"Cool glasses, where did you get them?" Generating Visually Grounded Conversation Starters for Human-Robot Dialogue

Ruben Janssens
IDLab
Ghent University - imec
Ghent, Belgium
rmajanss.janssens@ugent.be

Pieter Wolfert
IDLab
Ghent University - imec
Ghent, Belgium
pieter.wolfert@ugent.be

Thomas Demeester
IDLab
Ghent University - imec
Ghent, Belgium
thomas.demeester@ugent.be

Tony Belpaeme
IDLab
Ghent University - imec
Ghent, Belgium
tony.belpaeme@ugent.be

***Abstract*—Visually situated language interaction is an important challenge in multi-modal Human-Robot Interaction (HRI). In this context we present a data-driven method to generate situated conversation starters based on visual context. We take visual data about the interactants and generate appropriate greetings for conversational agents in the context of HRI. For this, we constructed a novel open-source data set consisting of 4000 HRI-oriented images of people facing the camera, each augmented by three conversation-starting questions. We compared a baseline retrieval-based model and a generative model. Human evaluation of the models using crowdsourcing shows that the generative model scores best, specifically at correctly referencing visual features. We also investigated how automated metrics can be used as a proxy for human evaluation and found that common automated metrics are a poor substitute for human judgement. Finally, we provide a proof-of-concept demonstrator through an interaction with a Furhat social robot.**

***Index Terms*—Human-Robot Interaction; multi-modal dialogue; conversational agent; Natural Language Generation; Natural Language Processing; situatedness; grounding**

I. Introduction

When we engage in a face-to-face interaction, we expect the other to see us and to understand the context that we are both in. For inter-human communication, making conversation is an important part of our daily lives. It is estimated that during our workday activities, 50 to 80% of our time is spent on communication with peers [1]. In these interactions, we establish common ground, i.e. the set of propositions in a conversation which we treat as 'true' [2]. In order to reach common ground in a conversation, we rely on the concept of *grounding*, a set of propositions on which we can build a conversation [3]. For this, situational awareness is important: the concept of knowing what is going on around oneself [4] and being able to use that knowledge effectively in an interaction with the environment and with social others.

Yet, when we try to have a conversation with a robotic partner, it is unlikely that the robot is able to meet the criteria we have for human interlocutors, which in turn leads to a less natural interaction and experience. Getting robots to both understand their environment, and allowing them to reference it in a conversation, is a challenging objective in the fields of Human-Robot Interaction (HRI) [5] and Natural Language Processing (NLP). Nevertheless, having robots understand their surroundings, with the ability to weave that understanding into an open-domain conversation, is key to a successful HRI.

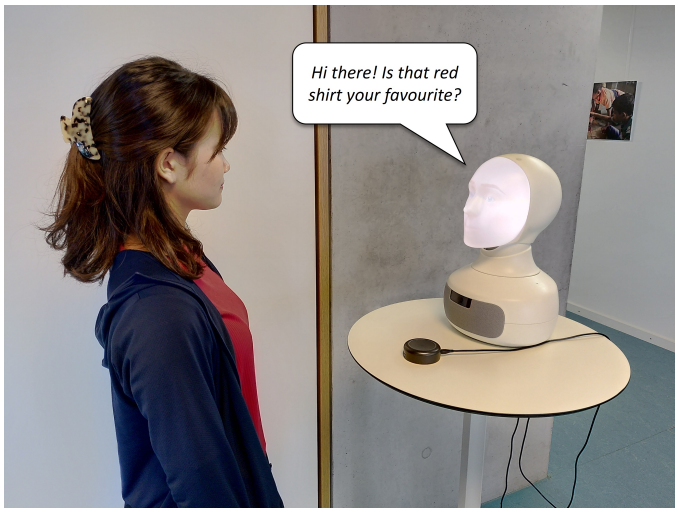


Fig. 1. Illustration of the interaction between a user and a Furhat robot, with the robot inviting the user to have a conversation through a polite question referring to a visual feature of the user.

