding are prerequisites for dialogue that humans would describe as "intuitive" and flexible. In contrast to systems like (Narechania et al., 2020) which are restricted to visual output and systems like (Adiwardana et al., 2020) which respond verbally, dialogue systems in visual models should be able to handle multi-modal responses, as illustrated in Figure 2. The highlighting, scaling, coloring or fading of certain visual properties is an important part of the response which not only contains text but rather text and a visual action combined.

Mixed-Initiative Dialogue Figure 2 shows an example for a collaborative, mixed-initiative interaction where the the dialogue flow is not entirely centered on user queries. Tang et al. (2020) used dialogue modelling for generating visual story lines in collaboration with a user and obtained promising results in leveraging visual exploration scenarios. In contrast to concise, goal-based visual analytics (see Section 3.1), hard-coded grammars might be too restricted to handle the high uncertainty and the complex underlying reasoning in explorative scenarios. The visual analytics task differs from the visual exploration task considering the fact that it is not driven by a concise goal. Introducing mixed-initiative dialogue in visual exploration enhances the users' ability of communicating uncertainties and supports experimenting and iterative engagement with the environment as applied in active visual problem solving or in educational settings. When no concrete goal can be formulated, a NLI has to adapt to the user and present contextual information like hints, explanations or react to expressed uncertainties (e.g. 'What does this blue region here show me?', 'How can I slice the vessel and investigate the thickness?', 'How does this spot evolve over time?') which is a form of guidance, an evolving field in visual analytics (Ceneda et al., 2020). This interaction also is not bound to text-language interaction, but furthermore accommodates gestures, movements or glances and therefore can be categorized as multi-modal. Here, recent advances in NLP could extend the interface flexibility by providing better context-awareness using visually grounded dialogue techniques and contributing to solve the user goal inference problem by re-framing the task as an iterative goal alignment task executed via mixed-initiative dialogue. We think that especially mixed-initiative dialogue would be a challenging but very promising direction and well-suited for inferring user intentions in complex VIS settings because of the usage of direct user feedback and iterative alignment. Furthermore, mixed-initiative dialogue methods could support the setup of user-centred evaluation of more complex visualization techniques, such as in (Lawonn et al., 2014). In sum, we argue that the role of NLP in interfaces with visual models is to enrich the dialogue between a system and a user (Dimara and Perin, 2019) with more flexible and intuitive ways of dialogue that might include touch-based or controller-based interaction.

5 Conclusion

In this paper, we gave a brief introduction on how interaction is understood and modelled in the fields of NLP and VIS. We found that existing work on NLIs in the Visualization domain heavily relies on query-based interactions. We argued that for interacting with highly complex visual models these strict interaction protocols might not be sufficient. Recent developments in Language & Vision investigate dialogue in visual contexts and reach promising results. We believe that this holds interesting research gaps for future work in integrating different variations of NLP-backed dialogue methods into visualizations enabling multi-modal interaction with visual models.

Acknowledgments

We thank the Michael Stifel Center Jena for the funding of this work, which is part of the "A Virtual Werkstatt for Digitization in the Sciences" project funded by the Carl Zeiss Foundation (062017-02).

References

D. Adiwardana, Minh-Thang Luong, D. So, J. Hall, Noah Fiedel, R. Thoppilan, Z. Yang, Apoorv Kulshreshtha, G. Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *ArXiv*, abs/2001.09977.

Robert Amar, James Eagan, and John Stasko. 2005. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 111–117. IEEE.

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics (TACL)*, 8.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Franscesca Bacci, Federico Maria Cau, and L. D. Spano. 2020. Inspecting data using natural language queries. In *ICCSA*.

Ashutosh Baheti, Alan Ritter, and Kevin Small. 2020. Fluent response generation for conversational question answering. *arXiv preprint arXiv:2005.10464*.

