Conceptual Processing System for a Companion Robot

Artemiy Kotov[1, 2*] [0000-0003-3353-5549], Liudmila Zaidelman[1, 2] [0000-0002-2941-144X], Anna Zinina[1, 2] [0000-0001-9575-1875], Nikita Arinkin[1, 2] [0000-0003-2303-2817], Alexander Filatov[1], Kirill Kivva[1,3]

[1] Russian State University for the Humanities, Moscow, Russia
[2] National Research Center "Kurchatov Institute", Moscow, Russia
[3] Bauman Moscow State Technical University, Moscow, Russia
kotov@harpia.ru

Abstract. Companion robots should perceive speech, recognize objects in the real world, and further react with speech utterances and nonverbal communicative cues. Robots should also remember the interaction history and accumulate knowledge from external text sources: news, blogs, and e-mails. We designed a conceptual representation system for a companion robot, able to support this list of interactive tasks. The system includes a speech processing component and operates with semantic representations: sets of semantic markers, assigned to valencies within sentence predications. The reaction support system inherits a classic *production* architecture and consists of scripts, sensitive to rational or emotional stimuli. The general architecture is based on parallel processing of scripts, it may trigger several behavioral reactions in response to a stimulus, and after that combines speech output and nonverbal reactions of these scripts on the robot, thus, constructing compound and rich behavior. Semantic representations and scripts are also used to index incoming utterances in a memory base.

Keywords: Conceptual Processing, Knowledge Base, Production Logic, Semantic Representation, Emotional Computer Agent.

1 Introduction

While traditional linguistic models deal with text representations on syntactic and semantic levels, conceptual processing systems should operate a cognitive behavioral model (or a robot) and thus should handle visual recognition, speech reactions, question answering, problem-solving and support discussions on actions in the problem space. These systems should operate with situational conceptual representations, shared by the modules of speech comprehension, problem solving and memory. We represent a project aimed at the development of a conceptual processing and natural communication system for a personal companion robot with some extensions to the domains of problem solving and visual comprehension.

The support of natural dialogue and natural question answering is a recently fast-growing area, utilizing numerous approaches, and in particular – scripts and conceptual representations. Dialogue systems are usually divided into rule-based, information-retrieval and statistical systems [1; 2; 3]. However, as shown at the 2017 Amazon Alexa

Prize competition [4], most of the participating teams used rule-based approach in their chatbots, while boosting the approach with neural networks and machine learning. In this sense, a dialogue support system may be treated as a production architecture, where an input utterance triggers the scripts and the best response script is further selected.

In this work we mostly follow the SOAR architecture, designed to execute diverse natural cognitive tasks, and communicate to humans in this respect [5]. SOAR uses scripts as the operational base, as it offers problem solving by the enumeration of scripts from *the given state* to *the target state* (solution) of a problem. The operation with scripts allows the system to communicate to a user the present internal state and the proposed moves, discovered in the problem space.

2 Design of reactions

R. Shank's classic works introduce *scripts* as a model for natural inference and suggests a solution to question answering basing on the inferred representations [6]. We use a set of scripts to process an input. For each input, the closest script is selected. In the proposed architecture the scripts are also used as (a) models of emotional and rational reactions, both – verbal and nonverbal, (b) representations of "regular situations" for the resolution of ambiguity, (c) indexes for semantic representations in the memory base. The constructed semantic representations, further, can be used in several components of an extended cognitive model: linking speech processing with visual input recognition and with memory. We combine scripts with neural networks: the latter are used to evaluate syntactic links and to select a set of syntactic trees in the cases of ambiguity. Further, the calculation of distance between an input and script premises is based on multiplication of weights of different semantic markers – similar to the neural networks.

2.1 Competing reactions

Emotion processing by a limited set of reactions – *proto-specialts* – was suggested by M. Minsky [7]. According to his view, a set of *proto-specialts* dominate during the processing of each stimulus and define the reactions of an organism in case of danger or urgent lucrative opportunity. A. Sloman has further suggested, that emotional processing competes with "rational" inferences, and has distinguished the competing cognitive levels of *reactions/alarms* and the *deliberative reasoning* within CogAff architecture [8]. It was noted that the *units for rational processing* are more accurate in the classification of stimuli and provide better planning, while *reactions* are faster and ensure survival of an organism in critical situations. Sloman has noted, that compound emotional processes may engage both rational end emotional units, like *secondary emotions*, which trigger emotional processing by a rational inference. As noted in [9] an emotional event may trigger diverse emotional and etiquette speech reactions, which linearize to compound combinations of (a) interjections (b) emotional evaluations (c) emotional classifications (d) acquisition of speaker's responsibility, and (e) etiquette replies. This suggests that the competing scripts do not completely suppress each other but can be distributed in time. Within the suggested architecture the scripts are divided into 3 groups: (a) *d-scripts or dominant scripts*, responsible for the emotional processing [10], (b) *behavioral rules* – etiquette and oughtness, and (c) *r-scripts or rational scripts*, responsible for rational classification of an input stimulus. Several scripts may be invoked by a stimulus and send their BML packages to the robot, providing the combination of an emotional exclamation with an etiquette reply. Further, the compound behavioral patterns can be created by diverse internal states: a reaction to an incoming phrase and internal anxiety, expressed by automanipulation.

