Linguistic approaches to robotics: from text analysis to the synthesis of behavior[1]

A. Kotov[1,2], N. Arinkin[1], L. Zaidelman[1], A. Zinina[1]

[1] National Research Center "Kurchatov Institute", Moscow, Russia, pl. Kurchatova, 1
[2] Russian State University for the Humanities, Moscow, Russia, Miusskaya pl., 6
kotov_aa@nrcki.ru

Abstract. We examine the problem of "understanding robots" and design an F-2 emotional robot to "understand" speech and to support human-like behavior. The suggested system is an applied implementation of the theoretical concept of robotic information flow, suggested by M. Minsky ("proto-specialists") and A. Sloman (CogAff). This system works with real world input – natural texts, speech sound – and produces natural behavioral output – speech, gestures and facial expressions. Unlike other chatbots, the system relies on semantic representation and operates with a set of d-scripts (equivalents to proto-specialists), extracted from advertising and mass media texts as a classification of basic emotional patterns. The process of "understanding" is modelled as the selection of a relevant d-script for the incoming utterance.

Keywords: Syntactic parsers, semantic processing, robot companions, emotional agents

1 Introduction

M. Minsky has suggested that the basic architecture underlying information processing of living beings and future robots should be based on "proto-specialists": simple reactive units, detecting negative or positive situations on one side and executing behavioral reactions on the other [1]. According to Minsky, "proto-specialists" are responsible for basic drives and emotions (hunger, thirst, aggression, flight etc.). They constantly compete for the body's resources: the winning proto-specialist controls the behavior (e. g. direction of movement), while the others remain on standby. Further, A. Sloman has extended the architecture with a triple layer model with sophisticated conflict handling [2, 3]. While basic drives and emotions remain on the first (basic) level, their activation may be suppressed by deliberative reasoning (second level) and meta-management – reflective processing (third level). This describes the interaction between rational and emotional processes: while rational processing may control emotions to a certain degree, reflexes grab control when an organism needs to recoil

[1] Design of the syntactic parser was supported by RFBR grant 16-29-09601 ofi_m, development of negative emotional patterns was supported by RSF grant 17-78-30029, and design of the F-2 robot was executed within the research program of the "Kurchatov Institute".

from a snake or a falling object. This architecture was implemented in the CogAff project for the design of emotional agents [4], and underlies the architecture of many agents in artificial societies. Such agents have very limited sets of input events and output actions – although designed as models for living beings (humans) or as software prototypes of future robot companions, which will live in the real world and should handle natural language texts (and other semiotic systems) as their input, showing believable emotional behavior as their output. In this work we demonstrate the application of this theoretical architecture to a robot companion, which receives natural language speech as its input and generates speech, gestures and facial expressions for its output.

In our studies we solve two main problems:

(a) develop a system for natural text understanding, which also draws conclusions from the received statements and suggests possible reactions for an utterance;

(b) develop a system to execute behavioral reactions on a robot companion.

The general architecture of the system is represented in Fig. 1

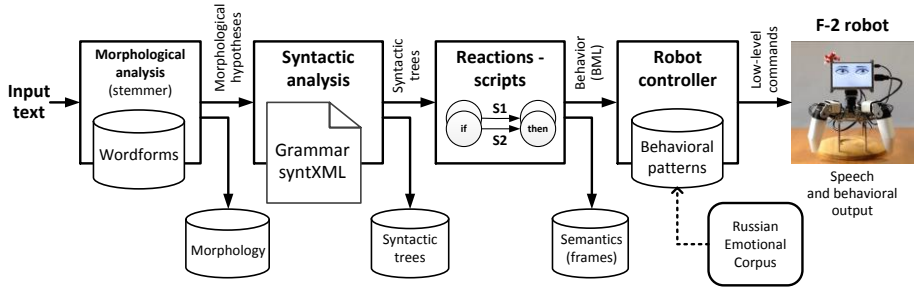

Fig. 1. Processing workflow for the emotional robot F-2.

