F-2 Robot Toolkit: Model for Speech Comprehension and Communicative Behavior

Artemy Kotov
Kurchatov Institute, Russian State University for the Humanities
Moscow, Russia
kotov@harpia.ru

Anna Zinina
Kurchatov Institute, Russian State University for the Humanities
Moscow, Russia
zinnia_aa@nrcki.ru

Liudmila Zaidelman
Kurchatov Institute, Russian State University for the Humanities
Moscow, Russia
luda.zaidelman@yandex.ru

Nikita Arinkin
Kurchatov Institute, Russian State University for the Humanities
Moscow, Russia
nikita.arinkin@gmail.com

Keywords—multimodal communication, speech understanding, robot-to-human interaction, pointing gestures

I. INTRODUCTION

We design cognitive architecture, software and hardware for a robot F-2 – an experimental base to simulate text comprehension and natural communication via speech, gestures and facial expressions. The robot receives speech sounds or written text, extracts text semantics and selects behavioral strategies or emotional reactions in different communicative situations. The robot performs speech parsing on several levels of a linguistic model: it performs morphological and syntactic analysis, assembling a set of syntactic trees with semantic representations. Further it compares the semantic structures with a set of *scripts* – *if-then* productions, used as models for emotions and speech inferences. The activation of scripts simulates emotional dynamics and basic natural-language inferences. In a communication scripts generate behavioral patterns to be executed on a physical robot (Fig. 1).

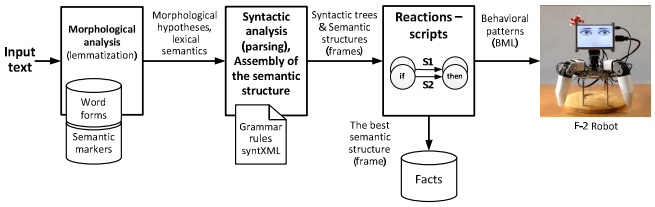

Fig. 1. Processing architecture of F-2 robot

These components work as follows. The robot may receive oral or written texts at its input. Oral texts are converted to written form via the Yandex Speech API service. Written texts may be fed as text files (like fiction) or via the RSS subscriptions. Speech component of the robot is operated daily in a standalone mode: it downloads and processes about 7,000 sentences from top news and blogs, extracted facts are stored to a database. Morphological analysis associates each wordform to a lexeme and assigns a set of grammemes: *gender*, *case*, *number* etc. For the morphological analysis we use a dictionary, based on OpenCorpora project [1], it contains 100,000 lexemes (1.5 mln wordforms) saved to an SQL database. Regular expressions unit and a neural network are used to process compound words (like: *47*, *3-rd, 5%*) and unknown words. 30,000 lemmas are annotated by semantic markers basing on the semantic primitives in [2] and semantic dictionary [3]: a list of 606 semantic markers is used for the annotation. Words are characterized by markers of their semantic class (hyperonyms) and potentially emotional markers as described in [4].

Syntax processing is implemented as a left-to-right syntactic parser. Wordforms are sequentially fed into a stack, stack head is compared to the set rules on syntXML language [5]. The rules reduce the stack head and establish a syntax binding. In case of syntactic homonymy, when several rules can apply to a stack, a separate stack is created for each applied rule. All the stacks are evaluated and a constant number of stacks (usually top 256) is preserved for further steps. We expect, that for compound emotional reactions like irony [6, 7], and for studies of the architecture of consciousness [8] – several trees (stacks) may be used from the heap, where one tree constitutes the "accurate" understanding, while other trees may form "ironical" or "hypothetical" understanding of the text. Each tree receives its semantic representation: a set of semantic markers of the corresponding words, distributed between semantic valencies: *agent, patient, instrument, time, place* etc., based on [9]. These semantic representations are further evaluated within the scripts component.

Scripts component is designed to model simple inferences and communicative reactions to an incoming event. Script units are defined as *if-then* operators (productions) and distributed into emotional and rational scripts. From the theoretical point of view the model inherits Sloman's CogAff architecture for emotional agents [10], while an emotional script as a concept corresponds to the notion of *proto-specialist* by Minsky [11].