2.2 Situational representation

L. Barsalou within the research of conceptualization has noted that the structure of a notion – e. g. *chair* – depends of the situation, where the notion appears – e. g. *kitchen, cinema, hotel* [11]. This observation has been categorized as a set of assertions, describing the rules of conceptualization [12]. Following the assertions, the set of semantic features within a notion (actant or verb) can be affected by the frame of the situation. A similar process was noted within the psychology of emotions: during an emotional *top-down* processing an invoked emotion or drive can change subjective representation of a situation [13], e. g. a person tends to overrate his ability to eat while hungry

Within the suggested system, a semantic predication (premise of the script) contains a list of markers, significant for the recognition of an incoming event. Although not all the sought markers may appear in the stimulus to invoke the script, the markers, marked within the script as *focal* [14], are applied to the incoming representation on a "top-down" basis, thus changing the initial representation and making it "more emotional". This ensures, that the representation of an incoming situation converges to the script, and is appended by markers, significant for the script.

2.3 Speech processing

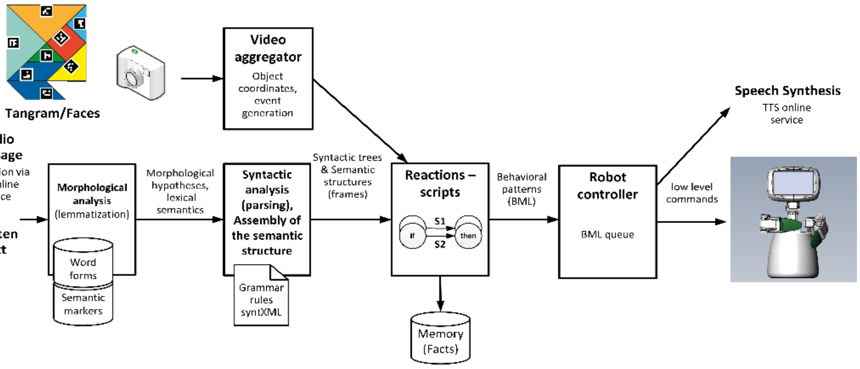
Tangram/Faces
Video aggregator
Object coordinates, event generation
Speech Synthesis
TTS online service
Audio message
recognition via STT online service
Written text
Morphological analysis (lemmatization)
Word forms
Semantic markers
Morphological hypotheses, lexical semantics
Syntactic analysis (parsing), Assembly of the semantic structure
Grammar rules syntXML
Syntactic trees & Semantic structures (frames)
Reactions – scripts
If
S1
S2
then
Memory (Facts)
Behavioral patterns (BML)
Robot controller
BML queue
low level commands

Fig. 1. General architecture of the conceptual processing system.

Speech processing module should represent the utterances in a form, sufficient to calculate the distances to the premises of scripts. We rely on shallow semantic representations: semantics of a single clause is represented as a *semantic predication* – a set of semantic markers, distributed between semantic valencies, with possible lexical ambiguity. In the same way, the premises and inferences of scripts are also represented as *semantic predications*, with no ambiguity (see Table 1). A script is searching for the premise semantic predication, moves to the inference semantic predication and may execute actions in BML format [15], assigned to both predications. The general architecture of the system is represented on Fig. 1. Input text can be received from a text source or as an oral speech, in this case Yandex Speech API is used to decode the signal to the written form. Each recognition result is processed by syntactic parser with the Russian grammar in syntXML format [16], containing about 600 syntactic rules, and a dependency syntactic tree is constructed.

During the processing, a predicate is assigned to *predicate* valency, and the actants are assigned to an *agent*, *patient*, *instrument* etc. – following a modified list of semantic valencies by Fillmore [17]. In case of a compound sentence, several semantic predications with co-reference links are constructed. In case of morphological or syntactic ambiguity, a set of syntactic trees may be processed in parallel. A neural network is used to evaluate each syntactic link and to select the best trees. Semantic representation of each valency is constructed as an aggregation of semantic markers from all the words within the subtree of the valency, e. g. for a *preposition phrase* a semantic representation is a sum of markers from the *preposition*, *noun* and all the *adjectives*. In case of lexical ambiguity, a Cartesian product of semantic sets is constructed.