The system accepts written texts in Russian as the input or decodes input audio instructions with Yandex Speech API. Further, the text is processed by morphological and syntactic components. Syntactic trees are transferred to the reaction component, which extracts from each tree a shallow semantic representation and compares it to the scripts – models of emotional and rational reactions. The winning script(s) form output behavioral reactions with utterances in BML (Behavior Markup Language) – which are combined and processed by the robot controller, and further executed by the F-2 robot. During the execution of BMLs, the robot controller retrieves the behavioral patterns from a database, designed following the research on real behavior in emotional dialogues in the Russian Emotional Corpus (REC).

2 Text analysis and sentiment extraction

Present technologies of sentiment analysis usually rely on the "bag-of-words" method, in which a text is represented as a set of wordforms [5-7]. "Emotion" is calculated as an average score of emotional words present in the text. Although this approach is suitable for text classification and even for some dialogue systems, it omits an im-

portant aspect: emotion is triggered by an event and in text comprehension emotion is triggered by an event representation in text semantics. Semantic roles are crucial in this representation: the message *X conquered Y* has different emotional contents for *X* and for *Y*. Further, *You are an idiot!* should sound like an insult to the addressee, not like a neutral statement of the evaluation 'N is bad/stupid'. Several approaches in automatic text analysis and sentiment extraction rely on partial syntax parsing and extraction of T-expressions (three-element tuples <*subject, relation, object*>) [8] or four-element sets (subject, relation, actant, negation) [9]. Modern compound projects, like ABBYY Compreno parser [10] or Sentilo project [11], suggest constructing a syntax tree of the whole sentence and adding the semantic information to the nodes of the tree. In these cases semantic facts or emotional assertions can be found as substructures of the extended syntactic trees. Within our project we design a system which should recognize a set of "emotional" semantic patterns in the incoming text. Therefore the text parser should analyze the morphological/syntactic structure of the text and construct its semantic representation – suitable for the recognition of emotional meanings.

2.1 Semantic parser for natural text analysis

At the first stage of analysis, the parser divides incoming text into wordforms and retrieves possible morphological hypotheses from a dictionary that contains approximately 737,000 wordforms for 48,000 lexemes. Wordforms from the analyzed text are represented as lexemes and are assigned grammatical and semantic attributes. Semantic attributes are retrieved from a semantic dictionary; it assigns to each lexeme a set from a list of 660 semantic markers – main semantic categories, like 'human' or 'object', and emotional markers like 'intensive' or 'inadequate'. 28,000 words are annotated by semantic markers: from 1 to 18 markers per word (on average 2). Homonymic and polysemic meanings are distinguished: a marker can be assigned to a specific meaning of a word.

At the second syntactic stage the parser builds a syntactic tree for each sentence. At each step it adds the next wordform to a stack and tries to reduce the stack head with the available rules. The syntactic component relies on the rules described in syntXML format [12]. A rule is represented as a possible reduction, in which the right-hand side is reduced to the left-hand head ***h*** (1). Head ***h*** can also be a member of the right-hand side (2):

$$\boldsymbol{h} \rightarrow \langle a, b, \ldots n \rangle \tag{1}$$

$$\boldsymbol{h} \rightarrow \langle a, b, \boldsymbol{h}, \ldots n \rangle \quad \text{or} \quad \langle a, b, \boldsymbol{h}^{\text{head}}, \ldots n \rangle \tag{2}$$

This rule structure allows us to combine rules from immediate constituent grammars as well as from dependency grammars. In our case we use "immediate constituent" rules for conjunction groups (*John and Pete* are reduced to a "virtual head" – [John_and_Pete]) and dependency rules for most other cases (*red house* is reduced to [house]).