Scripts *if*-conditions (*premises*) are defined as *semantic predications* – similar to a semantics of a simple sentence. For each semantic predication in an incoming phrase the parser calculates the distance to each script. The best tree with semantic predications, closest to the scripts, is selected as the most emotional (closest to the emotional scripts) or as a tree with regular semantics (corresponding to rational scripts). Each script is activated by a stimulus (a) proportionally to the similarity between the stimulus and script *if*-condition and

978-1-7281-3891-6/19/$31.00 ©2019 IEEE

(b) proportionally to the sensitivity of the script. Each activated script outputs a behavioral packet on BML – Behavior markup language [12].

The behavior of the robot and gesture design are developed basing on the studies in natural human communication, as described in [13]. We rely on REC corpus [14] with annotated video recordings of natural emotional dialogs – 295 talks at university oral exams, 600 recording of talks with clients at a municipal office. Within the corpus we annotated over 300.000 speech and nonverbal behavioral elements: facial expressions, gestures, head and body movements. The prototypic patterns are drawn on a 3D-figure of the robot in Blender 3D-editor and exported to MySQL database, the gesture dictionary contains 220 gestures. Each script may be provided with a piece of behavior in BML format. Upon the activation of a script, it outputs one of its BMLs to the *robot controller*, which manages all the BML packets from different scripts, and combines the packets: e. g. head movement from one script may be executed simultaneously with a hand gesture from another script, expressing two emotions or communicate functions at the same time. This design allows the robot to execute blending gestures [15].

F-2 is an advanced experimental device to design and verify the emotional reactions, invoked by the processing of text semantics as well as communicative behavior during various tasks, in the suggested demo experiment – pointing gestures during assisting in solving a problem.

II. Demo Experiment: Pointing Gestures

In the suggested demonstration robot helps participants to complete a Tangram puzzle – a well-known experimental media for studies of natural human communication [16], development of linguistic resources [17, 18] as well as for the design of robot communicative strategies [19]. The puzzle consists of 7 elements of different color, shape and size. The task of a participant is to arrange the elements on a table or within the contour on a white sheet. The robot detects the position of each puzzle element by a visual recognition system and suggests the next move.

In an actual experiment we have tested the role of pointing gestures of the robot: the goal was to verify if the participant follows the gestures of the robot, and if he recognizes the pointing behavior of the robot. The experiment involved 31 participants (12 female, 19 male), at a mean age 27. During the experiment the participant was to complete 4 figures. The whole experiment was recorded from two viewpoints: from the side with the view on a player (Fig. 2a) and the top view of the playing field (Fig. 2b).

Before each task the game elements were placed in front of the participant on the left and right sides of the playing field (Fig. 6). Two paired elements (large triangles; small triangles) were always placed on different sides of the playing field. In its speech instructions the robot has always been referring to an element by its shape and size (not by color). Thus, an ambiguous reference in speech had been appearing when the robot mentioned one of the paired elements, like *Take a big triangle!* (a triangle on the left or on the right?).


Fig. 2a. Side view

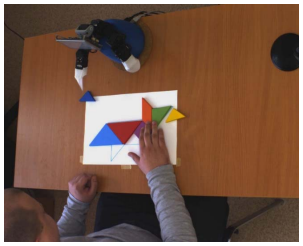
Fig. 2b. Top view

In the first experimental condition the robot accompanied its instructions by oriented communicative actions: it used pointing hand gestures, head and eyes movements (Fig. 3a). In the second experimental condition the robot did not use pointing gestures, its speech instructions were accompanied by non-oriented movements of hands, head and eyes (Fig. 3b).

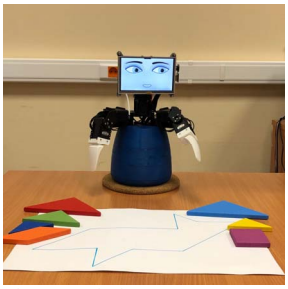
Fig. 3a. Pointing gesture

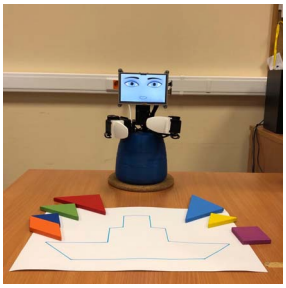
Fig. 3b. Symmetric gesture

According to the results of the experiment the participants significantly more often (chi-squared, $p < 0.01$) prefer the robot which uses oriented gestures (1st condition). But the difference between the experimental conditions was not obvious for participants: only half of the players (15 people; 48.4% of the total group) noticed the difference between robot's performance with and without pointing gestures. At the same time, a nonverbal instruction was significant even for the participants, who did not notice the difference. We have examined the cases, when two paired elements are not yet used and are placed on the two sides of the playing field, the robot may refer to such an element ambiguously as *big triangle* or *small triangle*. When the robot used a pointing gesture with the speech reference (both on left and right side), the participants followed this nonverbal indication in 91.1% of cases and took the element, the robot was pointing at. When the robot did not show a pointing gesture, the participants took a random element (48% – left and 52% – right). This fact shows a fairly stable influence of pointing gestures on participants' behavior, even if they do not explicitly notice the difference between *pointing* and *non-pointing* behavior.