2.4 Design of lexical semantic dictionary

The construction of semantics relies on the sets of semantic markers, assigned to words in the semantic dictionary. To annotate the words we use a set of 4778 markers consisting of (a) 209 focal markers of d-scripts, (b) 669 markers from a semantic dictionary [18], and (c) 3900 semantic markers, assigned to words after clustering of word2vec vectors. To generate markers, basing on *word2vec*, we have clustered nouns to 2000 clusters and 600 “superclusters”, and assigned to all the words one marker for the cluster and one marker for the supercluster. This procedure provides a two level “ontology” with one “low level” and one “high level” (supercluster) marker for each word. In the same way verbs were divided into 1000 clusters and 300 superclusters. Table 1 shows a semantic representation for un utterance *Our feelings betray us*. Lexical semantics is distributed into “1 1”/ “1 2” ambiguous slots (*betray* may be a communicative or behavioral action). Markers assigned after word2vec clusters are indicated by “@”.

Table 1. Semantic representation (predicate structure) of *Our feelings betray us (/lie to us)*.

Predicate	Agens	Patient
1 1, 1 2 present tense	1 1 many	1 1 somebody
1 1, 1 2 assertive	1 1 abstract	1 1 egocentric – me
1 1 to communicate	1 1 negative emotions	1 1 other person
1 1, 1 2 DECEPT attribute	1 1 positive emotions	1 1 physical object
1 1 to report	1 1 @feeling_NOUN	1 1 principal – speaker

1 1, 1 2 @to_simulate	1 1 @173_NOUN	1 1 set of people
1 1, 1 2 @214_VERB		
1 2 social action		

2.5 Design of r-scripts

Rational scripts should cover and classify most of the situations that appear regularly in incoming texts or multimodal events. Scripts contribute to the resolution of ambiguity and provide inferences (perspective component). Scripts should distribute the incoming events in classes, where each class has a generic semantic representation and provides reactions (inferences), relevant to all the inputs, attributed to this class. While designing the scripts, we have to aggregate the case frames (predicative structures), where all the words in each valency constitute a coherent set, e.g. {*team, sportsman, champion, player, hockey_player*} *defeated* {*host, guest*} and {*candidate, promotee*} *defeated* {*mayor, governor*}. Then for each case frame we look for a prototype example to represent the situation. On the first stage the following two methods were applied to construct the premises of scripts:

1. For most of verbs we have defined the most frequent words in each valency in the text corpus (over 80 million wordforms). The results were manually inspected in case words from several different metaclusters occupied a valency, the whole case frame was divided into two or more case frames.
2. For each verb we have automatically selected facts, where (a) all the words in the *agent* valency belong to the same supercluster and (b) words in the *patient* valency (for transitive verbs) belong to the same supercluster. So, we have got different case frames for different meanings of the verb, as well as the different case frames for direct and metaphoric meanings:
   (i) {*finger, palm, hand*} *gripes* {*shoulder, finger, palm, hand*}
   (ii) {*anxiety, fear, depression*} *gripes* {*neck*}

Script premises were defined for the selected case frames as the semantic predications withholding the semantics of the most frequent words in each valency. 1619 scripts were constructed after the procedure, defining “prototypical” situations, to be considered in incoming texts. A typical answer was designed for each script as the sequence of words in the valencies – this simulates a strategy, where the agent replies with a generalization or a typical example for an incoming phrase.

R-scripts also contribute to the resolution of ambiguity: for the set of syntactic trees, constructed after audio recognition (ambiguous audio signal), syntactic processing (syntactic ambiguity) and the construction of semantics (lexical ambiguity), we select the tree with semantic predications on average closest to the set of scripts. It means, that the system selects the semantic representation, which is the most emotional (is the closest to d-scripts) or the most regular (is the closest to r-scripts). The most probable semantic representation is saved to a database and is indexed by the script, used to select this representation.

In a multimodal dialogue mode, an incoming semantic representation can invoke several scripts, used to generate a compound reaction to the input. The scripts are divided into groups and are associated to *microstates*. A script sensitivity is proportional to the activation of microstate: by modifying the microstate one can design (a) an *emotional* agent, (b) a *rational* agent, or (c) an aggressive or depressive agent.

