Towards narrative generation of spatial experiences in service robots *

Ivan Meza
IIMAS-UNAM
ivanvladimir@turing.iimas.unam.mx

Jorge Garcia Flores
LIPN-Université Paris 13-CNRS
jgflores@lipn.fr

Aldo Gangemi
LIPN-Université Paris 13-CNRS
aldo.gangemi@lipn.univ-paris13.fr

Luis A. Pineda
IIMAS-UNAM
lpineda@unam.mx

Abstract

In this paper we propose a first approximation to the generation of narrative experiences by a service robot. The goal of such narratives is to communicate in a brief, structured and natural way what the robot recently experienced while solving a task. We structure the narrative's elements into blocks guided by the spatial areas the robot visited. We also include a mechanism to mention failures, and template-based natural language generation. At this state of our research we include movement, manipulation and visual events. We evaluate our approach comparing the generated narratives with descriptions produced by humans after having followed the same robot's trajectory and performed the same robot's activities. We use the ROUGUE-L automatic evaluation metric and a survey in order to measure the narrative quality of the robot's description. Results show that while the robot generated narratives have close to human completeness scores, they still lack of readability and fluency.

1 Introduction

As the tasks that service robots solve continue to increase in their complexity and the robots become more autonomous it becomes harder to keep track of the activities that robots perform. From the point of view of human-robot interaction, a desirable skill in a service robot would be that it could tell us its daily experience in solving a task. For instance, after a day cleaning the house it would be natural to ask the robot about its day, and it would also be natural that it could reply to us. There are many possible answers to such a question, however we consider that an informative and natural answer should consider the following aspects:

- *It should be from the point of view of the robot*. Since the trigger of the skill is a direct question by the user, the narrative should be expressed as the robot experienced the task. It must have the robot as the main character of the experience.
- *It should communicate the status of the task*. The narrative should provide an overall review of the successful completion of the task and subtasks, but it should also mention any problems the robot faced in performing the task.
- *It should be about the relevant events of the task*. The narrative should touch those events which are relevant, and leave out actions and details which are not important to communicate the task status.
- *It should time align*. The narrative should be clear about which events happened first and which latter, or which events happened in a repetitive cycle.
- *It should include the context in which the task was performed* A narrative composed of a sequence of events might look plain and confusing. The narrative should indicate the structure of the task and provide details about the scenario in which actions were performed.

In this work we present a method for generating spatial narratives from the experience of a service robot. As a starting point we focus on tasks which include navigation through the space and the manipulation objects. The robot would generate a first person narrative about the places it visited and the actions it performed (for instance: the objects it moved from one place to other). While the robot describes the events following the sequence in which they happened, we propose to organize these events by groups based on the navigation trajectory in sake of the narrative's understandability. In order to provide some context to the user, we extend the current capabilities of the robot to include information from the scenes such as descriptions, readings from the spatial cues (signs) and complementary knowledge about things or message it has never seen before.

As there is no current gold standard that would allow to evaluate this capability, we have collected a small amount of human produced descriptions about a task similar to what the robot would have performed. These descriptions have been used as a gold standard against which we have compare the robot's narratives. Additionally, we evaluate the general quality of the robot generated narratives by applying a survey to the participants subjects of the the evaluation.

The paper is organized as follows. In section 2, we present related work on storytelling and robotics, and narrative generation as independent tasks as well. We also describe our robot platform and framework. In section 3, we define the main elements of our approach for the narrative generation

*

of spatial experiences. In section 4, we present the evaluation settings and section 5 discuss the results of this work. Finally, section 6 proposes some perspectives and further work.

2 Related Work

The robotic skill of narrating has been proposed with different goals in mind. For instance [Jensen *et al.*, 2003] proposed to combine multiple sensor readings from multiple robots to create a narration of a scene through time. This narrative consisted in a sequence of short descriptions about activities perform by persons and the resulting spatial configurations (e.g., *person 01 approached R03 front side*). This research did not address the problem of generating a structured narrative: short descriptions came one after another as soon as their robots detect a change in the scene.

