Being Data-Driven is Not Enough: Revisiting Interactive Instruction Giving as a Challenge for NLG

Sina Zarrieß and **David Schlangen**
Dialogue Systems Group // CITEC // Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
{sina.zarriess,david.schlangen}@uni-bielefeld.de

Abstract

Modeling traditional NLG tasks with data-driven techniques has been a major focus of research in NLG in the past decade. We argue that existing modeling techniques are mostly tailored to textual data and are not sufficient to make NLG technology meet the requirements of agents which target fluid interaction and collaboration in the real world. We revisit interactive instruction giving as a challenge for data-driven NLG and, based on insights from previous GIVE challenges, propose that instruction giving should be addressed in a setting that involves visual grounding and spoken language. These basic design decisions will require NLG frameworks that are capable of monitoring their environment as well as timing and revising their verbal output. We believe that these are core capabilities for making NLG technology transferrable to interactive systems.

1 Introduction

The past decade has seen substantial progress in data-driven methods for natural language generation (NLG). It is now widely agreed that data-driven techniques are needed to obtain NLG systems that are adaptive and human-like (Belz, 2008), domain-independent (Wen et al., 2016), and – with recent methods from vision & language cf. (Bernardi et al., 2016) – suitable for agents that interact with humans in a physical environment (such as dialogue systems or robots) (Kazemzadeh et al., 2014). Despite this progress, however, data-driven NLG is rarely used in current real-world interactive systems, where more traditional (template-based) approaches for generating verbal output still persist.

In this paper, we argue that existing methods in data-driven modeling for NLG are heavily tailored to textual data and, therefore, fail to meet the requirements of dialogue systems, social agents or robots which target fluid interaction and collaboration in the real world. In the traditional view, the NLG task is usually framed as follows: given some non-verbal piece of data as input (e.g. sensor data, a meaning representation, facts from a knowledge base), the system needs to decide *what* to say (do content selection, text or sentence planning, micro-planning), and *how* to say it (do lexicalization, surface realization), cf. (Reiter and Dale, 1997). While recent data-driven systems have mostly overcome previous modular architectures that assigned these decisions to separate components in the processing pipeline (Konstas and Lapata, 2013), they still follow basic assumptions related to how the system processes its non-linguistic input and verbal output:

- static input: NLG systems are usually trained to map a given input to some written output, meaning that the environment does not change while the system is producing output
- perfect input: NLG systems are often trained on perfect representations of an environment or a knowledge base
- one-shot output: NLG systems do not need to monitor whether the listener has actually understood the output, strategies that are frequent in conversation (revision, correction, installments) do not have to be considered
- no temporal dimension: NLG systems assume that their output is not immediately consumed, i.e. it does not need to be packaged or timed (e.g. a text is produced as a whole)

Proceedings of the Workshop on NLG for Human–Robot Interaction, pages 27–31,
Tilburg, The Netherlands, November 8, 2018. © 2018 Association for Computational Linguistics

These assumptions are convenient when framing NLG tasks as machine learning problems (e.g. as ranking, classification or sequence-to-sequence learning), but they are highly problematic for interactive systems. To illustrate this point, we propose to revisit instruction giving as a challenge for data-driven NLG in interactive systems: here, a human instruction follower (IF) and an agent as the instruction giver (IG) have to achieve a common goal in a visual environment (e.g. find a route or treasure, assemble an object). The IG knows how to complete the task (e.g. where the treasure is, how the object looks like) but cannot affect the environment. The IF can affect the environment and the objects in it, but needs the IG's instructions to achieve the goal. In the context of the GIVE challenge (Byron et al., 2007), this setting has received considerable attention in the NLG community for some time (Byron et al., 2009; Striegnitz et al., 2011), but has not been developed further since then.

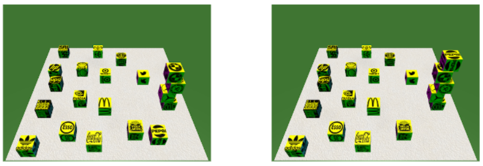
Generally, we believe that future approaches to instruction giving in NLG should extend GIVE along the following dimensions, in order to enable transfer of NLG technology to real-world applications like robots or dialogue systems:

- vision: generating instructions from a low-level visual representation of the environment, i.e. without perfect access to visually present objects and their properties
- spoken language: generating spoken instructions, such that the IF's non-verbal actions can happen concurrently with the IG's verbal utterances
- timing and information delivery: going beyond traditional NLG approaches focussing on content selection and/or surface realization, and move to real-time incremental processing that captures the affordances of spoken language and fluid interaction

In the following, we will show that these points constitute considerable challenges for the state-of-the-art in data driven NLG research and outline directions for how they could be addressed.