A syntactic rule indicates a semantic role for an actant in the future semantic representation. For example, if *John walks* is reduced by subject+verb rule, *John* is excluded from the stack and is assigned an *agent* (or **ag**) semantic role. Other actants of a verb fill in other semantic valences – *patient, instrument, cause* etc.

The syntax component outputs syntactic trees: numerous trees for a sentence can be constructed, taking into account morphological or syntactic homonymy. No trees are generated if the sequence of words is considered to be "ungrammatical" and cannot be reduced by the syntactic rules to a single head.

2.2 Model of emotional reactions

Labels of semantic roles, assigned by syntactic rules, are used to construct a semantic representation: the semantic markers of each word are transferred to the corresponding semantic valence. For a single clause, a semantic representation (semantic predication) is a table of semantic valencies, filled with semantic markers for each valency. Each additional clause in a sentence adds another semantic predication. As shown in [12] the utterance *A real man is always interested in the life of the beloved girl* gives the following semantic predication:

Table 1. Semantic representation (semantic predication) for the utterance *A real man is always interested in the life of the beloved girl*

p (predicate)	ag (agent)	pat (patient)
think, pay-attention, frequently	object, somebody, man, positive	abstract, time-period, existence, object, somebody, woman, of-minimal-age, positive

Each semantic predication is compared with the inventory of scripts, the typical emotional reactions of the robot. Proximity to one of the scripts allows the parser to resolve homonymy and choose a preferred tree from a set of alternatives, if multiple trees have been constructed at the syntactic stage.

The system of scripts is represented in [13, 14]. Scripts are extracted by content analysis of publicity materials, leaflets and advertisements, etc. The inventory contains rational scripts (or **r-scripts**, corresponding to deliberative reasoning in CogAff architecture) and dominant scripts (**d-scripts**, corresponding to reactions or alarms in CogAff architecture). D-scripts are represented by 21 positive scripts – for compliments, expression of joy, advertising – and 13 negative scripts – for conflict, lamentation, or negative influence.

The scripts model emotional reactions to an incoming utterance and produce output behavioral patterns. For each sentence, a measure of proximity to each script is calculated. The script is activated depending on its affinity with the incoming utterance and proportionally to the "mood" or "temperament", simulated by the robot (the scripts have current activation levels and "sensitivity", which allows the robot to imitate a temperament). In particular, the semantic predication in Table 1 activates negative d-scripts (PLAN: somebody plans something frightening against me – 'man makes an

evil plan against woman' – and SUBJV "Subjectivity": somebody is narrow-minded, thinks only about one thing – 'all men think about are women') and positive d-scripts (ATTENTION: subject is pleased, because somebody pays him attention – 'woman is happy because of the men's attention' – and APPROVAL: somebody acts like a hero, does something right – 'real men do well to pay attention'). The tendency to prefer negative or positive scripts can be controlled by the simulated robot character. The most activated script produces a robot output – a speech utterance and an element of nonverbal behavior to be executed by the F-2 robot.

3 Development and execution of emotional behavioral patterns

Future robot companions should communicate fluently with people, attain and maintain emotional contact, and maintain an interface, similar to normal personal interaction. In order to execute the generated reactions we developed a control system of behavior for the personal robot F-2 (Fig. 3). We extract behavioral patterns from a multimodal corpus and combine these patterns during the execution by the robot to enrich its behavior and the expression of emotions.

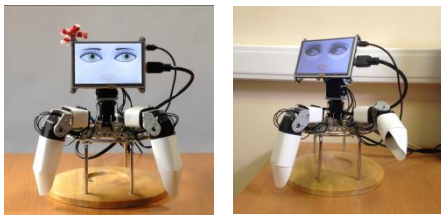

Fig. 2. F-2 emotional robot.

3.1 Research and design of communication patterns

Activated scripts generate elements of communicative behavior, represented by packages in BML – Behavior Markup Language [15, 16]. These packages contain speech utterances and expressive patterns (gestures, facial expressions). BML packages refer to gesture movements stored in a database – during the execution these movements are retrieved from the base, linearized, and executed by the F-2 robot.