Another application of narratives in robotics has been to provide the robot with the storytelling capabilities. This case has been extensively studied: [Mutlu *et al.*, 2006] propose a method to model and evaluate the storytelling ability by a robot. However, in this case the robot tells a predefined story, not necessary about its experience or even itself. Another aspect of storytelling that has been studied is the effect of robot interaction on children's language acquisition. In particular collaborative storytelling has been show to have a positive impact on children vocabulary acquisition [Kory and Breazeal, 2014; Kory, 2014; Leversund *et al.*, 2014]. The inclusion of gestures and gaze during the storytelling has improved understanding [Ham *et al.*, 2011] as well. For further examples in this area, refer to [Gwo-Dong Chen and Wang, 2011] which provides a brief survey in this field.

On the other hand, recent progress in computer vision and the generation of descriptions for images, videos and movies [Torabi *et al.*, 2015; Rohrbach *et al.*, 2015] have made accessible a new set of tools for the generation narrative experiences. Most of these methods are based on recurrent neural networks (RNN) [Yao *et al.*, 2015] trained using large corpus mapping images or video sequences to natural language descriptions (for instance, audio descriptions intended for visually impaired). Although this advances represent a new window of opportunity in order to program a narrative skill into a robot by coupling such systems into a task, to our knowledge such approach has not been tried in a service robot before, mainly because a flat description is different from a narrative. In this direction, [Ting-Hao *et al.*, 2016] propose a dataset which can be used to generate narratives from a sequence of images. Although, this later approach is quite promising, at this moment we are faced with the problem of how to incorporate actions performed by robot into the narrative, for this reason we choose a more symbolic approach that allows to incorporate knowledge from the robot. Just recently we learn about the [Rosenthal *et al.*, 2016] which proposes an alternative approach to the one here presented.

2.1 The robotic platform

For the generation of narratives we use the robot Golem-III. This is the third generation of a service robotic platform. The previous version, Golem-II+, and Golem-III have been extensively programmed to participate in the RoboCup competition [Wisspeintner *et al.*, 2009 12 01T000000] and to perform research in the field. Both platforms were programmed using SitLog [Pineda *et al.*, 2013], a programming language for service robot tasks, and they rely in a collection of basic behaviors which are assembled by the robot task programmers [Pineda *et al.*, 2015]. Some of these tasks include following a person, obeying complex commands, assisting in a restaurant [Pineda *et al.*, 2016], being a waiter [Rascon *et al.*, 2015], playing Marco Polo, taking assistance of a class [Meza *et al.*, 2016] among others.

Table 1 lists the main components of Golem-III and Figure 1 shows a picture of the robot.

PatrolBotTM robot	Sick LMS-500 Laser
On-board computer	Two laptops
Kinect 2 camera	Point Grey Flea camera
8SoundsUSB audio interface	RODE directional mic
Two-Way loudspeaker	5-bumper arrays
In house arms	In house torso

Table 1: Golem-III hardware.

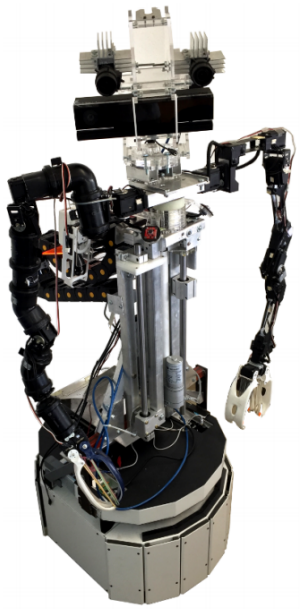

Figure 1: Golem-III.