At the start of each successful conversation is a proper conversation starter. Conversation starters are often polite questions or remarks, which are not necessarily communicatively informative, but are instead a way of inviting the other to an interaction [6]. Most human relationships are formed through such face-to-face interactions, and conversation is central to the construction of a relationship between people. Often, this relationship-forming conversation is lacking a task-oriented aspect and instead is a polite exchange of social talk, also known as phatic communication [7] or small talk. One way of doing this is, for example, to comment on something

This research received funding from the Flemish Government (AI Research Program) and was supported by the Flemish Research Foundation grant no. 1S95020N.

perceptible, like asking why a conversation partner wears their sunglasses on a rainy day. Evidence shows that when we approach a stranger to talk, we frequently rely on publicly displayed cues as a source of material for initial conversations [8]. Small talk has been suggested as being equally important to establish a relationship or bond with artificial agents [9], [10]. When the robot personalises its interaction, by referring to an individual's features or preferences, the perception of the robot improves and secondary outcomes, such as learning with the robot, increase [11], [12]. This requires the artificial agent to have visual perception, but also the skills to place it in a temporal and linguistic context.

In this paper, we take a step towards solving this challenge using data-driven NLP on a social robot to generate conversation-starting questions based on visual information. We present a new crowd-sourced data set with image/question pairs, models to generate questions given novel visual input, various approaches to evaluate them, and an implementation on the Furhat social robot. Fig. 1 shows how these questions are used in a real-world setting, and an overview of the entire system is given in Fig. 2.

II. Methods

A. Visual Conversation Starters Data Set

In order to generate questions to start a human-robot conversation in a data-driven fashion, we required a data set of visual inputs and corresponding conversation starter questions. These visual inputs must correspond with what a robot would sense, and the questions should be related to the content of the input.

1) Images: We created a data set of images that reflect what a robot could encounter in an interaction with a human. We decided on using the YFCC100M data set [13], as it is a very large and varied set of images collected from Flickr and is also used in related work [14], [15]. Following this, we selected images with the keyword 'person'. We used face detection to only select images that contain one face covering at least 5% of the image. Only images having a width of at least 300 pixels were kept, and finally, unrealistic and unusable images were discarded through a manual inspection. This led to 7928 images to be included in the data set.

2) Questions: Questions were crowd-sourced via Amazon Mechanical Turk (AMT). Each participant was shown a unique image, and their task was to come up with three conversation-starting questions. In order to ensure quality, only participants that had already completed at least 5000 tasks on AMT, at least 98% of which approved, were allowed to participate. We also manually removed questions which were not of sufficient quality. Remuneration was in line with the US minimum wage, and ethical guidelines of the university were followed.

Using this method, 3471 images were annotated. We additionally annotated 529 images ourselves, leading to a total of 4000 annotated images and 12000 corresponding questions. These were split into a training, validation, and test set.

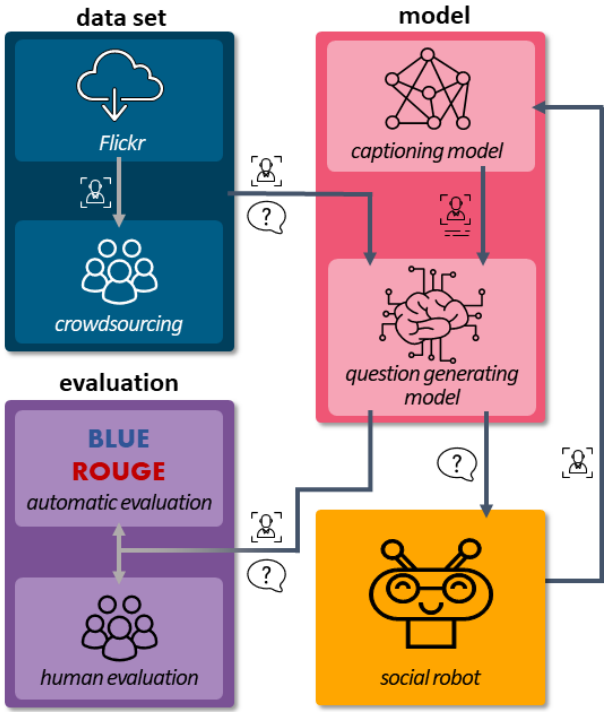


Fig. 2. Overview of the complete system and data flow, consisting of the collection of the data set, training of the model, evaluation and demonstration.

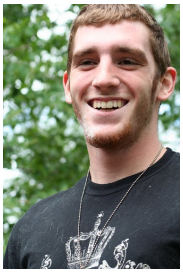
B. Models for Visual Conversation Starters

With our data set containing HRI-appropriate images and questions, we built and trained a system to accurately generate questions for novel images. The first part of the system contains a trained captioning model, that generates a textual description of the input image without the need for fine-tuning. The second part is a model that produces a question based on this image caption, for which we trained a baseline retrieval model and a generative model.

