Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 169 (2020) 63–69

Procedia Computer Science

www.elsevier.com/locate/procedia

Postproceedings of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2019 (Tenth Annual Meeting of the BICA Society)

Speech Understanding System for Emotional Companion Robots

Artemiy Kotov[a,b]*, Nikita Arinkin[a,b], Liudmila Zaidelman[a,b], Anna Zinina[a,b]

[a]*Kurchatov Institute, Moscow, Russia*
[b]*Russian State University for the Humanities, Moscow, Russia*

Abstract

Within the project of emotional robot F-2 we develop a natural text parser for automatic speech comprehension. It is aimed at the construction of semantic representation and the selection of "scripts" – units for inference modelling and the selection of emotional reactions for the robot. The design of the speech understanding system follows the traditional concept of linguistic levels: it consequently constructs morphological, syntactic and semantic representations. Unlike neural networks, these representations are readable for a developer, so the accumulated data is available for statistical and analytical processing. Parser accepts written text after speech recognition (during actual talks with the robot) or during processing of everyday news and blogs from the internet. Parser may save text representations on each level of linguistic model to a database. In particular, semantic representations from daily internet processing are used as an accumulated knowledge database for the robot. In this paper we discuss the approach to the model of understanding through a competition of scripts on a robot companion.

Keywords: : Emotional robot; semantics processing; text comprehension

1. The Approaches to the Problem of Machine Understanding

The problem of machine understanding of natural text is one of the cornerstone problems of cognitive psychology, with the main discussion centered around the *Chinese room argument* by John Searle [1]. In his thought experiment Searle has shown, that a computer program has no possible room to hold or to implement the natural function of

* Corresponding author. Tel.: +7-499-196-71-00 # 31-10
E-mail address: kotov_aa@nrcki.ru

10.1016/j.procs.2020.02.115

understanding. In the following discussion numerous counter-arguments were suggested by his opponents, showing, for example, that the function of understanding may engage numerous procedures and be simulated within the entire machine, even if the operator (observer) does not *understand* the core of incoming stimulus through his management actions. Many approaches to the understanding are closely connected to the notion of consciousness. Sometimes it is suggested that consciousness provides a kind of internal *space* to support the function of understanding. So theoretical approaches to these two functions are closely interconnected.

In our design of cognitive architecture, we rely on the following principles:

(1) Understanding is a communicative function: one feels that another understands him, if the latter shows some accurate communicative responses. Following this principle, we develop a natural multimodal corpus – Russian Emotional Corpus (REC) with annotation of communicative actions in natural emotional situations: speech, gestures, facial expressions, head and body movements, etc. We also design a physical robot F-2, capable of showing a number of communicative responses, and reproducing key behavioral patterns, extracted from REC.

(2) Understanding is a procedure on the basis of text semantics. So, the process of understanding should rely on some text representation, considered as semantically reliable: e.g. this representation should be similar for synonymic texts. This representation should also be understandable to a researcher, so we do not rely on neural networks or word embeddings as representations of text semantics. For the purposes of (1) this representation should be sufficient to invoke communicative responses, considered as an *understanding communicative reaction* by an observer.

Understanding is a compound reaction. In order to simulate the understanding, we design (a) a speech processor (parser) with the construction of explicit semantic representation, and (b) a companion robot F-2 to simulate communicative reactions. Further design is executed on the basis of speech processing simulation by the parser and the simulation of communication by the robot.

2. F-2 Robot Architecture

The approach to the simulation of understanding is based on the development of theoretical and software architecture to process incoming text and generate behavioral responses for F-2 robot (Fig. 1).