3 Narrative Generation of Spatial Experiences

For a service robot, we define the generation of spatial experiences as the ability to narrate in a retrospective way a set of events E experienced by the robot during a specific navigation path P. This is intended to be listened by a human, so it has the form of a first person short navigation narrative which includes information from the robot knowledge base (K) and it is enriched by contextual information (C). Here is

an example of a narrative produced by human subject 2 after having performed the same task than the robot:

> *"First I went past the bathroom, I saw the sign saying that it was the Computer Science Department and the range of cubicles in the department. Then I went to Ernesto Bribiesca's cubicle, it was shut so I couldnt pick up the water nor move the bottle. After that I got into Ivan Meza's cubicle, I read the SOPA poster and I took the water that was on the desk, I left it again and I saw on the desktop a computer with keyboard, mouse, a portrait, speakers. After that I went to the meeting room, where I found one person presenting, the room was busy, there was a Golems project meeting, and finally I got to the lab."*

This description corresponds to the following task: the robot starts at the lobby of our department, then it has to go to an office and move a water bottle from one desk to another; then it has to do the same but in a different office. Next it will visit a meeting room and if busy it had to read why it is busy or what is happening there. Finally, it will visit the lab and finish the task there. In the narration, it can be noticed that the robot was not able to visit the first office.

Tasks as the previous one are programmed using SitLog programming language; the programmer of such tasks decides which actions to log and how to enrich them by adding visual events; for instance after it just started a task, arrived to a point or having failed. Once a sequence of events E is defined there are three processes which intertwine in order to generate the narrative:

1. Extraction of discourse blocks from the sequence of events
2. Generation of propositions
3. Fix the discourse for failing events

Discourses blocks

In order to create a coherent narration we propose to group the events into blocks B representing the different areas the robot visits. At the log, we only have motion actions to specific points in the building. For instance, in the example task, the visit of the first office is triggered by the action $go(p_2)$. Using this information and the knowledge base K we are able to infer that this particular point is in a certain office. Every other event that happens will be grouped in the same block. Blocks are put together using connector phrases such *afterwards* or *after that*.

Generation of propositions

We have identified four kinds of propositions which can be directly translated from a sequence of events inside a block (i.e., $E_1 \in B_i$): first, spatial statements describing the robot's navigation movements; second, actions performed by the robot; third, statements describing visual objects on the room; fourth, statements describing sentences read by the robot in context. We consider movements events different from other actions of the robot (e.g., grab an object) since we used them to mark when we enter to a new area as explained before.

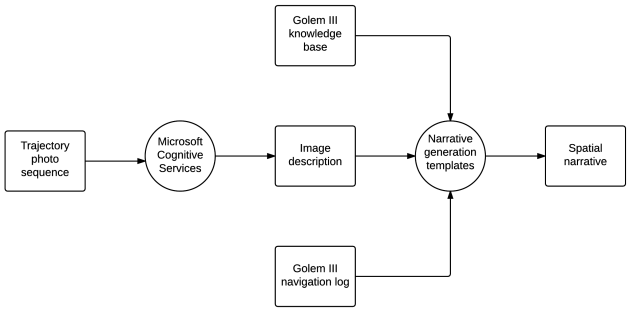

Figure 2: Narrative generation system architecture

At this moment in our research we have chosen to use a simple template realization in order to generate propositions from events. The robot knowledge base (K) is augmented with templates for each type of event. For instance an action of type such as $go(loc)$, has associated the following template *i went to* loc. In order to deal with the monotony of saying always in the same thing, we use a weighted set of templates per event type, and we randomly draw from them at the generation step. At this point, we have a simple set of thumb rules in order to produce the nouns with the right determination ($adesk$ vs $thedesk$). Events within an block are put together, using the connector *then*.

In case of visual events, we have used a commercial visual description system[1]. The confidence of the description is used in order to decide when to complement the information with an automatic description. Finally, when a text is detected in the visual event we include this information as something the robot read. Similarly to events, we use the weighed set of templates per description to avoid monotony.