2 Visual grounding for instructions

A fundamental design decision in GIVE was to use a virtual environment such that the NLG systems had access to a perfect symbolic representation of the visually present objects and their properties. In the meantime, a lot of research in human-robot interaction has be done on modeling instructions in more realistic visual environments, though this community has often focussed on grounding verbal instructions to robot actions, cf. (Chai et al., 2018). Bisk et al. (2016) have proposed a nice formulation of a move-by-move instruction following task in an object assembly domain (see Figure 1): given an image of the current state of an environment (left image) and a verbal instruction, the task is to predict the target state of the environment after executing the instruction (right image). This move-by-move setting abstracts away from the internal action representations of a robot and also from general aspects of planning.



place the block that is to the right of the stella block as the highest block on the board. it should be in line with the bottom block .

Figure 1: Instruction example in the BLOCKS data set (Bisk et al., 2018)

We believe that this set-up is promising for NLG as well, where the task would be to generate a verbal instruction that enables the IF to execute a particular action or achieve a state change of the environment, while the system (the IG) is given the current and the goal state of an environment as an image. This would be natural extension of existing language generation systems that are able to generate descriptions of real-world images (Bernardi et al., 2016), or referring expressions to objects in real-world images (Yu et al., 2017). At the same time, it would require systems to go beyond the commonly used CNN-LSTM architecture (Vinyals et al., 2015; Devlin et al., 2015; Mao et al., 2016; Yu et al., 2017) as these currently only map visual representations of single images or objects to verbal output. Instead, a visually grounded instruction generation system needs to reason about expressions that relate the current visual state to a target state, such as *place the block to the right (source state) as the highest block on the board (target state)* in Figure 1.

Conceptually, the problem of generating instructions in object assembly domains is similar to generating relational referring expressions which have been a notorious challenge for referring expression generation in general (Krahmer and Van Deemter, 2012). Relational expressions are also challenging for neural architectures (Hudson and Manning, 2018), and grounding (understanding) of relational referring expressions has been addressed in some recent work (Cirik et al., 2018; Hu et al., 2017) following the idea of modular networks based on syntactic structures (Andreas et al., 2016). However, none of these models is designed for generating relational structures in verbal expressions, such as instructions.

3 Spoken language dynamics

From research on situated spoken dialogue, it is well known that spoken and written language bear very different affordances. In spoken communication, listeners react, both non-verbally and verbally, to what speakers are saying, while they are saying it; and speakers adapt what they are saying, based on the reactions (or lack thereof) that they get, while they are speaking. The field of Conversation Analysis (see (Stivers and Sidnell, 2012) for a recent overview) and, taking up and further developing some of their ideas, the work of Herbert (Clark, 1996) has done much to shed light on the intricate strategies that interactants follow to co-construct dialogue in this way.

Figure 2 illustrates some prominent strategies that speakers use to achieve task success in spoken communication, with an instruction giving example taken from our PentoRef data (Zarrieß et al., 2016). Here, the IF has to assemble an object out of Pentomino pieces while the IG observes his actions over a camera feed. During a time span of approximately 30 seconds, the IG produces 18 short utterances in total that instruct the IF what to do next (e.g. *turn to the left*), confirm the IF's action (*exactly*), or repair what she is currently doing (*to the left, this is to the right*). Also, interestingly, the final step of the instruction (i.e. how to put the target piece to its target location, image 10-12 in Figure 2) is left underspecified by the IG as it is obvious to the IF how to complete the task. This level of coordination and adaptation between speakers and listeners is impossible in written communication where verbal and non-vernal actions cannot happen concurrently.