Elements of communicative behavior are developed with the help of emotional patterns from the multimodal corpus REC, Russian Emotional Corpus [17]. The REC includes video records of real emotional interactions: university exams, customer service in a municipal office, interviews with people engaged in art. The corpus is marked up with the help of ELAN software – a professional tool for complex annotation of video and audio resources. We marked up speech of participants as well as the facial expressions and gestures of one of the interlocutors (student, client, respondent). A separate level of markup sets the communicative function of a mimic move-

ment or gesture if this function can be definitely determined [18]. Thus the corpus allows us to select typical elements of behavior to express a specific function. These elements are drawn in the Blender 3D editor and saved to the database. Such a library of BML reactions is developed for each script.

Fig. 3. Conversion of an emotional gesture from REC corpus to Blender 3D character for further use of the gesture by F-2 robot to express a certain communication function, in this case joy.

3.2 Behavior execution

The developed system simulates the behavior of the robot depending on the script activation and expresses the modeled gestures using the robot's motors, screen and audio system. As the input the system receives messages with the following information:

- script identification;
- weight of script activation,
- BML, produced by this script.

This information is processed and queued in descending order of activation. The system tries to activate tags of BMLs from the top of the queue. During BMLs scanning tags check if the execution device (hand, head etc.) for this tag is free. Tags with free devices become active while others stay in the queue waiting for the corresponding device to be unlocked. Preparation of tags consists of three mechanisms:

- acquisition of the instructions from the database and their parameterization;
- generation of instructions based on the environment model;
- generation of speech.

Prepared data represents instructions based on time stamps. Each tag has its own timer, which returns instructions associated with the active time stamps. Received instructions are transferred to the corresponding device (gesture instructions to robot motors, facial expressions to screen, utterances to audio system). After the processing of all time stamps for each tag according to the timer triggers, the tag completes, which lets the system find further tags to execute. If all tags of a BML are complete then the information on the script that has generated the BML is deleted from the queue and the activation of the script is reduced. If the robot activates several scripts then the BMLs from different scripts can be combined: robot can demonstrate a gesture of denial by hand and reverie through its head (turning the head and eyes up and

sideways) (Fig.2). To produce miscellaneous combinations we have formed a hierarchy of body parts, which allows us to describe instructions for one or several expression devices (e.g. for one hand or both hands). This hierarchy is used in the process of tag activation to choose the most convenient way to show a behavior on the robot.

Moreover the elements of behavior can be synchronized in between. For each gesture, facial expression or utterance it is possible to allocate the following main points:

- beginning;
- end;
- peak – marked as "stroke".

For example, the robot can start to think by looking away and turning its head left, right or up, then speak with an appeal – directing its hand to the interlocutor. In this way the gesture "appeal" is synchronized with the utterance, which allows the robot to express emotions more accurately. In general, such combinations of elements allows us to make the communicative behavior of the robot richer, more complex and more attractive to people. Thus we develop a cycle for the information processing, which constructs and executed compound behavior on different robot actuators for the given activation of scripts – in particular, expressing conflicts in the behavior, when contradictory scripts are activated and expressed at the same time (similar to script activation in 2.2).

4 Conclusion

The suggested system is an applied implementation of the theoretical concept of robotic information flow, suggested by Minsky and Sloman. This system works with real world input – natural texts, speech sound – and produces natural behavioral output: speech, gestures and facial expressions. Unlike other chatbots, the system relies on operations with semantics and operates with a set of d-scripts (equivalents to proto-specialists) extracted from advertising and mass media texts as a method of classification of basic emotional patterns. The process of "understanding" in this system can be modelled as the selection of a relevant script for the incoming utterance. At the same time, emotional relevance is subjective – one may be happy and/or confused, while being the center of attention. Here we suggest that the selection of one "most" relevant reaction is not reasonable – several highly activated emotional reactions may be mixed in the output, forming compound behavior of an artificial agent (robot), able to express emotions with numerous output devices.