An example of a proposition generated by our systems: for the event $start(p_1)$, which signals that the robot started the task in the point p_1, our system is able to generate the following proposition: *I woke up at the hall, there was a sign on the door there it said "computer science department"*. The *start* event is associated with the *I woke up at* loc template. From the knowledge base, it is able to infer that such point is located at the *hall*. Then the proposition is enriched by the description of the visual event *a sign in the door*, and finally it is completed by reading the text contained in the sign.

Generation of failing events

At this point we are generating events as if they were successful, however when something fails it is usually logged after the fact. In order to narrate whatever fails, we have to go back in the description and fix it. When a fail event happens, we pop the latest proposition generated and rephrase in order to indicate that that event did not happened and actually failed. In the narration example, this happened when the robot couldn't open the first office's door (see table 2).

[1] https://www.microsoft.com/cognitive-services/en-us/computer-vision-api

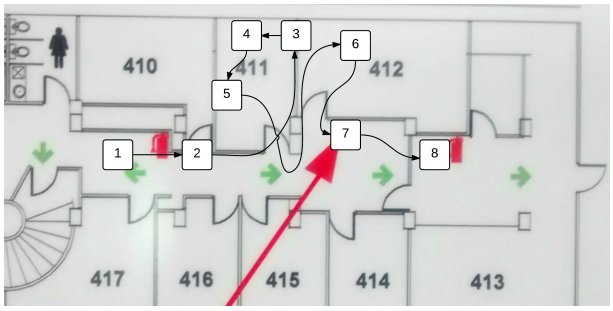

Figure 3: In vitro trajectory

4 Experimental settings

To evaluate the spatial narratives generated by the robot we implemented an experimental protocol where both humans and the robot had to follow the same trajectory, stop at the same points and perform the same actions. At the end of the trajectory, we asked the human participants to describe their trajectory for a recorder. The transcription of these records were compared with the robot's narratives in order to evaluate similarity, readability and fluency.

In vitro trajectory

We established an 8 stop trajectory at IIMAS lab (see figure 3). At each point, the participants were asked to perform the same actions than the robot: observe a sign, describe a scene, get into an office, read a poster. We also include a failed action (get into a closed office, see table 2) with the intention of including frustrated actions into the narrative description.

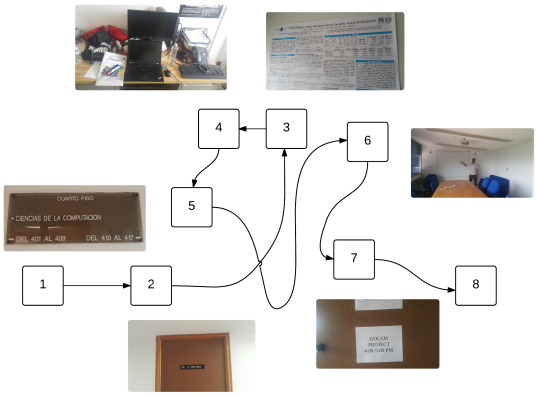

Figure 4: Photo sequence of points 1, 3, 5, 6 and 7

This experimental trajectory is described as *in vitro* because narratives were generated using the robot's system outside the robot, by means of a web interface we developed for testing purposes. A photo was taken at each step of the trajectory, exactly at the point were the robot would be programmed to stop. The photo sequence of the trajectory (see figure 4) was uploaded into the web application interface to start the narrative generation process.

Human and robot narratives

Five human participants were asked to stop at each point of the trajectory and perform the action specified in the protocol (see table 2). At the end of the path, we record the participants describing their trajectory. The protocol asked the participant to focus on descriptions, signs and both successful and unsuccessful actions. We translated the recordings into English and the transcripts were compared, both automatically and manually, against the robot narratives.

Three different version of the same robot narrative were evaluated. For **Robot-v0** the templates took the raw output of the image description service without any fine tuning. **Robot-v1** has a confidence score threshold to filter which descriptions would be included in the final narrative. **Robot-v2** includes the confidence threshold and was fine-tuned to improve sentence segmentation and punctuation.