1) Captioning model: We use a captioning model to textually describe the image, removing the need for the question generating model to interpret visual information. We made this choice because training a model to interpret images requires a large amount of data: models for object detection or image captioning are often trained using hundreds of thousands of training examples [15], [16]. Based on the performance of recent captioning models, we expect that they are of sufficient quality for this task, even without fine-tuning.

We chose the 'Dense Captioning with Joint Inference and Visual Context' model [17]. This model is designed to perform "dense captioning", meaning it generates multiple short sentences, each mentioning a different aspect of the image. We joined the different short sentences that had a confidence score higher than 0.7, to provide a short descriptive text. An example of an image with this text can be seen in Fig. 3.

2) Retrieval-based baseline model: Retrieval models produce output by pulling up a suitable response from the training data. In our task, such a model will retrieve the image from the training data that has the most similar caption to that of the



man with brown hair. the eyes are brown. the man has a beard. the man has short hair. the man is smiling. the man is wearing a black shirt. the man is wearing a necklace. the shirt is white. white letters on shirt.

Fig. 3. An example image from the test set. The format in which the captions for this image will be used as input for our question generating model is displayed to the right of the image.

input image and return a question paired with that image. This baseline model is based on bag-of-words text representations, using the term frequency - inverse document frequency (TF-IDF) formula for calculating the cosine similarity on vector representations of captions. The model was implemented using the ParlAI framework [18].

3) Generative model: Generative models produce text directly, rather than retrieving training examples. We chose to use the sequence-to-sequence model BART [19], which achieves state-of-the-art performance on comparable tasks and is embedded in ParlAI. It was pre-trained on large text corpora and we fine-tuned it on our data set. This transfer learning approach allowed us to make use of the knowledge that was extracted from those corpora, beyond our limited-sized visual conversation starters data set. The model was trained on an NVIDIA Tesla V100 GPU, with 32GB of VRAM.

C. Human evaluation strategy

We performed a human evaluation of the generated questions, using crowdsourcing. For this, we considered four relevant dimensions that reflect the quality of the model output: *Syntax* (whether the question contains a language mistake), *Visual reference* (whether the question correctly references a visual feature of the image), *Interestingness* (whether the question is interesting and not generic), and *Appropriateness* (whether the question is appropriate and not offensive). The first two dimensions were scored on a three-point scale, while the latter two used a binary scale. Each point on the scale was accompanied by a brief textual description for the annotators. During analysis of the results, the scales were mapped to a numeric score of 0, 0.5, or 1 (with 1 being better). In order to summarize the four dimensions into a single metric, we calculated the harmonic mean of the respective scores[1], referred to as the SVIA metric (given the labels of the respective dimensions).

As we relied on human raters to evaluate the quality of the responses, we calculated the Inter-Annotator Agreement (IAA) on the SVIA dimensions. The IAA is reported as Randolph's κ, which should be used in situations where a fixed number of raters assign ratings on a nominal scale but do not know the marginal distribution of the categories [20]. Randolph's κ ranges from -1 to 1, with $\kappa < 0$ and $\kappa > 0$ indicating agreement lower and higher than chance, respectively. Given that Syntax and Visual reference scales are rather static (i.e. knowing a sentence is syntactically correct or not), we expect these to have a high IAA. Whether something is interesting or appropriate lies more in the eye of the beholder, and therefore we expect a lower IAA for the matching scales.

To evaluate the quality of generated conversation starters we crowd-sourced participants using Amazon Mechanical Turk, under the same constraints as in Section II-A. The evaluation covered 50 images from the test set and, for each image, the 5 best questions generated by the generative (BART) model and the same for the retrieval (TF-IDF) model. This resulted in 250 image-question pairs per model. Each unique image-question pair was evaluated by four separate participants, and each participant had to pass an implicit and explicit attention check during the evaluation, which were inspired by [21]. The implicit attention check was a normal-appearing image-question pair, but with an objectively clear Visual reference score, allowing us to check whether the participants selected the correct option. The explicit attention check contained an image with instructions on which scoring options to indicate.