M. Brehmer and T. Munzner. 2013. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19:2376–2385.

Peter William Scott Butcher, Nigel W John, and Panagiotis D Ritsos. 2020. Vria: A web-based framework for creating immersive analytics experiences. *IEEE Transactions on Visualization and Computer Graphics*.

Davide Ceneda, N. Andrienko, G. Andrienko, T. Gschwandtner, S. Miksch, Nikolaus Piccolotto, T. Schreck, M. Streit, Josef Suschnigg, and C. Tominski. 2020. Guide me in analysis: A framework for guidance designers. *Computer Graphics Forum*, 39:269 – 288.

I. Chowdhury, Abdul Moeid, Enamul Hoque, M. Kabir, Md. Sohorab Hossain, and M. M. Islam. 2021. Designing and evaluating multimodal interactions for facilitating visual analysis with dashboards. *IEEE Access*, 9:60–71.

H. Cuayáhuitl, Kazunori Komatani, and Gabriel Skantze. 2015. Introduction for speech and language for interactive robots. *Comput. Speech Lang.*, 34:83–86.

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

Evanthia Dimara and Charles Perin. 2019. What is interaction for data visualization? *IEEE transactions on visualization and computer graphics*, 26(1):119–129.

A. Figueiras. 2015. Towards the understanding of interaction in information visualization. *2015 19th International Conference on Information Visualisation*, pages 140–147.

Daniel Fried, Jacob Andreas, and Dan Klein. 2017. Unified pragmatic models for generating and following instructions. *arXiv preprint arXiv:1711.04987*.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pages 3314–3325.

Siwei Fu, Kai Xiong, X. Ge, Y. Wu, Siliang Tang, and W. Chen. 2020. Quda: Natural language queries for visual data analytics. *ArXiv*, abs/2005.03257.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and D. Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using gpt-2. ACL.

Sen He, Wentong Liao, H. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault. 2020. Image captioning through image transformer. *ArXiv*, abs/2004.14231.

Simao Herdade, Armin Kappeler, K. Boakye, and J. Soares. 2019. Image captioning: Transforming objects into words. In *NeurIPS*.

Kasper Hornbæk and Antti Oulasvirta. 2017. *What Is Interaction?*, page 5040–5052. Association for Computing Machinery, New York, NY, USA.

Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Aligned dual channel graph convolutional network for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7166–7176, Online. Association for Computational Linguistics.

Zhaosong Huang, Ye Zhao, Wei Chen, Shengjie Gao, Kejie Yu, Weixia Xu, Mingjie Tang, Minfeng Zhu, and Mingliang Xu. 2019. A natural-language-based visual query approach of uncertain human trajectories. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1256–1266.

N. Ilinykh, Sina Zarrieß, and D. Schlangen. 2019. Meetup! a corpus of joint activity dialogues in a visual environment. *ArXiv*, abs/1907.05084.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2020. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3985–4003, Online. Association for Computational Linguistics.

Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. 2020. Learning to execute instructions in a Minecraft dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2589–2602, Online. Association for Computational Linguistics.

V. Kazemi and A. Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *ArXiv*, abs/1704.03162.

Mahmoud Khademi. 2020. Multimodal neural graph memory networks for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7177–7188, Online. Association for Computational Linguistics.

Abhinav Kumar, Jillian Aurisano, B. D. Eugenio, and A. Johnson. 2020. Intelligent assistant for exploring data visualizations. In *FLAIRS Conference*.

Kai Lawonn, Alexandra Baer, Patrick Saalfeld, and Bernhard Preim. 2014. Comparative evaluation of feature line techniques for shape depiction. In *Proc. of Vision, Modeling and Visualization*, pages 31–38, Darmstadt.

Kai Lawonn, Erik Trostmann, Bernhard Preim, and Klaus Hildebrandt. 2016. Visualization and extraction of carvings for heritage conservation. *IEEE transactions on visualization and computer graphics*, 23(1):801–810.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Ziming Li, Julia Kiseleva, and Maarten de Rijke. 2020. Rethinking supervised learning and reinforcement learning in task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3537–3546, Online. Association for Computational Linguistics.

Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, and Nan Duan. 2020. RikiNet: Reading Wikipedia pages for natural question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6762–6771, Online. Association for Computational Linguistics.

Shixia Liu, Weiwei Cui, Yingcai Wu, and Mengchen Liu. 2014. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393.

Monique Meuschke, Wito Engelke, Oliver Beuing, Bernhard Preim, and Kai Lawonn. 2017. Automatic viewpoint selection for exploration of time-dependent cerebral aneurysm data. In *Bildverarbeitung für die Medizin*, pages 352–357. Springer.

Monique Meuschke, Kai Lawonn, Benjamin Köhler, Uta Preim, and Bernhard Preim. 2016. Clustering of aortic vortex flow in cardiac 4d pc-mri data. In *Bildverarbeitung für die Medizin*, pages 182–187. Springer.

Monique Meuschke, Samuel Voß, Bernhard Preim, and Kai Lawonn. 2018. Exploration of blood flow patterns in cerebral aneurysms during the cardiac cycle. *Computers & Graphics*, 72:12–25.

Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2019. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *arXiv preprint arXiv:1907.09358*.

Arpit Narechania, Arjun Srinivasan, and John Stasko. 2020. Nl4dv: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics*.

Libo Qin, Xiao Xu, Wanxiang Che, Yue Zhang, and Ting Liu. 2020. Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6344–6354, Online. Association for Computational Linguistics.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134, Online. Association for Computational Linguistics.

Sashank Santhanam and Samira Shaikh. 2019. A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions. *arXiv preprint arXiv:1906.00500*.

V. Setlur, S. Battersby, Melanie Tory, R. Gossweiler, and Angel X. Chang. 2016. Eviza: A natural language interface for visual analysis. *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*.

Vidya Setlur and Arathi Kumar. 2020. Sentifiers: Interpreting vague intent modifiers in visual analysis using word co-occurrence and sentiment analysis. *arXiv preprint arXiv:2009.12701*.

Yutong Shao and Ndapa Nakashole. 2020. ChartDialogs: Plotting from Natural Language Instructions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3559–3574, Online. Association for Computational Linguistics.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online. Association for Computational Linguistics.

A. Srinivasan, J. Stasko, Daniel F. Keefe, and Melanie Tory. 2020. How to ask what to say?: Strategies for evaluating natural language interfaces for data visualization. *IEEE Computer Graphics and Applications*, 40:96–103.

Tan Tang, Renzhong Li, Xinke Wu, Shuhan Liu, Johannes Knittel, Steffen Koch, Thomas Ertl, Lingyun Yu, Peiran Ren, and Yingcai Wu. 2020. Plotthread: Creating expressive storyline visualizations using reinforcement learning. *IEEE Transactions on Visualization and Computer Graphics*.

C. Tominski. 2015. Interaction for visualization. In *Interaction for Visualization*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Yue Wang, Shafiq R. Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and S. Hoi. 2020. Vd-bert: A unified vision and dialog transformer with bert. In *EMNLP*.

Fumeng Yang, Jing Qian, Johannes Novotny, David Badre, Cullen Jackson, and David Laidlaw. 2020. A virtual reality memory palace variant aids knowledge retrieval from scholarly articles. *IEEE Transactions on Visualization and Computer Graphics*.

J. S. Yi, Y. Kang, J. Stasko, and J. Jacko. 2007. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13:1224–1231.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Bowen Yu and C. T. Silva. 2020. Flowsense: A natural language interface for visual data exploration within a dataflow system. *IEEE Transactions on Visualization and Computer Graphics*, 26:1–11.

Wang Zhu, Hexiang Hu, Jiacheng Chen, Zhiwei Deng, Vihan Jain, Eugene Ie, and Fei Sha. 2020. BabyWalk: Going farther in vision-and-language navigation by taking baby steps. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2539–2556, Online. Association for Computational Linguistics.