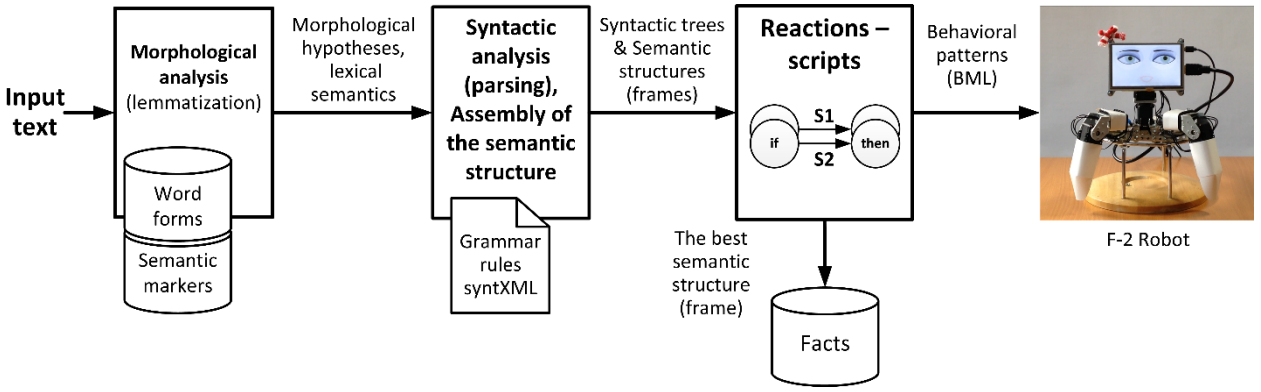

Fig. 1. Architecture of F-2 cognitive processing. Input text is processed on morphological and syntactic levels. The generated semantic structures are used to activate scripts, which generate behavioral patterns in BML format to be executed by F-2 robot.

2.1. Input processing and morphological analysis

The robot may receive oral or written texts at its input. Oral texts are converted to written form via the Yandex Speech API service – several possible recognition results can be received from the service, each is processed to the semantics level, where the ambiguity should be solved. Written texts may be fed as text files (like fiction) or via the

RSS subscriptions. Speech component of the robot in operated daily in a standalone mode – without the physical robot. It downloads about 7,000 sentences from top news and blogs, construct the semantic representations, suggests a possible reaction (*script*) for the robot and saves the results to a database (4.5 million sentences are processed for over a year, 9.2 million facts are extracted).

Morphological analysis associates each wordform to a lexeme and assigns a set of grammemes: *gender*, *case*, *number* etc. Several morphological hypotheses may be assigned to one word as the disambiguation takes place on further linguistic levels. For the morphological analysis we use a dictionary, derived from OpenCorpora project [2, 3], it contains 100,000 lexemes (1.5 million wordforms) saved to an SQL database. Numbers and compound words (like: *47*, *3-rd, 5%*) are processed by regular expressions unit, unknown words are processed by a neural network unit, which outputs from 6 to 10 hypotheses for further syntactic analysis.

The most frequent words in the main dictionary are linked to a set of semantic markers (semantic set): from 1 to 18 markers are assigned to 30,000 lemmas. We rely on the semantic primitives [4] and on lexical annotation in semantic dictionary [5]: 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 [6].

2.2. Syntax processing

We implement left-to-right syntactic parser on a fixed list of rules to construct a syntactic tree (or trees) basing on the input wordform sequence. Wordforms are sequentially fed into a stack. In case of lexical ambiguity an extra stack is created for each ambiguous form. Syntax parser contains a description of grammar as a set of rules on syntXML language [7]. A rule is defined as a sequence of segments with a possible reduction. Rule head ***h*** can replace the right-hand side of the rule as a new segment (1), or be a member of the right-hand side (2) and subordinate all the other right-hand side segments.

$$\boldsymbol{h} \rightarrow \langle a, b, \ldots n \rangle \quad (1)$$

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

For each segment within a rule one can (a) verify if the segment is a specific lexeme, (b) verify the presence of a particular grammeme in the segment, (c) check grammatical agreement with the other segments of a rule, (d) assign new grammemes or rewrite the existing grammemes, (e) copy specific grammemes to the rule head ***h***. This variety allows us to describe in syntXML form both immediate constituent grammars – where the head ***h*** is generated by the rule and is not contained in reduced sequence, like *VP*, *NP* nodes, as well as the dependency grammars, where the whole sequence is reduced to one of its elements, like *Verb* – head of the verb phrase. Right-hand side of a rule may have a variable number of segments: 1 or more. This allows us to define not only binary relations, but also syntactic relations with numerous segments. For example, noun conjunction phrase can be defined as a reduction to a noun phrase (3) or as a reduction to the head noun (4).

$$NP \rightarrow \langle Noun_1, Conj, Noun_2 \rangle \quad (3)$$

$$Noun_1 \rightarrow \langle Noun_1, Conj, Noun_2 \rangle \quad (4)$$

The grammar allows optional segments with the right-hand side of the rule. For example, a noun conjunction can be defined as a co-occurrence of two nouns with an optional conjunction (5).

$$NP \rightarrow \langle Noun_1, [Conj,] Noun_2 \rangle \quad (5)$$

In the actual grammar description, we combine the two approaches: verbs and nouns are considered as the heads of the corresponding syntactic groups – as in dependency grammars, while conjunctions are described as a new syntactic node – as in immediate constituent grammars (5). A sample tree is represented on Fig. 2.