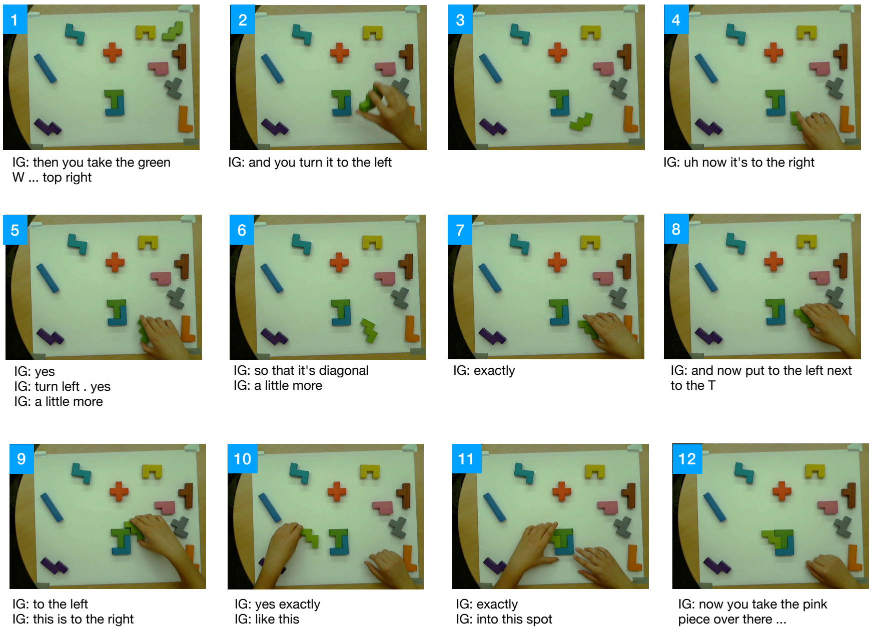
Unfortunately, most research on data-driven NLG still focusses entirely on written text or typed utterances, even in the domain of dialogue, as existing platforms and workflows for data collection are radically more efficient for text as compared to speech. Also the GIVE setting used typed communication. An interesting pilot study on a spoken version of the GIVE challenge was carried out by (Striegnitz et al., 2012) who found that interactions between participants were faster, more natural and rich of conversational phenomena (e.g. installments) that cannot be observed in text or typed chat. Another promising the direction here is the platform developed by (Manuvinakurike and DeVault, 2015), which extends the standard procedure for collecting chat interactions via crowdsourcing to spoken dialogue.

4 Monitoring, timing, revision

When facing uncertainty through visual grounding and dynamics through spoken language, NLG systems will need to address a range of decisions that, currently, completely fall out of the scope of research in this area. In the interactive world, NLG needs to monitor the listener's reaction in real-time and be able to quasi-continuously decide *when* to produce verbal output and *how* to potentially revise previous or future output. Thus, in order to generate fluid instructions as in the interaction shown in Figure 2, it is precisely the combination of *when* to speak and *what* to say that matters: an utterance that is appropriate at a particular point in time, might already be perceived as inappropriate or confusing shortly after.

To the best of our knowledge, aspects of monitoring and timing have not been addressed in data-driven NLG frameworks, though incremental processing has been shown to be highly effective in experimental or rule-based settings, cf. (Skantze and Hjalmarsson, 2013; Skantze et al., 2014; Buß and Schlangen, 2010). In the dialogue community, specific tasks that involve timing have been modelled in a data-driven way, such as barge-in detection (Selfridge et al., 2013), *end-of-utterance* detection (Raux and Eskenazi, 2012; Maier et al., 2017)), or *turn-taking* (Skantze, 2017) .

Even less work has been carried out on NLG systems that are able to produce revision, repair or correction utterances which can be essential to achieve task success, as shown in Figure 2. In (Zarrieß and Schlangen, 2016), we have explored an installment-based approach in a referring expression generation system for objects in real-world images, and found that even simple,



IG: then you take the green W ... top right

IG: and you turn it to the left

IG: uh now it's to the right

IG: yes
IG: turn left . yes
IG: a little more

IG: so that it's diagonal
IG: a little more

IG: exactly

IG: and now put to the left next to the T

IG: to the left
IG: this is to the right

IG: yes exactly
IG: like this

IG: exactly
IG: into this spot

IG: now you take the pink piece over there ...

Figure 2: Example for task-oriented conversation in shared visual space from (Zarrieß et al., 2016): the joint task for the IF and IG is to build a puzzle out of Pentomino pieces where the IF can manipulate pieces on a physical gameboard and the IG sees the outline of the puzzle, observes the IF's actions in real-time (over a camera feed) and instructs the IF over headphones; the overall interaction time shown here is approx. 30 secconds; utterances have been translated to English from German transciptions

hand-crafted strategies for repair and revision very clearly improve the referential success of the system. (Villalba et al., 2017) propose a formal approach to generating contrastive referring expressions which is designed for similar scenarios. What is clearly missing to date, however, is a data-driven NLG framework that encompasses these various aspects of conversational grounding and timing in interaction.