References

1. Minsky, M.L.: The Society of Mind. Touchstone Book, New-York, London (1988).
2. Sloman, A.: Beyond Shallow Models of Emotion. Cognitive Processing 2, pp. 177-198 (2001).

3. Sloman, A., Chrisley, R.: Virtual Machines and Consciousness. Journal of Consciousness Studies 10, 133-172 (2003).
4. Sloman, A.: Varieties of Affect and the CogAff Architecture Schema. In: Johnson, C. (ed.) Proceedings Symposium on Emotion, Cognition, and Affective Computing AISB'01 Convention, vol. 10, pp. 39-48, York (2001).
5. Su, F., Markert, K.: From words to senses: a case study of subjectivity recognition. Proceedings of the 22nd International Conference on Computational Linguistics. Volume 1, pp. 825-832. Association for Computational Linguistics, Manchester, United Kingdom (2008).
6. Chetviorkin, I.I.: Testing the sentiment classification approach in various domains — ROMIP 2011. Computational Linguistics and Intellectual Technologies. Issue 11, vol. 2, pp. 15-26. RSUH, Moscow (2012).
7. Poroshin, V.: Proof of concept statistical sentiment classification at ROMIP 2011. Computational Linguistics and Intellectual Technologies. Issue 11, vol. 2, pp. 60-65. RSUH, Moscow (2012).
8. Katz, B.: From Sentence Processing to Information Accesson the World Wide Web. AAAI Spring Symposium on Natural Language Processing for the World Wide Web (1997).
9. Mavljutov, R.R., Ostapuk, N.A.: Using basic syntactic relations for sentiment analysis. Computational Linguistics and Intellectual Technologies. Issue 12, vol. 2, pp. 91-100. RSUH, Moscow (2013).
10. Anisimovich, K.V., Druzhkin, K.J., Minlos, F.R., Petrova, M.A., Selegey, V.P., Zuev, K.A.: Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. Computational Linguistics and Intellectual Technologies. Issue 11, vol. 2, pp. 91-103. РГГУ, М. (2012).
11. Recupero, D.R., Presutti, V., Consoli, S., Gangemi, A., Nuzzolese, A.G.: Sentilo: Frame-Based Sentiment Analysis. Cognitive Computation 7, 211-225 (2014).
12. Kotov, A., Zinina, A., Filatov, A.: Semantic Parser for Sentiment Analysis and the Emotional Computer Agents. Proceedings of the AINL-ISMW FRUCT 2015, pp. 167-170 (2015).
13. Kotov, A.A.: Mechanisms of speech influence in publicistic mass media texts (in Russian). Ph.D thesis. RSUH, Moscow (2003).
14. Kotov, A.A.: Mechanisms of speech influence. Kurchatov Institute, Moscow (2017).
15. Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thórisson, K., Vilhjálmsson, H.: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. Intelligent Virtual Agents, pp. 205-217 (2006).
16. Vilhjálmsson, H., Cantelmo, N., Cassell, J., E. Chafai, N., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A., Pelachaud, C., Ruttkay, Z., Thórisson, K., van Welbergen, H., van der Werf, R.: The Behavior Markup Language: Recent Developments and Challenges. Intelligent Virtual Agents, pp. 99-111 (2007).
17. Kotov, A., Budyanskaya, E.: The Russian Emotional Corpus: Communication in Natural Emotional Situations. Computational Linguistics and Intellectual Technologies, Issue 11 (18). Vol. 1., pp. 296-306. RSUH, Moscow (2012).
18. Kotov, A. A., Zinina A. A.: Functional analysis of nonverbal communicative behavior (in Russian). Computational Linguistics and Intellectual Technologies. Issue. 14., vol. 1, pp. 299-310. RSUH, Moscow (2015).