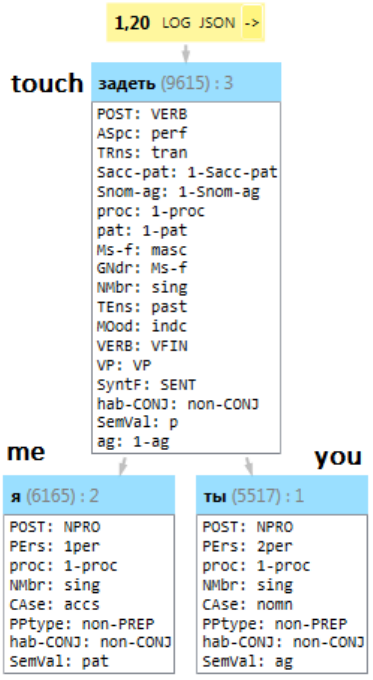

Fig. 2. Syntax tree for a sentence *You touched me*

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 is preserved for further steps, while stacks with lower scores are discarded (in actual tests we operate with a heap from 256 to 512 stacks). This method suggests a balanced approach to homonymy: on one side, all the cases of lexical and syntactic homonymy are processed, on the other side, only stacks with the best syntactic structures are preferred. By the end of a sentence the parser may generate a set of syntactic trees: a tree is generated within each stack, which is reduced to a single head. In case several trees are generated for a sentence, the preferred tree is selected in the *scripts* component depending of its semantic consistency. We expect, that for compound emotional reactions like irony [8, 9], and for studies of the architecture of consciousness [10] – several trees 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.

A semantic representation is constructed for each syntactic tree. A semantics of a syntactic clause is represented by a *semantic predication*: a number of valencies with semantic sets in each valency (Table 1). We use a set of 23 valencies: *agent, patient, instrument, time, place* etc., based on [11]. A predicate is assigned to a dedicated valency "*p*". In case of ambiguity, several meanings (sets of semantic markers) are generated for an ambiguous word or a sub-tree. As shown in Table 1, the word *touch* is processed as ambiguous, pronoun *you* is assigned to the listener ('egocentric') while the pronoun *me* is assigned to the speaker of the incoming phrase ('other_person', 'principal').

Table 1. Semantic representation for an utterance *You touched me*

Valency	Meaning: a set of semantic markers
Predicate (*touch*)	1. touch, assertive, past_time 2. cause_emotions, assertive, past_time
Agent (*you*)	1. egocentric (me), person, object
Patient (*me*)	1. person, other_person, object, principal

A semantic set within a valency combines the semantic sets of all the lexemes within the corresponding syntactic sub-tree – e. g. *noun phrase* (NP) is always assigned to a separate valency and combines semantic markers from

prepositions, adjectives and *nouns*. A semantics of a multi-clause sentence is represented by several semantic predications, combined by co-reference relations. These semantic representations are further evaluated within the scripts component to classify the input representation and to select a behavioral reaction, if running on a robot.

2.3. 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. Scripts are distributed into several levels: emotional and rational scripts. While an emotional script as a concept corresponds to the notion of *proto-specialist* by Minsky [12] the concept of level organization of cognitive units inherits Sloman's *Cognition and Affect* (CogAff) architecture for the emotional agents [13, 14]. Scripts *if*-conditions (*premises*) are defined as *semantic predications* – similar to semantic structure of a single clause. For each semantic predication in an incoming utterance parser calculates the distance to each premise of the existing scripts. 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 premise and (b) proportionally to the sensitivity of the script and the current state of the agent: prior activation of scripts by former incoming events. Table 2 represents top scripts, activated by the sentence *You touched me.*

Table 2. Top scripts, activated by an utterance *You touched me*

Distance to script	Script	Sample speech output
2.3482	ME*DANGER	*I can beat you!*
2.3482	ME*CARE	*I care about you!*
2.3482	RULE: action	*I shall compensate it!*
2.3482	RULE: etiquette	*I am sorry!*
2.3482	me*INADEQ	*I always make people suffer!*
2.3482	NEG*negative physical influence	*Go, get it!*
2.3482	RULE: make new rule	*I have to be careful!*
2.3482	CONSIL	*It did not make harm!*

Each script represents a classification basis for an incoming phrase: so, the agent calculates a variety of representations to classify the incoming event. After the evaluation of distances, the agent calculates the activation of scripts, proportionally to its personal profile – sensitivity of each script. E. g. an agent with *depressive* profile prefers scripts like **me*INADEQ** ('I make inadequate actions') wile a *polite* agent prefers the etiquette reactions from **RULE: etiquette** (*I am sorry!*). The agent may simultaneously invoke several scripts and, for example, distinguish output for "internal deliberation" (like *I always make people suffer!*) and "polite external output" (like *I am sorry! I shall compensate it!*).