Acknowledgments

The research is supported by the grant of the Russian Science Foundation (project №19-18-00547).

REFERENCES

[1] V. V Bocharov, S. V Alexeeva, D. V Granovsky, E. V Protopopova, M. E. Stepanova, and A. V Surikov, "Crowdsourcing morphological annotation," in Computer Linguistics and Intellectual Technologies, Issue. 12 (19). Vol. 1, Moscow: RSUH, 2013, pp. 109–114.

[2] A. Wierzbicka, Semantic primitives. Frankfurt/M.: Athenäum-Verl., 1972.

[3] N. Yu. Shvedova, Russian Semantic Dictionary. Moscow: Azbukovnik, 1998.

[4] A. Kotov. Mechanisms of speech influence in mass media. Ph. D. Thesis. 23.06.03. Moscow, 2003.

[5] A. Kotov, A. Zinina, and A. Filatov, "Semantic Parser for Sentiment Analysis and the Emotional Computer Agents," in Proceedings of the AINL-ISMW FRUCT 2015, 2015, pp. 167–170.

[6] S. Attardo, J. Eisterhold, J. Hay, and I. Poggi, "Multimodal markers of irony and sarcasm," Humor - Int. J. Humor Res., vol. 16, no. 2, pp. 243–260, 2003.

[7] A. Kotov, "Accounting for irony and emotional oscillation in computer architectures," in Proceedings of International Conference on Affective Computing and Intelligent Interaction ACII 2009, Amsterdam: IEEE, 2009, pp. 506–511.

[8] A. A. Kotov, "A computational model of consciousness for artificial emotional agents," Psychol. Russ. State Art, vol. 10, no. 3, pp. 57–73, 2017.

[9] C. J. Fillmore, "The Case for Case," in Universals in linguistic theory, E. Bach and R. T. Harms, Eds. New York: Holt, Rinehart & Winston, 1968, pp. 1–68.

[10] A. Sloman, "Beyond Shallow Models of Emotion," Cogn. Process., vol. 2, no. 1, pp. 177–198, 2001.

[11] M. L. Minsky, The Society of Mind. New-York, London: Touchstone Book, 1988.

[12] S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. Thórisson, and H. Vilhjálmsson, "Towards a Common Framework for Multimodal Generation: The Behavior Markup Language," in Intelligent Virtual Agents, 2006, pp. 205–217.

[13] M. Rehm and E. André, "From Annotated Multimodal Corpora to Simulated Human-Like Behaviors," in Modeling Communication with Robots and Virtual Humans, 2008, pp. 1–17.

[14] A. Kotov, E. Budyanskaya, "The Russian Emotional Corpus: Communication in Natural Emotional Situations," in Computer Linguistics and Intellectual Technologies. Issue 11(18). Vol. 1. Moscow, RSUH, 2012, pp. 296–306.

[15] M. Ochs, R. Niewiadomski, C. Pelachaud, and D. Sadek, "Intelligent Expressions of Emotions," in J. Tao, T. Tan, and R.W. Picard (Eds.) ACII 2005, LNCS 3784, Berlin, Heidelberg: Springer-Verlag, 2005, pp. 707–714.

[16] H. H. Clark and D. Wilkes-Gibbs, "Referring as a collaborative process," Cognition, 1986.

[17] T. Shore, T. Androulakaki, and G. Skantze, "KTH Tangrams: A Dataset for Research on Alignment and Conceptual Pacts in Task-Oriented Dialogue," in Proceedings of the 11th Language Resources and Evaluation Conference, 2018.

[18] M. Gnjatović and D. Rösner, "Inducing genuine emotions in simulated speech-based human-machine interaction: The NIMITEK corpus," IEEE Trans. Affect. Comput., 2010.

[19] D. Kirschner, R. Velik, S. Yahyanejad, M. Brandstötter, and M. Hofbaur, "YuMi, come and play with me! a collaborative robot for piecing together a tangram puzzle," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016.