III. Results and discussion

Results of the crowd-sourced human evaluation are presented in Table I. The table reports the mean of each of the SVIA dimensions over all test images that were evaluated (n=117 for BART and n=36 for TF-IDF). These metrics were calculated using the majority vote over all evaluators, averaged over all test images. In case of a tie, the mean of the most common values is taken. We also report the corresponding IAA using Randolph's κ.

Both BART and TF-IDF scored high on the Syntax measure and while BART scored satisfactorily on the Visual reference measure, TF-IDF scored significantly worse. Visual reference is clearly the objective that is the most challenging for the models, partially due to caption errors. The Interestingness score did not show a large difference between the BART and TF-IDF models. This is possibly an effect of Interestingness being only a two-point scale. This does, however, show that the evaluators are generally satisfied with the interestingness of the

TABLE I
HUMAN EVALUATION OF THE BART AND TF-IDF MODELS

		BART	TF-IDF
Syntax	Score	0.99	0.94
	Randolph's κ	0.88	0.63
Visual reference	Score	0.70	0.36
	Randolph's κ	0.51	0.51
Interestingness	Score	0.88	0.86
	Randolph's κ	0.48	0.26
Appropriateness	Score	0.97	0.94
	Randolph's κ	0.75	0.58
SVIA		0.87	0.66

[1]This is similar to the traditional F1 metric being the harmonic mean of precision and recall. We considered each of the dimensions equally important for the overall question quality evaluation, and decided on the harmonic mean which strongly penalizes a low score for any of the dimensions in the overall metric.

questions. Crowdsourcers also identified some inappropriate questions, but this amount remained limited.

With more consistently high scores over each dimension, the generative model achieved an overall higher SVIA score than the retrieval model. IAA on the four scales was also generally in line with our expectations. Syntax IAA was moderate-to-high and both Interestingness and Appropriateness IAA are low-to-moderate, as expected. However, Visual reference IAA was only moderate, while we expected this scale to be rather objective. In general, more evaluation data could provide stronger indications on the reliability of the SVIA scale.

A. Relationship between human and automatic metrics

We also looked into whether automatic metrics (used during training of models), can be used as a proxy for human evaluation of the conversation starters. For this we explored the correlation of both BLEU-4 [22] and ROUGE-L [23] (both standard NLP metrics, based on resp. precision and recall) with the human, crowdsourced, evaluations on the SVIA metrics.

Table II shows the results of this analysis: Spearman's ρ shows the correlation between the mean score on each of the SVIA dimensions and the BLEU-4 and ROUGE-L scores for the BART (n=117) and TF-IDF (n=36) models. Only the Visual reference evaluation shows a consistent weak-to-moderate correlation with the automatic metrics.

TABLE II
CORRELATION BETWEEN THE HUMAN AND AUTOMATED EVALUATION

	BART		**TF-IDF**	
	BLEU	ROUGE	BLEU	ROUGE
Syntax	-0.00	0.04	0.31	0.38
Visual reference	0.21	0.28	0.34	0.40
Interestingness	-0.00	-0.04	0.07	0.09
Appropriateness	0.08	0.13	-0.08	-0.07

IV. TECHNICAL DEMONSTRATOR

To demonstrate our system in the wild, we deployed it on a social robot, and observed which questions were generated by the model under varying circumstances (i.e. different attributes and people). Fig. 1 shows the setup with the Furhat social robot [24]. Furhat has a camera, text-to-speech software, three-dimensional face with pan-tilt neck, lip syncing, and gaze recognition through which it can 'engage' with people and follow their movements. The video accompanying this paper also showcases this demonstrator.

The questions are generated by the BART model. Both the captioning model and the question generating model are running in a separate VM. A local computer running a Python script connects all the different components by forwarding Furhat's camera feed to the captioning model, sending the caption to the question-generating model, and using that output to drive Furhat's text-to-speech system.

To showcase the demonstration, four situations, each with a different set of attributes, are shown in Fig. 4. We found that the system primarily focuses on any attributes the user is wearing, such as glasses, a watch, necklace, hat, or tie.

Fig. 4. Four situations where a user is wearing a different set of attributes. Generated questions for each situation: *(leftmost)* `Where did you get your glasses?` *(centre left)* `Where did you get your glasses?` *(centre right)* `Where did you get your dress?` *(rightmost)* `Do you like wearing black? What are you looking at? Where did you get your shirt?`

If none of these items are worn by the user, the system generates questions concerning the colour of a user's clothes, or the length of their hair. In some situations, items are not always recognised, which makes the system generate more generic questions. Finally, the system sometimes hallucinates attributes. It can refer to non-existing items such as earrings.

