Towards a storytelling humanoid robot

Q. Anh Le[1], C. d'Alessandro[2], O. Deroo[3], D. Doukhan[2], R. Gelin[4], J.C. Martin[2], C. Pelachaud[1], A. Rilliard[2], S. Rosset[2] (names in alphabetic order)

[1] Telecom ParisTech-CNRS, [2] LIMSI-CNRS, [3] Accapela Group , [4] Aldebaran Robotics

{quoc, catherine.pelachaud}@telecom-paristech.fr, {cda, doukhan, martin, rilliard, rosset}@limsi.fr, olivier.deroo@acapela-group.com, rgelin@aldebaran-robotics.com

Abstract

We describe the settings of a project that aims at giving a robot the ability to tell tales to children. The choices made to produce the linguistic and behavioral expressivity required to produce a credible storytelling are detailed.

Introduction

This paper reports on the ongoing work done in the GVLEX project. The aim of this multidisciplinary project is to design and test a storytelling humanoid robot. Ideally, the robot would be able to process automatically a given tale or short story, and to play it for a children audience.
Such a project is by nature interdisciplinary and involves: text analysis (discourse, expression, characters), expressive text-to-speech synthesis (particularly expressive prosodic synthesis), expressive posture and gesture synthesis, and the coordination, between all these levels and aspects.
The robot used is NAO, a medium scale autonomous humanoid robot depicted in Figure 1. The remaining of this abstract describes the different tasks and processing levels involved in the project, the methodology adopted, the multimedia and speech corpora designed and recorded, and the first results obtained.

The robotic actor: Nao

Nao is a 57cm high humanoid robot with 25 degrees of freedom that can give it very expressive gestures. It is equipped with a text-to-speech software from Acapela group and it can play wave files to provide audio illustration. As last expressive feature, Nao is able to control the color of its eyes. A complete description of Nao is available in Gouaillier et al. (2009).
Nao is delivered with a graphical programming tool, Choregraphe (Pot et al., 2009) that allows, in an intuitive way, the design of complex behaviors. The expertise of the Aldebaran Robotics' developers relies in the ability to use Nao's features to transform a nice humanoid robot into a robotic actor. One aim of the GVLEX project is to provide tools to non-expert developers that can share this expertise of robotic actors' "director".

The speech of tales

Linguistic structure of tales

The main questions addressed in this part of the project are: What is the linguistic structure of tales? How can we characterize them? Or more precisely, what are the structural and linguistic information that can be useful for expressivity (at speech or gesture level)? The useful information is obviously multilevel. In this work we are not willing to design complete analysis for each level of interest but rather to design a multi-level analysis able to point out the interesting parts of the tale. Based on the classical studies by Propp (1928) and Greimas (1966), we defined two different levels of analysis, namely the "structural" level and the level of "lexical elements".
As for the structural level, a tale is a story that contains a title, an exposition, potentially followed by a scene, including a triggering event, from which the story unfolds. The core of the story can be structured in a series of scenes optionally interleaved with refrains. The story then ends within the epilogue. As for lexical elements, an extended named entity (ENE) definition has been proposed which is adapted to short stories: "person" which implies all kind of characters in the stories, that means human person but also animals or plants etc., "localization" and "time" information which are more standard.
The ENE person is useful in order to provide to Nao information concerning the speaker (is the narrator speaking or one of the story's character?) and then adapting speech and gesture synthesis to the speaker characteristics.
Moreover, as the speech synthesis is involved, some purely linguistic information, as syntactic information could be useful. Then, phrases boundaries and multi-word expressions segments are considered of interest.

Prosody of storytelling

Automatic text to speech (TTS) synthesis of tales is a new and difficult problem that involves several different levels: expressive prosody, narrative structure, character rendering, prosody beyond the sentence,
Expressive prosody means the prosody of attitudes and emotions: emphasis, accentuation, changes of registers and tempo, pauses, etc. The narrative structure indicates important prosodic aspects like character changes,

proximity/distance to the audience, specific prosodic patterns associated to recurrent structures (repetition, characters, refrains…).
Current TTS systems generally address only sentence-level prosody. In this project, prosody is considered at a global level: paragraph level prosody, discourse level prosody.

Speech and gesture corpora

We produced a corpus of 89 short stories, manually annotated on these different (narrative, syntactic, discourse) levels. Each story contains an average of 907 words (from 407 to 1318). Then 12 selected tales have been recorded by a professional speaker and annotated using a multi-layer framework (phonemes, syllables, pitch, rhythm, voice quality). Moreover, one tale has been produced by 6 actors and videotaped. The resulting 80 minutes have been annotated in order to specify a lexicon of about 500 gestures used in narrative situations.