Evaluation Metrics

The automatic metrics we used for comparing the generated narratives with the human gold standards is ROUGE-L [Lin, 2004], which was originally intended for automatic summarization evaluation [Dang and Owczarzak, 2008]: it measures lexical similarity between texts and it has a high correlation with human judgments. A Turing-like survey was used to measure accuracy, completeness, fluency and readability. Three days after the experience, the participants subjects who performed the robot task were asked to evaluate two texts each: one text was from another fellow participant, the other was robot-generated (the judge ignored the text author).

Point	Instruction
1	Start and read the sign
2	Get into Mr. Bribiesca's office and move the water bottle
3	Get into Mr. Meza's office and read the poster
4	Take the water bottle from the desk under the poster
5	Leave the water on the desk behind and describe what's on it
6	Go to the meeting room and describe the scene
7	If the meeting room is busy, look for a sign on the door and read what's going on
8	Go to the lab and stop

Table 2: Task procedure for human participants

5 Results

Table 3 shows the scores of the automatic evaluation using the ROUGE-L metric. The five human produced descriptions are taken as a gold standard for each of the robot generated narratives. The upper part of the table evaluates each of the three robot versions against this gold standard. The lower part of the table shows a cross-evaluation of human narratives (4 gold standards texts for each evaluated peer).

ROUGE-L evaluates recall and precision of lexical similarity between texts based on the longest common subsequence. In this case, the best recall score was obtained by **Robot-v0**, which had no fine tuning at all but had almost 20% more words than the fine tuned narratives v1 and v2. The same was observed in the cross-evaluation of the human descriptions: the best recall was obtained by Subject-5 description, which has the highest word count. On the other hand, **Robot-v3** was the most accurate according to the precision measure having

the lowest word count, and the same relation was observed within the human texts, where Subject-2 description had the highest precision and the lowest word count. On ROUGE-L precision evaluation the **Robot-v2** system outscored Subjects 3, 4 and 5 precision.

Narrative	Words	Recall	Precision	F
Robot-v0	164	**0.30515**	0.45663	**0.36583**
Robot-v1	139	0.25362	0.45324	0.32524
Robot-v2	139	0.28341	**0.50647**	0.36344
Subject-1	175	0.43587	0.58409	0.47559
Subject-2	119	*0.31863*	0.66469	0.42227
Subject-3	211	0.49374	0.50227	0.47204
Subject-4	307	0.60205	0.40819	0.46578
Subject-5	392	0.61060	*0.31663*	*0.40657*

Table 3: ROUGE-L automatic evaluation for robot and human narratives.

The fine tuning process was triggered by the fact that **Robot-v0** has evident semantic and interpretation flaws (like taking a poster as a TV or finding *two giraffes in a fenced area* in the corridor of a robotics lab). Here's **Robot-v0** narrative:

> *"I begun at in the hall, I saw a a sign on the door, I read it said COMPUTER SCIENCE DEPARTMENT. Then, I tried to, I walk towards the hall there, but couldn't, I noticed a clock mounted on a wall, it said DR. ERNESTO BRIBIESCA. Then, I went to the office of Ivan there, I saw a a picture of a tv, I read it said SOFA: Random Forests Regression for the Semantic Textual Similarity task, I grabbed a bottle, I moved to the desk there, I dropped the bottle, I noticed a cluttered desk with a laptop,it said sausenc &aceta para consejoros - ASO. When finish, I walk towards the meeting room there, I noticed a man standing in a room, I moved to the door there, I noticed two giraffes in a fenced in area,it said GOLEM. At this point, I went to the laboratory there, there was a a desk with a computer and a chair, I stopped at laboratory."*

And this is the content of **Robot-v2**, after a confidence threshold and segmentation and fluency improvements on the templates:

> *"I woke up at the hall there was a sign on the door there it said COMPUTER SCIENCE DEPARTMENT. After that I wanted to walk towards the office of Ernesto but couldn't, after that I walked towards the office of Ivan to a desk there it said SOFA: Random Forests Regression for the Semantic Textual Similarity task I took a bottle then I moved to a desk I leaved the bottle I noticed a cluttered desk with a laptop there it said sausenc &aceta para consejoros - ASO. Afterwards I moved to the meeting room to a table I saw a man standing in a room then I went to a door there it said GOLEM. After that I moved to the laboratory to a desk I noticed a desk with a computer and a chair afterwards I finished."*

From the simple reading of the two examples we can suppose that the confidence score of the image descriptions service reduces semantic noises (like *giraffes* or *clocks mounted on the wall*). Furthermore, the segmentation and fluency template improvements seem to increase the text's readability. Table 4 shows the results of a survey where five judges evaluated the quality of one robot **Robot-v1** and one human generated narrative each.

The survey's results seems to contradict ROUGE-L evaluation: human judges considered that the best robot's quality was completeness (the equivalent of ROUGE's recall, where the robot's narratives were low). On the other hand, they found robot's narratives lacking precision, unlike ROUGE, where **Robot-v3** got its highest score. Human judges also found robot's texts lacking of fluency and readability, and in this evaluation there was no quality criteria were the robot outscored human generated narratives.

	Humans avg.	Robot
Precision	4.0	2.8
Completeness	4.2	3.6
Readability	4.0	2.6
Fluency	3.8	2.8

Table 4: Human evaluation (five judges; 1=low quality; 5=high quality)

6 Perspectives

This paper presents an original approach for generating spatial experience narratives in service robots. Our goal is to enhance service robot's with experience capabilities, that is, to be able to summarize past activities, movements and perceptions for an absent human user in a narrative way. With this goal in mind, we traced an experimental trajectory where the robot had to perform a set of activities (like reading a poster, moving a water bottle or describing a scene) and produce a summary at the end of his path. We implemented a spatial narratives generation system based on an image description service and natural language generation templates. An evaluation method based on comparing robot generated narratives with first person descriptions of human performing the same robot activity trajectory was implemented. An automatic evaluation showed that the robot's narratives got poor recall (or completeness) scores but better precision, with scores outperforming three human produced narratives. A text quality evaluation by human judges showed the robot's narratives getting close to human scores on completeness, but lacking readability and fluency. Further improvements will include a more sophisticated method for text representation and generation. We will be adding FRED [Gangemi *et al.*, 2014] machine reading system in order to amplify the robot's knowledge base and, for instance, be able to find the meaning of *Semantic Similarity* or *Random Forest* in a poster read by the robot. Another improvement will be the use of *Geni* natural language generation system [Gyawali and Gardent, 2014] that might increase readability and fluency of the robot's experience narrative. With the inclusion of FRED and

Geni in our methodological frame we hope to enrich the behavior description set of the robot and thus be able to describe more complex actions, like person following and face or voice recognition. Finally, the automatic evaluation methods need some improvements: ROUGE-L evaluation seemed to be biased by the length of the text, so we might make use of the Pyramid evaluation method [Nenkova *et al.*, 2007] to improve the experimental settings. Code and test set is available under an open source license.[2]

References

[Dang and Owczarzak, 2008] Hoa Trang Dang and Karolina Owczarzak. Overview of the tac 2008 update summarization task. In *In TAC 2008 Workshop - Notebook papers and results*, pages 10–23, 2008.

[Gangemi *et al.*, 2014] Aldo Gangemi, Valentina Presutti, and Diego Reforgiato Recupero. Frame-based detection of opinion holders and topics: a model and a tool. *Computational Intelligence Magazine, IEEE*, 9(1):20–30, 2014.

[Gwo-Dong Chen and Wang, 2011] Nurkhamid Gwo-Dong Chen and Chin-Yeh Wang. A survey on storytelling with robots. *Edutainment Technologies*, page 450, 2011.

[Gyawali and Gardent, 2014] Bikash Gyawali and Claire Gardent. Surface realisation from knowledge-bases. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 424–434, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[Ham *et al.*, 2011] Jaap Ham, René Bokhorst, Raymond Cuijpers, David van der Pol, and John-John Cabibihan. Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power. In *Social robotics*, pages 71–83. Springer, 2011.