The actual behavior of the robot is composed of the behavioral packets in BML format [15, 16], linked to the robot gestures, designed on the basis of communicative pattern in the REC emotional corpus [17]. An activated script generates a desired BML behavioral response, and all the generated BML packets compete for the execution on the robot. Several most activated scripts may mix their BMLs while performing on the robot, resulting in compound and rich communicative behavior

3. Emotional classification of utterances

To verify the speech processing system, we run it on a server in daily mode. The system accumulates utterances from popular news and blogs via an RSS subscription. Each utterance is processed up to the semantic representation, which is saved to a database with the information on script activation. This allows us to (a) search the database for the utterances, matching specific semantic pattern, and (b) study the general patterns on script activation.

The search for semantic patterns surpasses traditional keyword search, as it allows parser to distinguish actants

(*agent*, *patient* etc.) with different semantic markers: Table 2 represents typical utterances, retrieved for a semantic pattern ‘state limits a person’.

Table 3. Sample utterances for a semantic pattern ‘State suppresses a person’

Semantic pattern	Utterances (examples)
Predicate: ‘limit, control’ (markers, typical for LIMIT script) Agent: ‘state’ Patient: ‘person’	*According his words, the state will control any American leader.* *France obliged the Russian billionaire Roman Abramovich to pay € 1.2 million in taxes for the mansion of the Château de la Croix in the Provence-Alpes-Cote d’Azur region.* *But they are controlled by the state, so the problem still remains.* *The United States restricts money transfers and trips to Cuba, imposes sanctions against the central bank of Venezuela, as well as against the son of the president of Nicaragua and Banco Corporativo.* *The USA doesn’t even control half of Syria.* *Accordingly, now the United States will freeze any property of the new figures from the sanctions list, and will also block any financial transactions involving them.*

This representation allows the robot to accumulate emotional or situational markers for agents: like to judge whether a ‘state’ is more likely to ‘limit’ people, or more likely to ‘protect’ people, retrieving text examples for the both contradictory judgements. The aggregation of actants, occupying the same valency for similar predicates, allows to build a semantic category. For example, ‘beverages’ can be represented as a set of patients for predicates with semantic marker ‘to drink’. Basing on this data an actant can be evaluated to more or less typical member of a category (*tea* and *glass* are more typical beverages than *blood*).

For over one year of daily operation the parser has accumulated 9,2 million facts for 4,5 million sentences (a fact corresponds to a clause in a sentence). This gives the following picture of the representation of scripts across the set of sentences (Table 4).

Table 4. Sample utterances for a semantic pattern ‘State suppresses a person’

Script	Script pattern	Number of cases	Percent
INCOMPREHENSION	‘the situation is unclear’, ‘this is something unbelievable or stupid’	4854554	52,7
DANGER	‘the situation affects our life or health’	578317	6,3
CREATION (other agent)	‘they create a beautiful situation’	508839	5,5
LIMIT	‘they limit us’	332073	3,6
COMFORT	‘we are in a comfort situation’	320657	3,4
PLAN	‘they make some plans against us’	310498	3,3
APPROPR (to move)	‘something valuable moves away’	294430	3,2
DECEIVE	‘they lie’	261222	2,8
UNUSUAL (moving to a place)	‘we get into an unusual place’	261043	2,8
APPROPR (to take)	‘they withdraw some values from us’	233428	2,5
MANIPULATION	‘they try to control us’	222569	2,4
APPROPR (to get)	‘we receive a great thing’	118543	1,3
CREATION (me agent)	‘I create a beautiful situation’	65578	0,7
APPROPR (to have)	‘they have some values’	58930	0,6
HAPPINESS	‘this is a funny situation’	48298	0,5

As evident from the table, most of the utterances are not classified by the parser as corresponding to any emotional situation – robot does not ‘comprehend’ them and tends to react as to something ‘unbelievable’ or ‘stupid’. These are mostly “rational” utterances with a low similarity with any emotional situation (scripts premises). Other utterances are classified as positive emotional situations (e. g. ‘we are in a comfort situation’) or negative ones (‘they make some

plans against us'). This variety defines the range of possible interpretations of an incoming situation and possible reactions of the robot.