Speech, Gestures, postures coordination and data flow

The control of the robot behavior is done through the real-time platform Greta, designed to control the multimodal behavior of embodied conversational agent. It follows the SAIBA flow (Kopp et al., 2006). It takes as input what the robot or agent aims to communicate and outputs the corresponding multimodal nonverbal behaviors. The input text is augmented with communicative and emotional information encoded through FML (Function Markup Language) (Heylen et al, 2008), while the output behavior is represented with BML (Behavior Markup Language) (Vilhjalmsson et al, 2007). Both FML and BML are XML languages. BML is body-independent, i.e. it is not constraint by a particular body type or animation parameters. We use BML to represent the behavior of the NAO robot and of the virtual agent.
Thus the flow of our algorithm is as follow: it takes as input the story to be told augmented with communicative functions and prosodic tags. Greta computes the synchronized nonverbal behaviors to be played by the robot or by the virtual agent. So, for a same FML input, the behavior of the virtual agent and of the robot should convey similar meanings. The difficulty arises as the robot Nao and the virtual agent Greta do not share the same modalities. For example Nao has no facial expression and almost no finger while Greta does not walk. As a consequence, several BML tags outputted by the system cannot be displayed either by the robot or by the agent. It could result in different meanings conveyed by the robot and the agent animations. Our solution is to use two lexicons, one for the robot and one for the agent where their respective entries should convey similar meanings, as illustrated in fig. 1. To build these lexicons, we rely on the notion of gesture variant and gesture family introduced by Calbris (1990). A gesture family encompasses several instances of behaviors, which may differ in shape, but convey similar meanings. Thus the entries in the lexicons of the robot and of the agent are part of a same gesture family, even if they differ in shape.

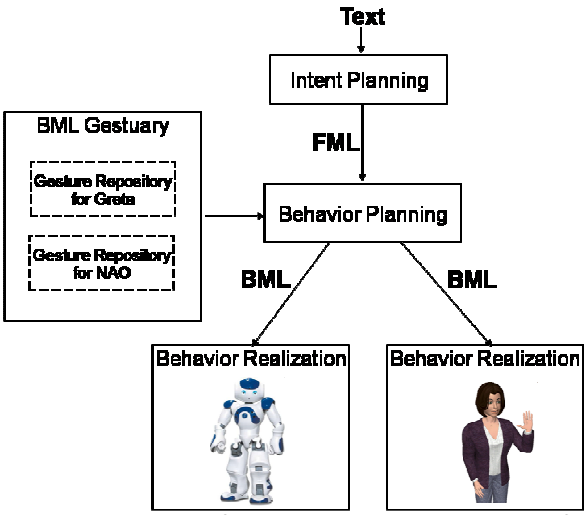

Fig. 1: The Greta platform controlling the behaviors of the Nao robot (left) and virtual agent (right).

Acknowledgements

This work has been partially funded by the French project GVLex (ANR-08-CORD-024 http://www.gvlex.com/).

References

Calbris, G. 1990. *The semiotics of French gestures*. University Press, Bloomington: Indiana.
Gouaillier, D., Hugel, V. et al., 2009. Mechatronic design of NAO humanoid. IEEE Int. Conf. on Robotics and Automation, Kobe, Japan.
Greimas, A.J., 1966. *Sémantique structurale : recherche et méthode*, Larousse, 1966
Heylen, D., Kopp, S., Marsella, S., Pelachaud, C., Vilhjalmsson, H., 2008. The Next Step Towards a Functional Markup Language, Intelligent Virtual Agents, IVA'08, Tokyo.
Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thorisson, K, Vilhjalmsson, H., 2006. *Towards a Common Framework for Multimodal Generation in ECAs: The Behavior Markup Language*. 6th Int. Conf. on Intelligent Virtual Agents, Marina del Rey.
Pot, E. et al., 2009. Choregraphe: a Graphical Tool for Humanoid Robot Programming. Proceedings of Roman
Propp, V., 1968 (orig. 1928). Morphology of the Folktale, Austin University, Texas Press.
Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thorisson, H. van Welbergen, R. van der Werf, 2007. The Behavior Markup Language: Recent Developments and Challenges, Intelligent Virtual Agents, IVA'07, Paris.