5 Conclusion

This paper has discussed the task of interactive instruction giving from the perspective of data-driven NLG. We have argued that, if this task is set up so that it involves visual grounding and spoken language, it will constitute an interesting and considerable challenge for existing data-driven NLG frameworks. We believe that addressing this challenge and coming up with data collections and modeling methods for it will substantially forward the state-of-the-art in NLG, and foster transfer of NLG technology to real-world interactive systems.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48.

Anja Belz. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14(4):431455.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.(JAIR)*, 55:409–442.

Yonatan Bisk, Kevin Shih, Yejin Choi, and Daniel Marcu. 2018. Learning interpretable spatial operations in a rich 3d blocks world. In *Proceedings of AAAI 2018*.

Yonatan Bisk, Deniz Yuret, and Daniel Marcu. 2016. Natural language communication with robots. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 751–761.

Okko Buß and David Schlangen. 2010. Modelling sub-utterance phenomena in spoken dialogue systems. In *Proceedings of the 14th International Workshop on the Semantics and Pragmatics of Dialogue (Pozdial 2010)*, pages 33–41, Poznan, Poland.

Donna Byron, Alexander Koller, Jon Oberlander, Laura Stoia, and Kristina Striegnitz. 2007. Generating instructions in virtual environments (give): A challenge and an evaluation testbed for nlg. *Position Papers*, page 3.

Donna Byron, Alexander Koller, Kristina Striegnitz, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander.

2009. Report on the first nlg challenge on generating instructions in virtual environments (give). In *Proceedings of the 12th european workshop on natural language generation*, pages 165–173. Association for Computational Linguistics.

Joyce Y Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. 2018. Language to action: Towards interactive task learning with physical agents. In *IJCAI*, pages 2–9.

Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. 2018. Using syntax to ground referring expressions in natural images. *arXiv preprint arXiv:1805.10547*.

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

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China. Association for Computational Linguistics.

Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4418–4427. IEEE.

Drew A Hudson and Christopher D Manning. 2018. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.

Ioannis Konstas and Mirella Lapata. 2013. A global model for concept-to-text generation. *J. Artif. Intell. Res.(JAIR)*, 48:305–346.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Angelika Maier, Julian Hough, and David Schlangen. 2017. Towards deep end-of-turn prediction for situated spoken dialogue systems. In *Proceedings of Interspeech 2017*, Stockholm, Sweden.

Ramesh Manuvinakurike and David DeVault. 2015. Pair me up: A web framework for crowd-sourced spoken dialogue collection. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 189–201. Springer.

Junhua Mao, Huang Jonathan, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions.

Antoine Raux and Maxine Eskenazi. 2012. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing (TSLP)*, 9(1):1.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams. 2013. Continuously predicting and processing barge-in during a live spoken dialogue task. In *Proceedings of the SIGDIAL 2013 Conference*, pages 384–393.

Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230.

Gabriel Skantze and Anna Hjalmarsson. 2013. Towards incremental speech generation in conversational systems. *Computer Speech & Language*, 27(1):243–262.

Gabriel Skantze, Anna Hjalmarsson, and Catharine Oertel. 2014. Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication*, 65:50–66.

Tanya Stivers and Jack Sidnell. 2012. Introduction. In Jack Sidnell and Tanya Stivers, editors, *The Handbook of Conversation Analysis*, chapter 1, pages 1–8. Wiley-Blackwell, Oxford, U.K.

Kristina Striegnitz, Hendrik Buschmeier, and Stefan Kopp. 2012. Referring in installments: a corpus study of spoken object references in an interactive virtual environment. In *Proceedings of the Seventh International Natural Language Generation Conference*, pages 12–16.

Kristina Striegnitz, Alexandre Denis, Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Mariët Theune. 2011. Report on the second second challenge on generating instructions in virtual environments (give-2.5). In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 270–279.

Martin Villalba, Christoph Teichmann, and Alexander Koller. 2017. Generating contrastive referring expressions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 678–687.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. 2016. Conditional generation and snapshot learning in neural dialogue systems. In *EMNLP*, pages 2153–2162, Austin, Texas. ACL.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR 2017*.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernandez, and David Schlangen. 2016. Pentoref: A corpus of spoken references in task-oriented dialogues. In *10th edition of the Language Resources and Evaluation Conference*.

Sina Zarrieß and David Schlangen. 2016. Easy things first: Installments improve referring expression generation for objects in photographs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 610–620, Berlin, Germany. Association for Computational Linguistics.