V. CONCLUSION AND FUTURE WORK

We presented an architecture, qualitative evaluation and HRI demonstration of a system to generate phatic expressions referring to the visual elements of the user. Further, we also described our novel data set with images and conversation starters that are appropriate for HRI, which is publicly available at https://github.com/rubenjanss/visual-conversation-starters. Human evaluation showed that generative models are better at producing questions that correctly refer to visual features than retrieval models. Next, we compared the human evaluation scores with automatic metrics, such as BLEU and ROUGE, and only found weak correlations. To show the effectiveness of our system in the wild, we implemented it on a Furhat robot. Initial results of this HRI indicate that our models can be used for live interaction, when being confronted with new data.

One possible improvement is to make use of transfer learning with a new generation of Transformer models. This might bring a more high-level reasoning to this task and a more diverse set of conversation-starter questions. Also, pre-trained image-to-text models could be fine-tuned end-to-end to investigate whether the captioning is still needed to interpret the image. Another improvement is that the data set currently in use has not been checked for possible biases, which is important when bringing these models and applications to a larger audience [25]. Future work also includes evaluating human-robot interactions that start with a visually situated question, as we only focused on the quality of those questions. We expect that this will lead to a more engaging interaction.

We recognize that building data-driven multi-modal HRI is a significantly larger task than described in this paper, but we believe we have demonstrated how existing technologies can be used for the optimization of HRI. The possibility of using pre-trained models in tasks aimed at HRI, holds a considerable promise as a method to achieve autonomous multi-modal HRI.

REFERENCES

[1] E. Klemmer and F. Snyder, "Measurement of time spent communicating," *Journal of Communication*, vol. 22, no. 2, pp. 142–158, 1972.

[2] R. Stalnaker, "Common ground," *Linguistics and philosophy*, vol. 25, no. 5/6, pp. 701–721, 2002.

[3] H. H. Clark and S. E. Brennan, "Grounding in communication.," 1991.

[4] M. R. Endsley and D. J. Garland, *Situation awareness analysis and measurement*. CRC Press, 2000.

[5] C. Bartneck, T. Belpaeme, F. Eyssel, T. Kanda, M. Keijsers, and S. Šabanović, *Human-robot interaction: An introduction*. Cambridge University Press, 2020.

[6] J. Laver, "Linguistic routines and politeness in greeting and parting," in *Conversational routine*, pp. 289–304, De Gruyter Mouton, 2011.

[7] V. Žegarac, "What is "phatic communication"?," in *Current Issues in Relevance Theory* (V. Rouchota and A. Jucker, eds.), pp. 327–362, John Benjamins, 1998.

[8] H. J. Wiener, *Conversation pieces: The role of products in facilitating conversation*. PhD thesis, Duke University, 2017.

[9] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 12, no. 2, pp. 293–327, 2005.

[10] T. Belpaeme, P. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu, *et al.*, "Multimodal child-robot interaction: Building social bonds," *Journal of Human-Robot Interaction*, vol. 1, no. 2, 2012.

[11] P. Baxter, E. Ashurst, R. Read, J. Kennedy, and T. Belpaeme, "Robot education peers in a situated primary school study: Personalisation promotes child learning," *PloS one*, vol. 12, no. 5, p. e0178126, 2017.

[12] H. W. Park, I. Grover, S. Spaulding, L. Gomez, and C. Breazeal, "A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 687–694, 2019.

[13] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[14] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, "Engaging image captioning via personality," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[15] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[17] L. Yang, K. Tang, J. Yang, and L.-J. Li, "Dense captioning with joint inference and visual context," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[18] A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston, "Parlai: A dialog research software platform," *CoRR*, vol. abs/1705.06476, 2017.

[19] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019.

[20] J. J. Randolph, "Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa.," in *Joensuu Learning and Instruction Symposium*, (Joensuu, Finland), ERIC, 2005.

[21] P. Jonell, T. Kucherenko, I. Torre, and J. Beskow, "Can we trust online crowdworkers?," *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, Oct 2020.

[22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[23] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81, 2004.

[24] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: a back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive behavioural systems*, pp. 114–130, Springer, 2012.

[25] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, *et al.*, "Bias in data-driven artificial intelligence systems—an introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1356, 2020.