4. Conclusion

We address the notion of understanding as a compound phenomenon, involving not only processing of incoming text, but also a reaction of the listener. In this chain understanding can be represented as a classification of the incoming text semantics by units, responsible for inference or reaction of the robot. Since only emotional output reactions are considered on the present stage of the project, the suggested approach covers less than 50% of incoming utterances. Further extensions may cover "rational" reaction as well as inferences to derivative text representations. At the same time the reduction of the input semantics space to the space of emotional representations is an important aspect of understanding, and the simulation of the process, as we suggest, can be considered as a shallow model of text understanding by a companion robot.

Acknowledgements

The development of speech parser is supported by RSCF project 16-29-09601.

The analysis of emotional text patterns (threats) in a wide corpus is supported by the Russian Science Foundation project 17-78-30029.

References

[1] Searle, J. (1980). "Minds, brains, and programs." *Behavioral and Brain Sciences* **3 (3)**: 417–424.

[2] Bocharov, V. V, Alexeeva, S. V, Granovsky, D. V, Protopopova, E. V, Stepanova, M. E., Surikov, A. V. (2013). "Crowdsourcing morphological annotation." *Computational Linguistics and Intelligence Technologies* **12 (19)**: 109–114.

[3] Protopopova, E. V., Bodrova, A. A., Volskaya, S. A., Krylova, I. V., Chuchunkov, A. S., Alexeeva, S. V., Granovsky, D. V. (2014). "Anaphoric annotation and corpus-based anaphora resolution: An experiment." *Computational Linguistics and Intelligence Technologies* **13 (20)**: 562-571.

[4] Wierzbicka, A. (1972). "Semantic primitives." Frankfurt/M., Athenäum-Verl.

[5] Shvedova N.Yu. (1998) "The Russian Semantic Dictionary." [Russkij semanticheskij slovar'. Tolkovyj slovar', sistematizirovannyj po klassam slov i znachenij]. M., Azbukovnik

[6] Kotov, A.A. (2003) "Mechanisms of speech influence in media texts." [Mekhanizmy rechevogo vozdejstviya v publicisticheskih tekstah SMI]. Dis. … kand. filol. nauk; 10.02.19; Zashchishchena 23.06.03. M.

[7] Kotov, A., Zinina, A., Filatov, A. (2015). "Semantic Parser for Sentiment Analysis and the Emotional Computer Agents." *Proceedings of the AINL-ISMW FRUCT 2015*: 167–170.

[8] Attardo, S., Eisterhold, J., Hay, J., Poggi, I. (2003). "Multimodal markers of irony and sarcasm." *Humor - International Journal of Humor Research* **16 (2)**: 243–260.

[9] Kotov, A. (2009). "Accounting for irony and emotional oscillation in computer architectures." *Proceedings of International Conference on Affective Computing and Intelligent Interaction ACII 2009*: 506–511.

[10] Kotov, A. A. (2017). "A computational model of consciousness for artificial emotional agents." *Psychology in Russia: State of the Art* **10 (3)**: 57–73.

[11] Fillmore, C. J. (1968). "The Case for Case." In E. Bach & R. T. Harms (Eds.) *Universals in linguistic theory*, New York: Holt, Rinehart & Winston: 1-68.

[12] Minsky, M. L. (1988). "The Society of Mind." New-York, London: Touchstone Book.

[13] Sloman, A. (2001). "Beyond Shallow Models of Emotion." *Cognitive Processing* **2 (1)**: 177–198.

[14] Sloman, A., Chrisley, R. (2003). "Virtual Machines and Consciousness." *Journal of Consciousness Studies* **10 (4–5)**: 133–172.

[15] Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Vilhjálmsson, H. (2006). "Towards a Common Framework for Multimodal Generation: The Behavior Markup Language." *Intelligent Virtual Agents*: 205–217.

[16] Vilhjálmsson, H., Cantelmo, N., Cassell, J., E. Chafai, N., Kipp, M., Kopp, S., van der Werf, R. (2007). "The Behavior Markup Language: Recent Developments and Challenges." *In Intelligent Virtual Agents*: 99–111.

[17] Kotov, A., Budyanskaya, E. (2012). "The Russian Emotional Corpus: Communication in Natural Emotional Situations." *Computational Linguistics and Intelligence Technologies* **11 (18)**: 296–306.