[Jensen *et al.*, 2003] Bjom Jensen, Roland Philippsen, and Roland Siegwart. Narrative situation assessment for human-robot interaction. In *Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on*, volume 1, pages 1503–1508. IEEE, 2003.

[Kory and Breazeal, 2014] J. Kory and C. Breazeal. Storytelling with robots: Learning companions for preschool childrens language development. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Washington, DC., 2014. IEEE.

[Kory, 2014] Jacqueline Jacqueline Marie Kory. *Storytelling with robots: Effects of robot language level on children's language learning*. PhD thesis, Massachusetts Institute of Technology, 2014.

[Leversund *et al.*, 2014] Anna Helen Leversund, Aleksander Krzywinski, and Weiqin Chen. Childrens collaborative storytelling on a tangible multitouch tabletop. In *Distributed, Ambient, and Pervasive Interactions*, pages 142–153. Springer, 2014.

[2] anonymized.com.

[Lin, 2004] Chin-yew Lin. Rouge: a package for automatic evaluation of summaries. pages 25–26, 2004.

[Meza *et al.*, 2016] Ivan Meza, Caleb Rascon, Gibran Fuentes, and Luis A Pineda. On indexicality, direction of arrival of sound sources and human-robot interaction. *Journal of robotics*, 2016. To appear.

[Mutlu *et al.*, 2006] Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Humanoid robots, 2006 6th IEEE-RAS international conference on*, pages 518–523. IEEE, 2006.

[Nenkova *et al.*, 2007] Ani Nenkova, Rebecca J. Passonneau, and Kathleen McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *TSLP*, 4(2), 2007.

[Pineda *et al.*, 2013] Luis A. Pineda, Lisset Salinas, Iván Meza, Caleb Rascón, and Gibrán Fuentes. SitLog: A Programming Language for Service Robots' Tasks. *International Journal of Advanced Robotic Systems*, 2013.

[Pineda *et al.*, 2015] Luis Pineda, Arturo Rodrguez, Gibran Fuentes, Caleb Rascon, and Ivan V. Meza. Concept and functional structure of a service robot. *International Journal of Advanced Robotic Systems*, 2015.

[Pineda *et al.*, 2016] Luis Pineda, Caleb Rascon, Gibran Fuentes, Varinia Estrada, Arturo Rodriguez, Hernando Ortega, Mauricio Reyes, No Hernndez, and Ricardo Cruz. The golem team, robocup@home 2016. Technical report, 2016. To appear.

[Rascon *et al.*, 2015] Caleb Rascon, Ivan Meza, Gibran Fuentes, Lisset Salinas, and Luis A Pineda. Integration of the multi-doa estimation functionality to human-robot interaction. *Int. J. Adv. Robot. Syst*, 12(8), 2015.

[Rohrbach *et al.*, 2015] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[Rosenthal *et al.*, 2016] Stephanie Rosenthal, Sai P. Selvaraj, and Manuela Veloso. Verbalization: Narration of autonomous robot experience. In *25th International Joint Conference on Artificial Intelligence*, 2016. To appear.

[Ting-Hao *et al.*, 2016] Ting-Hao, Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. Visual Storytelling. In *In Proceedings of NAACL'16*, 2016. to appear.

[Torabi *et al.*, 2015] Atousa Torabi, Pal Chris, Larochelle Hugo, and Courville Aaron. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint*, 2015.

[Wisspeintner *et al.*, 2009 12 01T000000] Thomas Wisspeintner, Tijn van der Zant, Luca Iocchi, and Stefan Schiffer. RoboCup@Home: Scientific Competition and Benchmarking for Domestic Service Robots. *Interaction Studies*, 10(3):392–426, 2009-12-01T00:00:00.

[Yao *et al.*, 2015] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. *ArXiv e-prints*, abs/1502.08029, February 2015.