2.6 Memory

The semantic representation of input sentences is indexed and saved to a database to support long-term memory and perspective question answering. Phrases are indexed by a script, used to select the meaning, and can be retrieved by the script index – all the predications, which correspond to the script premise, or by an arbitrary semantic pattern, for example, corresponding to the semantics of an input question. Table 2 represents sentences from the database, corresponding to the pattern ‘feelings deceive’. We suggest that question answering may retrieve utterances from the array of all the analyzed texts. An advantage of the method is that the knowledge of an agent is directly enriched through the automatic analysis of incoming texts, without any retraining. We also expect that scripts can provide more flexible Q&A by training on Q&A pairs, where a script premise corresponds to a question and script inference corresponds to an answer.

Table 2. Examples (facts) for *Our feelings betray us* semantic pattern from the memory base

Fact	Sentence
1094963	"anger troubled him, pushed him to insolence."
2527406	"the fact is that physical sensations deceive them, because…"
8446093	"no, you know that feelings do not deceive you."
146731	"Is it possible, that the sixth sense deceives him?"

3 Conclusion

In a conceptual processing system for a companion robot, semantic predication can be used as a basic data structure in internal interfaces between the modules of speech comprehension, visual perception, action, and memory. This architecture simulates balanced rational/emotional reaction, parallel processing of alternative reactions and the construction of a compound behavioral pattern, as suggested by linguistic observations in multimodal communication. It also manipulates semantic markers following the perception of speech and video recognition, as well as executes reactions on a bottom-to-top basis – activates scripts, and on top-to-bottom basis – converges an incoming representation with the premise of the selected script. In sum, the system offers a compound architecture for the main functions for a companion robot.

4 Acknowledgment

The present study has been mainly supported by the Russian Science Foundation, project № 19-18-00547.

The development of some of the components in the robot control system is supported by Kurchatov Institute project “Development of a semiotic system for multimodal remote control of the behavior of a semi-autonomous mobile robotic platform using natural language”.

References

1. Jurafsky, D.: Speech & language processing. Pearson Education India (2000).
2. Jurafsky, D., Martin, J. H.: Speech and language processing (draft). Chapter Dialogue Systems and Chatbots (Draft of October 16, 2019).
3. Cahn, J.: CHATBOT: Architecture, design, & development. University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science (2017).
4. Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., King, E.: Conversational ai: The science behind the alexa prize. arXiv preprint arXiv:1801.03604 (2018).
5. Laird, J. E., Newell, A., Rosenbloom, P. S.: SOAR: An architecture for general intelligence. Artif. Intell. vol. 33. № 1. pp. 1–64 (1987).
6. Schank, R. C., Abelson, R. P.: Scripts, plans, goals, and understanding: an inquiry into human knowledge structures. Hillsdale, N.J., New York: L. Erlbaum Associates (1977).
7. Minsky, M. L.: The Society of Mind. New-York, London: Touchstone Book (1988).
8. Sloman, A., Chrisley, R.: Virtual Machines and Consciousness. J. Conscious. Stud. vol. 10. № 4–5. pp. 133–172 (2003).
9. Sharonov, I. A.: Interjection in speech, text, and dictionary. M.: RGGU (2008).
10. Kotov, A. A.: Description of speech exposure in a linguistic. Computer Linguistics and Intelligent Technologies. M.: Nauka. pp. 299-304 (2003).
11. Barsalou, L. W.: Frames, concepts, and conceptual fields. Frames, fields, and contrasts: new essays in semantic and lexical organization. Hillsdale, N.J.: L. Erlbaum Associates. pp. 21–74 (1992).
12. Yeh, W., Barsalou, L.W.: The situated nature of concepts. Am. J. Psychol. vol. 119. № 3. pp. 349–384 (2006).
13. Clore, G. L., Ortony, A.: Cognition in Emotion: Always, Sometimes, or Never? Cognitive Neuroscience of Emotion. Oxford Univ. Press. pp. 24–61 (2000).
14. Glovinskaya, M. Ya.: Hidden hyperbole as a manifestation and justification of verbal aggression. Sacred meanings: Word. Text. The culture. M.: Languages of Slavic culture. pp. 69–76 (2004).
15. Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., Vilhjálmsson, H.: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. Intelligent Virtual Agents. pp. 205–217 (2006).
16. 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).
17. Fillmore, C. J.: The Case for Case. Universals in linguistic theory. New York: Holt, Rinehart & Winston. pp. 1–68 (1968).
18. Shvedova, N. Yu.: Russian semantic dictionary. Explanatory dictionary systematized by classes of words and meanings. M.: Azbukovnik (1998).